ATHABASCA UNIVERSITY

A CAUSAL MODEL OF WRITING COMPETENCE

BY

CLAYTON CLEMENS

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL

FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

MASTER OF SCIENCE IN INFORMATION SYSTEMS

SCHOOL OF COMPUTING AND INFORMATION SYSTEMS

ATHABASCA UNIVERSITY

JUNE, 2017

Approval Form

Athabasca University
FACULTY OF GRADUATE STUDIES

The future of learning.

Approval of Thesis

The undersigned certify that they have read the thesis entitled

"A Causal Model of Writing Competence"

Submitted by

Clayton Clemens

In partial fulfillment of the requirements for the degree of

Master of Science in Information Systems

The thesis examination committee certifies that the thesis
and the oral examination is approved

Supervisor:
Dr. Vive Kumar
School of Comuting and Information Systems
Athabasca University

Committee Members:
Dr. Maiga Chang
School of Comuting and Information Systems
Athabasca University

Dr. Kinshuk
College of Information
University of North Texas

External Examiner:
Dr. Ben Daniel
Higher Education Development Centre
University of Atago

June 13, 2017

1 University Drive, Athabasca, AB, T9S 3A3 Canada
P: 780.509-7536 | Toll-free (CAN/U.S.) 1.800.788.9041 (7536)
fgs@athabascau.ca | fgs.athabascau.ca | athabascau.ca

Abstract

Traditionally, assessing competence in English composition involves ignoring most of the steps in the writing process and focusing only on the final submission. The writing process cannot be assessed sufficiently in a traditional setting, and no formal structures exist for improving process-based writing competences.

In online distributed learning environments, students' steps to complete a composition can be recorded at granular levels using content capturing methods. For each step in the writing process, numeric qualities of the text can be determined using natural language processing. These metrics combine into measures of writing competence.

The volume of generated data is cumbersome to submit through distributed environments, and time-consuming to process. This research overcomes the difficulty by simulating writing traces from a corpus of completed essays. A writing analytics engine processes the traces, and the resultant metrics are examined using causal inference, producing multiple statistical models of writing competence as systems of interventions.

Keywords: writing, causal inference, simulation, analytics

Dedication

I dedicate this work to my family, the greatest support and encouragement anyone could hope for.

Acknowledgments

I would be remiss if I didn't first thank my Creator, the author and finisher of my faith, without whom I can do nothing.

Outside of the divine, I have had such an abundance of human help and assistance I hardly know where to begin. To my supervisor, Vive, thank you for your patience this last half-decade as I wobbled my way around learning how to manage life, causality, and thesis-writing. And thank you especially for your love of badminton as, in a sense, it established me in my career.

To the Quebec team, David Boulanger, Jeremie Seanosky, Isabelle Guillot, I would still be floundering around aimlessly if it weren't for you. You have been absolutely instrumental in this work, and I can't thank you enough for your brilliance and excellence in everything you put your hands to.

To Diane Mitchnick, thanks for the chats when we were working together on this project, and at all the conferencing we managed to do. Your passion for your own project has been an inspiration to me. To Colin Pinnell, Jason Bell, Steve Harris, thank you guys for your wealth of technical knowledge, which you so willingly share. I appreciate your unique perspectives, and the dash of humor you brought into an otherwise serious world. To Rahim Virani, thank you for standing with me, struggling with me, and randomly calling me on Skype to catch up. Peace, brother! To Stella Lee, thank you for your passion for academia and zest for life that have always been an encouragement to me.

To my wondrous collection of parents, David and Jennie McDaniel, Ron Clemens, Dana Harvie, I owe everything I have accomplished to you. Thank you for supporting me in every possible way, for being interested in what I was doing, for your constant prayers for me, for kicking my butt when I've wanted to quit, for thinking the world of me even when I didn't deserve it.

To Jilian Campbell, my boss and my friend, thank you also for not letting me quit, and for constantly inspiring me to make a difference every day.

Table of Contents

List of Tables

List of Figures

Chapter 1 - INTRODUCTION

Writing is unique among skills learned in school. Once a student graduates from high school he or she may never again be required to solve a polynomial equation, mix two chemical compounds, or recall the first law of thermodynamics. However, no one can say with certainty that once they graduate, they will never write another word. This is especially true in today's culture of Internet communication, social media, and the myriad other pervasive forms of non-verbal interaction. With access to computers and mobile devices ubiquitous almost worldwide, everyone writes, if only to send a text message or post a status on Facebook.

Those who pursue post-secondary education are held to a higher standard of writing competence. Certainly in literature studies, but even in the sciences, reports, essays, and presentations are essential to a passing grade in most courses. Further emphasis is placed on writing in graduate and doctoral studies, where research papers, journal articles and theses like this one are required to earn a degree.

Problem

Given that writing is so important to continuing education and to everyday life, it is imperative that it is a primary focus of student learning. This is certainly true in the early elementary grades, where reading and writing are among the primary skills that children need to learn [1]. Once students obtain these basic competencies, however, the conditions of their junior-high and high school learning environments must be of very high quality in order to promote high levels of writing competence following their secondary education. Without these particular conditions, students often do not transition to post-secondary schooling with the skills they need to write at the higher level demanded in degree programs [2]. Exacerbating the problem for students in today's society are the negative impacts of text messaging and social media on writing skills, which are both characterized by heavy abbreviation and a lack of proper mechanics [3].

This problem of underdeveloped writing competencies may stem from poor curriculum, or a lack of useful feedback. The traditional model of writing assessment involves a student writing about a proposed topic, in some cases having a guideline or template to follow (for example, introduction, conclusion, and three body paragraphs, each between 5-7 sentences). The professor receives the finished product only. That is, in the five steps of the writing process (pre-writing, writing, revising, editing, and publishing [4]), the professor sees, grades, and may comment on only the final stage. Unless a course is particularly designed to test a student on all five steps, the instructor is not getting the entire picture of a student's performance.

However, it is infeasible in a traditional classroom for instructors to monitor everything each student does as part of their writing work. It is conceivable, as mentioned above, to construct a course in which the students go through each phase of the writing process as a separate assignment. However, such a scheme would take a great deal of time and would be too cumbersome to become a standard method of assessment.

Even if the steps of the writing process could be separately assessed, there are still sub-processes within each step, especially the writing and revising stages, that couldn't be isolated. Moreover, the writing process is often recursive, with students looping back and forth between stages. Ideally, information on each student's writing process could be collected as they write, and given to the instructor to form a full picture of the student's task.

While the traditional classroom environment makes process-based data collection intractable, Internet-based mediums for content delivery provide a platform to implement it. Online learning is becoming a more common way to pursue post-secondary education, and even grade school.

Moreover, studies have shown that the use of computers for writing produces both a higher quantity and quality of work compared to pencil-and-paper environments [5]. By leveraging the power of online learning environments, it is possible to develop software that can do what a traditional classroom course cannot. In particular, keystroke logging can be used to trace every action the user makes on the keyboard while writing. Keylogging has been used for a long time in word processing applications in order to provide functionality to ‘undo’ and ‘redo’ buttons. Keylogging creates a stream of characters at particular timestamps, a complete history of exactly how a student went about a writing task.

Analyzing this stream using natural language processing techniques allows examination of the writing process by determining what the students’ competencies are and how they evolve over time. This collection of writing events, analyzed metrics, and inferred competence values make up a number of statistics that represent the evolution of the writing trace. Combining these for multiple students over multiple compositions, it is possible to determine how these competencies and concepts are related to each other developmentally over the course of a composition and over the course of multiple assignments. Statistical techniques can be used to propose and test several causal models of writing competence.

However, storing and processing detailed traces represents a large amount of network activity when multiple students and multiple compositions are involved. To convert this data to a usable form in real-time or even near-real-time is computationally intractable for anything but a very comprehensive suite of server hardware.

To combat these difficulties, the data can be simulated, as a first step into exploring the implications and usefulness of a causal model.

Research Question

Given the opportunity to increase knowledge of students' complete writing processes, and given that technology exists to trace those processes and determine competence information from them, the goal of this thesis is to address the following question:

How do the measurable factors that contribute to writing competence affect one another causally?

Data to answer this question will be simulated using a program called WriteSim that uses information from behavioral studies to determine how students act during the writing process. This simulator takes into account a number of behaviors, including student psychological profiles, keystroke timing, pausing, errors and revisions. The simulated data will be processed by an analytics engine called SCALE, optimized using a domain-specific set of tools that will analyze each text record and derive competence information from it.

The goal of this research is to develop and utilize this competence data in the creation of a formal model of competence development. Each of the competence factors output from SCALE using the simulated data will be analyzed for correlations and causal relationships using statistical inference methods.

A causal model of writing competence will assist in demonstrating how students learn concepts, and how they improve their skills over time.

Thesis Organization

This thesis is organized into five chapters. The first, this one, is an introduction to the topic. It briefly provides an overview of the problem, a description of the research question, and a small coverage of the methodology.

The second chapter will provide an in-depth review of the literature pertaining to writing, competence, analytics (both in general and specific to the domain), and the nature of causality (both from a philosophical and statistical perspective).

The third chapter explains in detail the methodology undertaken by the research, including an overview of the technology used and the detailed procedure.

The fourth chapter shows the result of the methodology, including proposed causal models for the competences and metrics being studied, the structural equation models underlying those models, and the comparative statistical viability of each model.

The fifth chapter gives the author’s conclusions based on the analysis, focusing on causal implications for each major competence, and gives recommendations regarding the pedagogical importance of the competences. The section includes proposed future research projects based on the results.

The back matter of the thesis includes an appendix with additional matrices determined in the results.

Chapter 2 - BACKGROUND

Writing is a discipline that has pervaded everyday life and education for centuries. Following is an in-depth coverage of the literature relating to writing as a craft, as a subject in education, how it can be quantified, and how competencies can be estimated and modeled.

Writing Education and Competence

Competence is classically defined in the dictionary as "possession of a required skill, knowledge, qualification, or capacity" [6]. Intuitively, to have competence in something is to be "good at it." This definition is somewhat ambiguous, however, leaving unanswered questions about exactly how "good at" something someone must be in order to be considered competent and how such proficiency can be measured. When speaking formally about competence, authors in the literature have developed a variety of meanings for the term. Sampson and Fytros [7] consider this literature and provide a consolidated definition: "a competence can be defined as a set of personal characteristics (e.g. skills, knowledge, attitudes) that an individual possesses or needs to acquire in order to perform an activity within a specific context."

This definition encapsulates three dimensions: context (the subject matter or area in which the subject should be competent), personal characteristics (an individual's skills, knowledge, abilities, attitudes, and other traits), and competence proficiency level (the degree of proficiency the individual demonstrates in the context). These dimensions are discussed in the literature to varying degrees.

The concept of profiling competence is millennia old. The Chinese, early Romans, and craftsman of the Middle Ages all had systems in place for determining how competent a given person was at their work in a variety of domains [7]. The modern study of competence began in the early 20th century with Taylor's book [8] on competence in the context of scientific management. He criticized the state of America, and the idea that competent individuals were simply ‘found’ or ‘born’, and proposed that such individuals should be trained instead.

McClelland’s 1973 article [9] sparked new interest in competence development in the context of human resources and education. The article is a strongly-worded criticism of America’s Scholastic Aptitude Tests, which correlated strongly with school grades, but didn’t address specific competencies. McClelland made a list of suggestions that urged testers to revise their methods to identify proficiency on a domain level. For example, to test someone’s competence in driving, they should be given a driving test rather than a pencil and a piece of paper.

The concept of creating competences using education, and subsequently testing and then certifying those competences, has become ingrained in society. In the example above, driver’s licenses serve as verification that someone is (ostensibly) a competent driver. Licenses, permits, and certifications all mark that a certain individual or group is able to do a specific thing well. Most often, testing of some sort is required to obtain these accolades. However, the Sampson and Fytros definition of competence is not solely based on knowledge or skills, but also attitudes. Further, it is difficult to agree on exact methods for measuring certain competences, such as leadership [10]. For this purpose, it is helpful to create competence models, which clearly describe or define what different levels of competence are in a particular domain.

A large body of research has been designated to the definition of competence, and the specification of various models within different lines of business [11, 12, 13]. Further, the definition has expanded to include competence for the entire organization; defining what skills, attitudes, and abilities a corporation has within the particular context of their business or market. In the literature, this is referred to as an organization’s ‘core competences’ [14, 15].

Some organizations have attempted to create standards for competences, so that organizations have a clear and common understanding of competences for a given domain. This type of competence model is a formalized structure that describes one or more of the three dimensions of context, individual characteristics, and proficiency levels. Some of the most popular formal models are IMS' Reusable Definition of Competency or Educational Objective (RDCEO) [16] and HR-XML [17], which is now part of the HR Open Standards Consortium.

Formal competence models give a more specific meaning in each context, and make it easier to objectively compare people who perform well at a particular task. Given wide adoption of such competence models, organizations and individuals would be able to easily assess how well they compare to the competition, or whether their unique mix of competences suit a given role or market share.

Education is concerned with the creation and assessment of competence. Traditionally, educators develop competence in students using learning materials, and assess it through the use of assignments and tests which verify the learned skills. This is a similar approach to certification, as discussed above.

Because education is a primary source of competences, the nature of learning is one of the most extensive research areas in existence. Within it, there are a variety of approaches and theories about the different ways people learn. For example, problem-based learning inspires knowledge exploration by focusing on a particular problem that students must acquire certain competences to solve [18]. The research has expanded even further since the late 1990s when the Internet began to become a learning environment, and an entirely new field called 'distance education' was born. This in turn sparked the study of self-regulated learning [19, 20], social learning [21, 22], and informal learning [23]. Additionally, education research has overlapped with computer science as more advanced automated tools and concepts for teaching and learning are developed alongside the study of online or distance environments [24, 25].

Writing is the act of using symbols to store information. These symbols take a wide variety of forms, and have evolved along with human societies. The mediums used for inscribing these symbols have also taken on a broad spectrum throughout history. Carvings and paintings were the earliest means, and the development of paper saw the quill and then the pencil becoming universal icons to indicate the act of writing. The printing press allowed the reproduction of large amounts of information, causing a monumental spread of knowledge throughout the world. A similar revolution has happened in the last century with a new medium, the personal computer. The development of computing and the Internet has thrust us forward into a global age of information sharing, where now the symbols used to store information have been digitized. This allows information to be stored on a massive scale, previously impossible to achieve [26].

Writing is an extremely important skill. Every member of a given society benefits from being able to translate their thoughts onto media for the purposes of sharing and retaining them [27]. In most cultures, writing is a minimum requirement for nearly every profession in existence. Written communication in office environments is essential, and even in labor-focused professions, clear descriptions are necessary for work orders, contracts, and incident reports. In the sciences and academia, technical and formal writing permeates research work. Moreover, in the modern world of mobile text communication and social media, the ability to write (however well) is necessary for day-to-day interactions [3]. It follows that writing competence should be well-understood and well-taught at all levels.

A challenge with competence in English composition is that not all of its aspects are necessarily a matter of objective correctness or incorrectness. In finite mathematics, for example, it is simple to grade each solution from a student as right or wrong. With such a domain, it is easy to assess competencies by simple questions. Once education goes beyond the level of mechanics (syntax, spelling, grammar, punctuation), many concepts in writing do not possess a binary correctness. Written communication can be subjective and open to interpretation. For example, the symbology or imagery of a poem as interpreted by a reader may be different than what the author intended.

This subjectivity requires the existence of rubrics and grading scales to provide a framework in which instructors can score different aspects of a piece of writing categorically and systematically.

In many rubric categories, instructors may not necessarily agree on whether a student has achieved a particular level of competence. These rubrics are used throughout the developmental years [28, 1]. To test adult writers for command of spoken and written English, the Common European Framework or Reference for Language [29] assigns a grade to different levels of competence. The framework is often used to test immigration applicants that don’t possess English as a first language.

Even with these models in place to measure and assess competences, they can be limited. A wide body of study still exists around the psychology and behavior of writers at many levels. There are a variety of best practices for fostering good learning environments for writing [30], and to ensure it is being assessed well [2, 31]. This research extends to focusing in on the specific needs of specific groups, such as those who speak English as a second language [32].

However, authors even as early as 1972 saw the true problem with assessing writing. Donald Murray [4] explains this in his article, which presents the perspective that writing should be taught as a process, rather than a product.

Today, the writing process is taught in its entirety. However, there is still a gap in the midst of curriculum, assessment, and competence: the finished product is always what is assessed in the end. With traditional tools, there is no efficient way to assess the student's entire writing process for a given composition, or to track a student's proficiencies in various aspects of the process over time. The field of learning analytics, however, may be able to build the bridge to cross this chasm.

Learning Analytics

The word 'analytics' can refer to the science of analysis, or it can refer to the metrics or statistics produced as a result of analyses. Often, the term is used in the latter sense, though both definitions are related to the concept of analyzing something. Analysis is defined as the act of separating a concept or entity into its constituent parts for the purpose of discovering more about its nature [33].

Analysis has its roots with the logic of ancient philosophers like Plato and Aristotle, and first used systematically by Euclid. The idea of analysis as a 'breaking down' of more complex thoughts influenced thinkers throughout the middle ages and became dominant during the scientific revolution [34]. The earliest forms of analysis in conjunction with mathematics and statistics to affect strategy and policy arose in the 18th century [35]. Following this, influential figures like Florence Nightingale increasingly used simple statistical analysis to make their cases [36].

Analysis is the root of the much more modern field called analytics, which has been generated by the ubiquitous availability of computing power, and the huge amount of data generated by the internet and large global organizations [37]. As a science, analytics is the methodology of analyzing data for patterns that allow the assessment of past events, informed decision-making in the present and prediction of the future given a set of actions. The word analytics can also refer to the patterns discovered from the data. There are a wide variety of applications of analytics. Business intelligence involves the evaluation of key performance indicators over time to drive profits or increase efficiencies. Web analytics are used by site owners and businesses to determine its degree of linkage from other sites, and the volume and nature of visitors to a site. Even more recently, social network analysis is used to discover the details of drivers behind social change on individual and collective levels [35].

The field of analytics that this thesis is primarily concerned with is learning analytics. The first formally recognized definition of learning analytics was given in 2011 in the call for papers of the first International Conference on Learning Analytics and Knowledge [38]:

“Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs.”

There is some question about the distinction of learning analytics from other similar disciplines such as educational data mining. Indeed there are many similarities between learning analytics and educational data mining, and the two communities often borrow from one another, sometimes only differing in their focus and origins [39].

Learning analytics grew up out of inspiration from big-data business intelligence, combined with the increasing popularity of online virtual learning environments. Such learning environments were built with the capability to log student browsing activity: their trace through a series of learning objects. Application of business intelligence to these logs led to further research in how to identify whether students were engaged or comprehensive in a distance education environment. In that way, learning analytics has ties to the study of self-regulated and coregulated learning as well [40]. This became even more relevant with the emergence of massive open online courses, in which registrations could occur in large volumes, but attrition rates were high [41]. Further, teachers and instructors became aware of the pedagogical value of learning analytics in assessing granular details of student work: the ability to assess process and not simply results of process [42].

The LAK 2011 definition above includes a series of actions (measurement, collection, analysis, and reporting), carried out upon a thing (data about learners and their contexts) with a certain goal in mind (understanding and optimizing learning and the environments in which it occurs). Research focuses on all three of these aspects, whether independently or as a whole. However, sets of data must be collected and measured before analysis, reporting, understanding, and optimization can occur.

Learning analytics data is often sourced from traces of student activity. In the infancy of learning analytics, this data was navigational in nature, reassembling a student's path through a given course based on the actions they took through the system. Recent developments, such as Purdue's Course Signals system, have expanded the premise of activity traces to include a broader variety of factors, including instructor grades, prior academic performance, and student characteristics [43]. Research has also begun to focus learning analytics on specific contexts; learning domains in which more detailed and relevant traces can be gathered.

Some of these interactions can give promising insights into student cognition. Training systems may include models that derive competence from minute traces like reaction time between events, how long a student lingered on a choice, and how often student changed answers during a test [44]. Other cognitive traits, such as degree of self- and co-regulation could be examined using these activity details as well.

While activity and navigational traces are promising, they are not often context-aware. For learning analytics to be effective for instruction on a particular topic, they must have knowledge regarding the domain the student is attempting to learn. Analytic systems must define the point at which something has been learned, should give a complete trace of the processes involved in learning, and provide a sense of whether the student conforms to pedagogical or cognitive principles.

However, a multitude of different subjects exist, each of them with unique knowledge that students hope to acquire. Moreover each of these domains have different needs, different means of gathering and representing the data, and different associated competences. Acquiring this specific learner data can be daunting when data on processes are not easily machine-readable. Application programming is an example of a field that is comparatively easy to trace. 'Records' are created with each compile, and program code already has a highly structured representation (via abstract syntax trees). Different attributes of the event are included within the object already. Depending on the language and environment, it's a matter of tracking the code at compile time, and then applying inferences on it to derive the user's construction process [45, 46].

Learning analytics is even broaching into subjects previously considered impossible for computers. Music can be captured via devices that support MIDI, and via audio recording with layers of software. Fuzzy logic can be applied to recordings in order to judge timing within reasonable parameters, and notes can be compared to expected tones for the composition [47].

With data available, there must be a motivation for using it. By the LAK 2011 definition, the motivation is to understand learners and optimize learning environments. Yet this understanding and optimization itself demands a motivation, and it is often an unspoken assumption in research that the true goal is to increase learners' performance and improve their experience by giving them and their instructors more pertinent and thorough information on their processes. Essentially, the true goal of learning analytics is to increase competences. Low-level domain data along with activity traces, when presented in a proper manner, can give a better sense of the relationship between learning objects and student performance. Understanding this relationship can allow educators to assess what works and what doesn't for different types of learners, opening a door to create more adaptive and personalized learning environments [48], and to foster self-regulation in learners [49].

A primary feature of these sorts of environments and thus, one of the goals of learning analytics as a discipline, is timely and helpful feedback. This has been a fixture of traditional learning environments and test-based assessment for some time [50], but is even more imperative in the faster-paced analytics-centric world.

As bright a horizon as learning analytics presents, there are some caveats. The use of business intelligence in organizations requires cultural adoption. If managers aren't trained and do not use the tools that analytics provide, their benefits will not be realized. The same is true of learning analytics in educational institutions. The existence and implementation of technology alone is not sufficient to create learner success. The systems have to be designed to be both functionally useful and have well-designed interfaces as a minimum. But more than that, instructors have to be promoters of the systems, there has to be a culture built around the analytics in order for them to be successful [51].

The technology for the implementation of learning analytics has a broad overlap with big data storage, retrieval, and inference techniques. Institutions must begin by developing a technology infrastructure to handle the amount of data that comes from learning traces, and in some cases, this has led to the invention of new protocols and techniques in distributed computing [52]. Once implemented, learning analytics systems are often informed by methods used in the industry to get insights into their data [53]. However, learning analytics demands that some techniques be adapted for the specific use. For example, particle swarm optimization can be used to cluster groups of students with similar learning styles to differentiate the type of content that engages them [54]. Returning to the discussion of competence from the previous section, learning analytics systems have been developed that provide numeric models of competence based on traces [55, 56]. The mechanisms to do so in the domain of interest for this thesis, writing, will be discussed in the next section.

Writing Analytics and NLP

Building from the definitions in the previous section, the goal of this research is to understand writers, and the environments in which they create compositions. The learning traces of interest, then, pertain to the measurable aspects of those compositions.

The first attempts to quantify aspects of language and automate language-specific tasks began as early as the seventeenth century, where philosophers like Descartes began a series of codes to translate words from one language to another. These works began the field of machine translation. Other than codifications, and advances in cryptography, nothing tangible was actually created until the 1930s, when some inventors began to create machines that featured bilingual dictionaries and grammar guides on microfilm [26].

In 1950, Alan Turing published his famous article, Computing Machinery and Intelligence [57] and therein defined the Turing Test for computer intelligence. If a person could have a conversation with a machine such that the machine's responses were indistinguishable from those of another human being, then the machine would be said to have passed the Turing Test. This idea that a machine could replicate the reactions, thought processes, and activities of human beings inspired an entirely new field: artificial intelligence. One of the overarching goals of natural language processing is to allow computers to 'understand' natural language and make use of it in the same way that humans do. In effect, the Turing Test is still the gold standard and the objective of NLP, and while a Turing-test-passing intelligence has not yet been constructed, technology is closer to producing one than it ever has been [58].

In 1957, Noam Chomsky published Syntactic Structures [59], which is deemed to be the most important early work on the analysis of grammars. The Cold War was underway during this time and the United States government was increasingly interested in machine translation, thus making funding available for research. Early results were promising. The Georgetown experiments saw the automated translation of 60 sentences from Russian to English [60]. A program called SHRDLU was developed to understand English instructions to manipulate objects in a limited domain called the 'blocks world', which was important to the field of robotics at the time [61]. However, due to computing limitations, it was impossible for computers to be truly aware of the context of language, or to solve a broader variety of problems reliably.

During the 1970s, several mathematical approaches to knowledge representation of natural language were developed. Researchers outside of the field of computer science began to look into ways to measure the quality or readability of a piece of writing algorithmically [62, 63]. Along with this, further analysis of linguistic structures and the idea of finite state automata for program design saw further development of machine translation.

In the 1980s, computing power became available more cheaply and abundantly than it had previously. The previously rule-based approaches to machine translations were superseded by statistical approaches, which used previously-tagged corpora in order to train learning algorithms to translate words and sentences [64]. These approaches are still mainstays of NLP techniques.

Natural language processing began to branch into a wide variety of tasks. One of the foremost was making computers 'understand' sentences by parsing and breaking them down into their constituent parts. Parsers usually took a piece of text as input and then produced a tree-like data structure from it that contained the base words as leaves, and larger noun- and verb-phrases as nodes. As part of parsing, the leaves could be identified as having a certain part of speech (POS). Large corpora, such as the Penn Treebank [65] were developed for the purposes of training algorithms to tag parts of speech accurately.

Basic part-of-speech tagging eventually evolved into semantic role labeling [66, 67]. Other problems in NLP include automatic summarization, named entity recognition, natural language generation, optical character recognition, relationship extraction, sentence tokenization, sentiment analysis, and speech recognition. These tools form the basis for a complete analysis of a given piece of writing.

Writing analysis begins with the basic data intrinsic to any composition. Writing in English is made up of a string of characters in the Roman alphabet. These characters are separated into certain structures to convey, organize, and transition between thoughts. Words are space-separated substrings. Each word represents a noun, verb, adjective, adverb, pronoun, preposition, conjunction, determiner, or some other part of speech present in the English language. Together, words are formed into sentences, which convey complete thoughts and usually contain a subject, a verb, and an object. Sentences are delimited by punctuation, such as periods, question marks, or exclamation marks. Sentences are further clustered into paragraphs. Paragraphs are often bundles of related ideas, delimited by new lines.

While obviously intrinsic to the English language, words, sentences, and paragraphs represent the most basic objects that writing analytics can make use of. Simple counts of the number of words in a document determine its length. Computing averages, such as average characters per word, or average words per sentence, can indicate complexity of a composition. In fact, one early measure of the readability of a document is based only on its total words, total sentences, and total syllables [63]. Simple equality tests for words can produce the number of unique words in a document as well; effectively an indicator of vocabulary usage. Combining the number of unique words with the total words gives a measure of lexical diversity; how broad the language in a composition is.

Adding a layer of external libraries allows further insight. Comparing each word in the document to a dictionary, a list of acceptable vocabulary for English and identifying which words are spelled correctly grants another metric: the total number of spelling errors in the document. Punctuation and capitalization can be tested for by examining characters within word tokens. It is expected that sentences end with a terminator, and that a single uppercase letter will follow terminating punctuation to begin a new sentence. However, simple tests for punctuation are insufficient, as they do not give any indication of whether a terminated sentence is actually complete, nor can they distinguish between sentence terminators and period-separated abbreviations (i.e. “N.L.P.”).

Grammar checking is often the work of rule-based systems, which are similar to dictionaries, except instead of single words they check for specific patterns in text. These patterns are defined by an exhaustive set of rules, defining exact words, character sequences, or regular expressions to check for within the document. These rules are often categorized to define of what sort of error a pattern match represents (typo, sentence fragment, etc.). Some systems also specify correct forms of the patterns matched by the rules, for easy access to possible corrections.

The above measures represent the most basic of processing in terms of computational costs: counts of characters, entities, and equalities, and rule matching. Their analytical power is limited in scope and correctness, so more advanced techniques are required to establish a more complete set of analytics over compositions.

Counts and equality metrics can be enhanced greatly if the various parts of speech of each word can be determined. While part of speech taggers can be built as rule-based systems, they are most often built on corpus linguistics. Large bodies of text can be tagged by hand with parts of speech for each word [65]. This serves as training data for various machine learning models. While the methodologies used differ widely, most such algorithms are centered around using a model that constructs probabilities for tags on words, usually given other lexical information in the text (like preceding and/or following words, or the phrase in which the word is situated). Bayesian inference is often the most common way to do this, using graphs of conditional probabilities, updated with the training data, to compute probability distributions for tags. Data structures like hidden or conditional Markov models are common in this purpose [68]. Stanford's CoreNLP collection uses a part of speech tagger that includes a conditional log-linear model to achieve 97.3% token accuracy on the Penn Treebank [69].

While POS tagging has its limits, obtaining counts of parts of speech for each word provides several additional statistical metrics that can be gleaned from the text. Proportions of personal pronouns to the size of the text, for instance, can serve as a measure of how personal or informal a composition is. A count of adverbs and adjectives can give a sense of modifier complexity, or a measure of function versus content words when compared to nouns or verbs. Going back to more basic metrics, it is now possible to categorize counts and look at average word size per part of speech, or the average density of form versus content words per sentence.

Natural language parsing breaks down a sentence structurally into its constituent parts. The most popular parsers rely on probabilistic context-free grammars (PCFGs), which use treebank training data to build a network of parse probabilities for each sentence [70]. These grammars are context-free in that they don't understand the contexts of language. For example, they cannot distinguish that the phrase 'eating chicken' is more probable in most cases than 'eating cats'. Some PCFGs are lexicalized, which means that the head words in each sentence are identified, giving some probabilistic weight to the words in each sentence. However this adds significant computational complexity [71].

Parsers add knowledge about the syntax of a sentence by splitting it into a tree structure of its noun- and verb-phrases. From these parse trees it is possible to determine noun-phrase complexity: how many noun-phrases there are in sentences on average. Additionally, looking at syntactic patterns enables counts of sentence types: simple, compound, and complex. Adding another layer of computation to parsers, it is also possible to resolve co-references within the text. That is, determine which phrases refer to the same subject or object [72, 73].

Sentiment analyzers determine the positivity or negativity of a sentence. Positivity affects the tone of a composition, and more positive writing is more likely to be well-received. Early sentiment analyzers assigned positive values to positive words, and negative values to negative words. However, because context is important in writing, many of these early tools did not correctly address the context of sentences, and how certain phrases can negate others. The best sentiment analyzers make use of parsing to break down sentence structures, and focus on how phrases modify others. These sentiment analysis tools are trained on sentiment treebanks, using neural tensor networks to learn the statistical functions to be used on new data [74, 75, 76].

Studies of individual words provide a basis to examine qualities within a piece of writing that would normally be considered subjective. Gilhooley and Logie's study rated 1,944 words for measures of age-of-acquisition, imagery, concreteness, familiarity, and ambiguity. The words were rated by students over the course of multiple sessions using a 7-point scale. Locating these words in the composition (using a parser or POS tagger to find their lemmas, or unconjugated roots), assigning these ratings, and then summing them across the document can give a sense of the amount of imagery or concreteness present in the writing, whether the writing is predominantly familiar or ambiguous, and what the reading level of the document might be [77]. Additional to word quality, knowledge bases have been constructed that provide information on the relationships between words. WordNet identifies 'synsets' or sets of cognitive synonyms, for each English word, defined by conceptual and lexical relations, and distinguishes between different forms and usages of words [78].

Advanced techniques like parsing, part-of-speech tagging, and sentiment analysis require training a machine learning algorithm on some large set of corpora, or using a hand-made database to associate meaning with language. Latent semantic analysis (LSA) eschews these methods, and still derives information about the relations and significance of words in a composition purely using mathematics. LSA involves creating a matrix whose rows are the content words in a document, and whose columns are context elements of some kind, usually sentences or paragraphs. The values in the matrix are the frequencies with which each word appears in each context. This matrix is then factorized into three matrices using singular value decomposition. One of these matrices is a diagonal matrix whose values can be changed to modify factorization coefficients, which allows the dimensional reconstruction of the original matrix in such a way that it gives relational measures of what the most important words are in the document, and how they correlate to one another within and across the contexts. LSA is an important technique for identifying topic flow within a composition [79, 80, 81].

By combining the output metrics from these varied tools, it is possible to begin painting a statistical picture of a composition that even encompasses elements that are not normally thought of as being numeric or computable. Concepts like imagery or connectivity or cohesion are considered to be subjective, but NLP tools in combination can provide estimates of these aspects of writing objectively [82].

Such tool combinations are used for automatic essay scoring, or AES. AES has roots going as far back as the 1960s, where even then it was possible to see the tremendous benefits of having computers assess student work. However, because of technological limitations at the time, the field laid dormant until the 1990s and the spread of microcomputers. Several AES systems began development once computing power was more readily available. These systems fall into several categories based on their approach to essay scoring. One category utilizes some or all of the tools mentioned above to define a number of metrics that approximate different aspects of the text. These measures are then taken for a large body of sample essays, which have been assigned grades by expert markers. The AES system is then trained on these essays, and that training data is used to evaluate other essays. Another category of AES makes use of latent semantic analysis to focus on the vocabulary used, and compare the results to a set of candidate criteria. There are also theoretical proposals to make use of a 'gold standard', to develop a statistical notion of what the best possible essay looks like, and compare a given essay to that standard. However, such an approach would require that the dimensions of gold standards are agreed upon, and that they are subdivided into appropriate classes for writing level and subject matter [83].

To date, AES has concerned itself only with grading completed essays. While automatic, these systems still follow traditional models of writing assessment, and were not designed to determine the details of the writing process and score them. Research is limited on how a student developed sentences or paragraphs over time, or how the flow of their composition improved or declined over time. This is where learning analytics systems, as discussed above, can assist. By analyzing traces of student compositions using the same metrics that AES systems use, it is possible to create a statistical trace of a student's path through a composition, and assess various competences from these metrics [84, 85].

Quantifying writing competence is an important endeavor from an efficiency standpoint in the case of AES, and from a developmental standpoint in the case of analytics from writing trace data. However, a large mass of raw data does not indicate anything useful about writing competences and the interaction between them. A statistical methodology is required to form conclusions from the observed patterns, but for pedagogical purposes, it is important to understand how the competence interactions work as a system of interventions. This is the realm of causal inference.

Causal Modeling

Causation is a constant of the human experience. The human brain understands causality intuitively and a moment's consideration can produce many examples of simple causal relations [86]. If Bob pushes on a chair, it falls over. Bob caused the chair to fall over. That is, if Bob had not pushed the chair, it would have remained upright. This seems straightforward. Intuitively, then, causation is an action (or intervention) that results in a change to something else. However, deeper consideration produces a number of questions on the nature of this causal claim. Did Bob really cause the chair to fall over? The push, the application of force, combined with the laws of gravity, caused the chair to fall. Was it a state of anger in Bob's brain that caused sufficient frustration to make Bob push over the chair? What were Bob's motivations? If Bob were pushing over the chair to demonstrate causation, could it be said that Bob's thesis caused the chair to tip over? If so, is it sound to make a universal causal claim that theses involving causation cause chairs to fall over?

These and other questions expose some of the complexities of causality. The human brain shortcuts through this complexity and focuses on causes and effects in simple terms. If A happens, B will happen. While useful for managing everyday occurrences, these claims aren't always valid. Many of them are based on personal beliefs and biases. It has thus been conjectured that causality does not exist at all, but is merely a human perception built on top of mathematical correlations [87]. However, there are some well-understood laws of nature and mathematics that are based in causality while not explicitly defining it.

For example, in classical physics, pushing the chair is an application of the physical law $m = Fa$. Intervening to apply force to a mass causes it to accelerate. Other branches of mathematics share this property of using causality without providing a definition: physics, geometry, formal logic, and probability theory.

Making scientific causal claims requires explicit knowledge of causality, and what it is. Intuitive or philosophical definitions will not suffice. However, the essence of causality does remain in the above definition: that taking an action upon a variable produces a change in another variable. For causality to be present, the effect of an intervention is different than if no intervention had occurred. Therefore, causality can be described as the difference between the actual results (what happened) and the counterfactual results (what didn't happen).

Unfortunately, without the use of time-travel technology, it is impossible to observe both the treatment effect and the counterfactual effect. The two are mutually exclusive. If one thing has already been done, then by nature, the other can't be done, and therefore both effects cannot be observed at the same time.

Classically, claims about causality require sampling from a population, and the performance of a treatment on some number of units in that sample. This is randomized experiment. Such experiments involve a treatment group and a control group. With a sufficiently large sample, it is assumed that these two groups collectively can represent the treatment and counterfactual effects on a population, giving the average causal effect [88].

Truly randomized experiments require that both variables are observable directly, that interventions to change one or the other are possible, and that there is a sufficiently large sample to correctly represent the counterfactual. In some cases, this might not be possible, plausible, or ethical. Often, it isn't possible to intervene, and sometimes phenomena can't even be directly observed. The problem, then, is how to establish causal claims in the absence of the ability to intervene, perhaps in the absence of the ability to observe, and how to establish when a causal relationship might have other unmeasured variables affecting it.

In cases where variables are observable, but cannot be intervened upon, statistics can provide answers to causal questions. Correlation is a simple statistical technique to compare whether two variables in a data set behave similarly. However, correlation does not imply causation. There can be any number of other effects on one or both variables that is causing the data to align. Further, while it is possible to see the 'goodness of fit' between two variables using correlation, it is impossible to determine causal direction, which variable causes which. Moreover, correlation is insufficient for causal claims because it doesn't answer the core causal question: the difference between actual and counterfactual data.

To overcome the weaknesses of correlation and other statistical techniques, and to apply causal interpretations to statistical problems, counterfactuals must be estimated. This is the concern of causal inference. Causal inference formalizes the way that causality works, makes assumptions about the nature of causality, and uses probability to translate statistical data into a causal structure.

Formal definitions for causal mechanisms are required before inferences can be made over data sets. These definitions are cited from Spirtes, Glymour, and Scheines' Causation, Prediction, and Search [89]. Causality is a relationship between two events: one event causes the other. An event may have multiple causes, no one of which can produce the event by itself. However, an event may also have different sets of causes that can produce it. Causation is transitive (if *A* is a cause of *B* and *B* is a cause of *C,* then *A* is also a cause of *C*), irreflexive (an event cannot cause itself), and antisymmetric (if *A* is the cause of *B*, then *B* is not the cause of *A*).

In a series of events, there may be direct causes and indirect causes. If *A* causes *B,* and *B* causes *C*, then we say that *A* directly caused *B* and *B* directly caused *C*. *B* becomes a causal mediary between *A* and *C*. In this way, direct causes have the effect of screening off indirect causes: once a direct cause occurs, indirect causes are not relevant.

In order to be treated as proper variables in a population, events have to be classified into types. When classified in this manner, variables can take on measured values depending on the resultant population. A variable *A* causes a variable *B* provided there is a set of values *a* for *A* and a set of values *b* for *B* and an event in which *A* taking on a causes *B* to take on the value *b*. When events are typified into variables, it becomes possible to define direct cause:

Variable *C* is a direct cause of variable A relative to a set of variables **V** provided (i) *C* is a member of a set **C** of variables included in **V**, (ii) there exists a set of values *c* for variables in **C** and a value *a* for *A* such that were the variables in **C** to take on values *c*, they would cause *A* to take on value *a* no matter what the values of other variables in **V**, and (iii) no proper subset of **C** satisfies (i) and (ii).

A formal definition of direct causes allows the further definition of other causal terms: common causes, causal chains, indirect causes, causal connection and causal structures. Each of these concepts is a configuration or extension of the notion of a direct cause. Exact definitions can be found in Chapter 3 (Section 3.2.2) of Causation, Prediction, and Search [89]. These definitions allow causal structures to be represented using directed acyclic graphs (DAGs).

A causal DAG consists of a number of nodes consisting of the variables in a structure. Two nodes *A* and *B* may have a directed edge between them if and only if *A* is a direct cause of *B*. Importantly, using a DAG does not give any notion of the weight or strength of common causes between variables. These values can only be determined from the underlying probability distribution of the causal system. Importantly, causal structures generate probability distributions even if the causal relationships are deterministic (i.e. the interventions upon the variables always have the same effects). Most causal systems, however, are pseudoindeterministic. That is, there is a system of variables that properly includes the members of a causal system that does represent a deterministic structure. Real-world causal structures are usually pseudoindeterminstic because unmeasured factors will usually exist (even if their effects are small) in a selected causal system. There are three axioms that govern the connection between causal structures and the probability distributions they generate. These axioms are fundamental to causal inference, and focus on the feature that is common to both causality and probability: independence of variables.

The Causal Markov Condition is an application of the Markov Condition in probability to causal structures. It is as follows:

Let *G* be a causal graph with vertex set **V** and *P* be a probability distribution over the vertices in **V** generated by the causal structure represented by *G*. *G* and *P* satisfy the Causal Markov Condition if and only if for every *W* in **V**, *W* is independent of $\mathbf{V} \setminus (\text{Descendants}(W) \cup \text{Parents}(W))$ given Parents(*W*).

Put more simply, every variable in a causal graph is independent of its non-descendants, conditional on its parents. The Causal Markov condition defines the sorts of independence relations necessary to connect causal graphs and generated distributions.

The Minimality condition further constrains the Markov condition by asserting that a causal graph cannot have any subgraphs that also satisfy the Markov condition: each causal graph must be the minimal representation of the Markov condition.

Let *G* be a causal graph with vertex set **V** and *P* a probability distribution on **V** generated by *G*. <*G*, *P*> satisfies the Causal Minimality condition if and only if for every proper subgraph *H* of *G* with vertex set **V**, the pair <*H*,*P*> does not satisfy the Causal Markov condition.

The Markov and Minimality conditions imply a set of independence relations for variables in a causal structure, and in a corresponding probability distribution. In some cases, the probability distribution can contain other independence relations that are not captured by the Markov condition as it applies to the causal graph. The assumption that this does not occur is called the Faithfulness condition.

Let *G* be a causal graph and *P* a probability distribution generated by *G*. <*G*, *P*> satisfies the Faithfulness Condition if and only if every conditional independence relation true in *P* is entailed by the Causal Markov Condition applied to *G*.

The Markov and Minimality conditions hold in nearly all practical cases of experimentation, as well as in the applied sciences. It is also implicit within other models of causal structure, including structural equation models. There are classes of problems in which care must be taken to properly define variables in the structure. For instance, when populations are mixed together, it is necessary to condition on an additional variable that separates the populations to ensure that the Markov Condition holds. There also exists a narrow class of problems pertaining to systems in quantum mechanics, in which the Markov condition does not hold.

Generally speaking, the faithfulness condition assumes that there are no exact coincidences in causal structures. It is, however, more likely to violated than the Markov condition. There is a subset of problems in which two variables are causes of a third, but they have a ‘cancelling’ effect on each other. They cannot be statistically correlated with the third variable. This is most common in backup mechanisms, where one variable ‘takes over’ a causal effect for another when it changes. The term for this type of violation is Simpson’s Paradox [90].

The axioms above will give a set of independence relations for a causal structure. However, there may be additional independence relations within a causal structure that are not represented by the Causal Markov condition alone. An additional tool to define these independence relations is a concept called d-separation [91].

If X and Y are distinct vertices in a directed graph G, and **W** a set of vertices in G not containing X or Y, then X and Y are d-separated given **W** in G if and only if there exists no undirected path U between X and Y, such that (i) every collider on U has a descendent in **W**, and (ii) no other vertex on U is in W.

Testing for d-separation is a fundamental part of causal inference. The procedure to check whether nodes X and Y are d-separated by **W** is as follows:

1. Check each undirected path between *X* and *Y*.
2. For each path, check each of its vertices to test whether it is active. A vertex *V* is active on a path relative to **W** if and only if *V* is a collider and it or any of its descendants are in W, or *V* is a non-collider and is not in **W**. A path is active if every vertex on the path is active.
3. If an active path between *X* and *Y* is found that is active in **W**, then *X* and *Y* are d-connected by **W**.
4. If all paths have been checked and are inactive, *X* and *Y* are d-separated by **W**.

With assumptions in place to ensure the reliability of causal claims using independence relations, a final piece of the puzzle is whether to assume that the variables present in the system of interest are the correct ones. As we mentioned, variable identification is important to ensure that a mixture of populations is not occurring. However, it is also important to identify causes at the correct level of granularity. It is possible to narrow down causal chains to the level of atomic movements, but this is rarely useful [92]. What matters predominantly is that all of the common causes of the variables of interest are identified.

The assumption that all of the common causes of the variables of interest are measured and included in the causal structure is called the causal sufficiency assumption. If causal sufficiency is not assumed, then an inference algorithm must allow for unmeasured latent causes within a structure.

A wide variety of causal inference algorithms have been developed for a variety of applications. Generally, the goal of these algorithms is to search for causal structure in a set of statistical data. The simplest of these algorithms are called pattern searches. Their goal is not to search for a single DAG representing the true causal structure for the data, but to search for patterns in the data that could produce several different DAGs. More specifically, these algorithms produce equivalence classes: a description of the set of graphs that represents the same set of independence relations that the data does, conforming to the causal Markov Condition and d-separation.

The independence relations within the data can be derived using a variety of established statistical tests [93, 94]. Once found, inference algorithms will apply the causal axioms along with d-separation searches to determine the equivalence class that fits those independence relations.

Equivalence classes have a graphical representation similar to DAGs. A directed arrow between *A* and *B* within an equivalence class pattern indicates that every DAG within the pattern contains that directed edge. An undirected edge indicates that in some graphs within the equivalence class, the arrowhead is into one variable, but into the other variable in some other graphs. The causal direction, in this case, cannot be determined and assuming or changing that direction has no effect on the fit of the structure to the data.

A number of possible DAGs can be represented by the patterns, and then tested for goodness of fit with the data. The set of possible DAGs can be constrained by background knowledge, which is a pre-imposed set of conditions that certain variables cannot or must cause others, or a constraint that different variables are different in time-order tiers, and thus can't be the causes of one another.

The archetypal pattern search algorithm is the PC algorithm which assumes causal sufficiency. It further assumes that all the data is either discrete or continuous, and in the continuous case, that variables are normally distributed and causal relationships are linear. PC works as follows:

1. Start with the complete undirected graph.

2. Remove adjacencies if they are independent given any set.
3. Collider test: look for adjacency triples of the form *X* – *Y* – *Z*, potential colliders. Check whether the set that separated *X* and *Z* contains *Y*. If it is not, then Y is a collider and the variables can be oriented *X* -> *Y* <- *Z*.
4. Away-From-Collider test: After colliders have been determined, look for any adjacency triples of the form *X* -> *Y* – *Z*. Because *Y* is not a collider, *Z* cannot cause *Y*. Thus, orient this triple *X* -> *Y* -> *Z*.
5. Repeat until there are no orientations.

Any edges that remain undirected in the structure produced by PC are uncertain. That is, there are no direct indicators of causal direction for the two adjacent variables.

There are a number of variants on the PC algorithm that apply differently to different circumstances and produce slightly different results.

Relaxing the assumption of causal sufficiency, the FCI algorithm works similarly to the PC algorithm. However, to account for latent common causes of variables, an addition to the DAG notation is needed. The partial ancestral graph (PAG) adds a circle symbol that can be placed at either end of an edge in the same way an arrowhead can be. The circle symbol indicates that a latent common cause may or may not be present. In the structure *X* o-> *Y*, the only conclusion is that *Y* is not an ancestor of *X*. In the case of an edge with two circles *X* o-o *Y*, there is not even a guaranteed adjacency: all it means is that no set d-separates *X* and *Y*. Double-headed arrows *X* <-> *Y* means there is a latent common cause of *X* and *Y*.

The FCI algorithm works as follows:

1. Start with complete undirected graph with all adjacencies set to o-o.
2. Remove adjacencies if they are independent given any set.

3. Collider test: look for adjacency triples of the form X – Y – Z, potential colliders. Check whether the set that separated X and Z contains Y. If it is not, then Y is a collider and the variables can be oriented X o-> Y <-o Z.
4. Away-From-Collider test: After colliders have been determined, look for any adjacency triples of the form X o-> Y o–o Z. Because Y is not a collider, Z cannot cause Y. Thus, orient this triple X o-> Y -> Z. The away-from-collider test shows that Y –> Z is unconfounded, as another variable would make a collider on Y.
5. Repeat until there are no orientations.

Once a pattern or a PAG has been produced from data, any single DAG that we select from the equivalence class will be equivalent in its causal structure. The causal structure itself has little use if not combined with statistical parameters. Parameterized models provide a framework to estimate causal strengths from the data.

In the case where all input data are discrete, a Bayes parameterized model can be used to probabilistically determine causal strengths based on priors from the data set. The joint distribution conforms to the Markov Condition that each variable is independent from its effects conditional upon its direct causes.

In the case where all variables are continuous, structural equation models (SEMs) can be used to estimate causal relationships in terms of a system of linear equations. There is an equation in a SEM for each variable in the model consisting of the value of its inputs (i.e. its direct causes), plus some error term. For example, the causal structure X -> Y <- Z can be represented by the following structural equation model:

$$X = \varepsilon_X$$

$$Z = \varepsilon_Z$$

$$Y = \beta_1 X + \beta_2 Y + \varepsilon_Y$$

These models are defined from the causal structure determined from an equivalence class and instantiated from the data. In a SEM, a probability distribution over the exogenous error terms will induce a probability distribution over the rest of the model. As a result, the computation of SEMs from statistical data begins by computing the mean and standard deviation of the values of the exogenous variables to acquire the initial terms. From this, the parameters of the model (specifically, the coefficients of the beta terms) can be estimated. Because of this, it is important that the exogenous error terms are pairwise independent, meaning that all off-diagonal terms in the covariance matrix representing the exogenous error terms is 0. The error terms are assumed to be normally distributed for SEMs, as they are effectively an aggregation of all of the noise affecting the variables.

Normally, SEMs are assumed to be linear equations with normally-distributed error terms. It is possible to relax these assumptions by creating generalized SEMs, which may have any equation specifications. However, it is difficult to learn these equations from data.

With a causal structure in the form of a parameterization like a SEM, and with experimental data, it is possible to estimate the statistical strengths of the relationships between the variables. Estimators make use of the correlation and covariance between variables to establish values for the parameters in a SEM. In the case of linear SEMs with normally-distributed error terms and no cycles, multiple linear regression can be used to estimate coefficients and residual variances. Otherwise, maximum likelihood functions are used to minimize the distance between the covariance of the variables produced by the model's error terms and coefficients, and the sample covariance matrix [95]. Normal tests for goodness-of-fit, namely the chi-squared value and the p-value for rejecting the null hypothesis, vary quite widely with sample size. While these are used as indicators of SEM model fit in situations with small sample sizes, other measures have been constructed based on the chi-squared and degrees of freedom that compensate for sample size. These include the Bayesian information criterion (BIC), comparative fit index (CFI), and root mean square error of approximation (RMSEA) [96, 97, 98].

Probability, conditional independence, causal axioms, and structural equation models represent a suite of statistical tools that can be used to estimate systems of interventions from observed data, and even provide some hints as to additional latent structures not present in that data. For this research, the data at hand comes from detailed writing trace data, analytical metrics over that data, and computed values of competence at each stage in the trace. Following is a methodology for combining these elements.

Chapter 3 - METHODOLOGY

The problem that this research addresses is in the nature of writing competence, and how different aspects of that competence relate to one another. The previous chapter described the vagaries of the act of writing, the nature of competence, the development and application of analytics (in general and as applied to natural language) and the statistical techniques involved in estimating causal systems from data. In this chapter, these concepts coalesce into a methodology for estimating writing trace data, analyzing it for quantifiable information, and applying causal inference to produce a number of models that will answer the research question.

WriteSim

The simulation and analysis of writing data begins with WriteSim, a program written in Java that takes completed papers from corpora (or any other source), and outputs a series of timestamped records for each document that represent a trace of the word-level transactions that might occur if a student had composed it.

The foundation of WriteSim is a research paper by Luuk Van Waes and Peter Jan Schellens [99], in which students were asked to complete writing tasks. Their keystrokes were individually logged and placed into long streams, which were then painstakingly analyzed to determine a number of writing behaviors. The study examined how, when, and why students paused during writing, and when and why they revised errors in their writing. For each composition, and each student, the number and length of pauses were considered, the location of the pauses within the linguistic area, and whether the pauses were due to formulation (the writer considering what to write next) or revision (correcting existing errors in the document). Pauses and revisions were broken up by the point in the document at which they were completed: whether mid-sentence, mid-paragraph, or at the end of the main composition. The level or linguistic structure of each revision was determined: word, sentence, or paragraph. Additionally, revisions could be mechanical or structural (correcting mechanics like spelling and punctuation, versus reorganizing the document for better topic flow). The statistical results of these parameters across the 40 students in the study were clustered to provide a number writing ‘profiles’:

Table 3-1: Writing Profiles

Profile Name	**Description**	**Statistical Averages** [99]
Initial Planner	Initial planners do most of their thinking up front. They will have very few revisions in the final stages of their writing process, mostly spending large amounts of time planning the perfect composition before they even begin writing it.	Stage I Revisions: 66.8 Stage II Revisions: 16.0 Initial Planning Time: 1310.0 s Average Pauses: 229.0 Average Pause Time: 21.0 s Recursion: 62.0
Average Writer	Average Writers tend to have an even amount of up-front and look-back revisions and average lengths and numbers of pauses. They represent a mean of writers, and are the largest group of writers.	Stage I Revisions: 88.7 Stage II Revisions: 28.0 Initial Planning Time: 332.0 s Average Pauses: 283.0 Average Pause Time: 12.0 s Recursion: 79.0
Fragmentary Stage I Writer	Fragmentary Stage I Writers spend less time up front than initial planners, but still tend to do all of their work in the first stage of the writing process. They will do a lot of short pausing and make a lot of revisions, but most of these actions will occur before they finish their first drafts.	Stage I Revisions: 137.5 Stage II Revisions: 17.0 Initial Planning Time: 235.0 s Average Pauses: 364.0 Average Pause Time: 12.0 s Recursion: 105.0
Stage II Writer	Stage II Writers make most of their revisions after they have quickly composed an initial draft. They will make a lot of revisions, specifically to sentences and paragraphs, rather than to individual words. They spent a reasonable amount of time on initial planning, but once they start, they didn’t pause often.	Stage I Revisions: 45.3 Stage II Revisions: 88.0 Initial Planning Time: 824.0 s Average Pauses: 201.0 Average Pause Time: 20.0 s Recursion: 33.0
Non-Stop Writer	Non-Stop Writers focus on moving quickly. They spend little time in initial planning, and pause and revise infrequently. They have a lower composition time than all other profiles, and tend not to revise much even at the end of the writing process.	Stage I Revisions: 45.6 Stage II Revisions: 10.0 Initial Planning Time: 319.0 s Average Pauses: 173.0 Average Pause Time: 15.0 s Recursion: 42.0

The parameters used in the experiment are the same parameters used as inputs for the simulator in its attempt to mimic the writing process. The simulator begins by generating the attributes of simulated 'students' using the results of the Waes & Schellens paper as a basis. The profile object within the program is responsible for generating student objects for each composition. Common to all profiles is the basis, which contains overall averages for the following statistics from the study:

Table 3-2: Profile Basis

Metric	**Categories**	**Values**
Average Number of Pauses	Pause Type	Formulation (171.5) Revision (142.2)
	Linguistic Location	Within sentence (245.9) At a sentence boundary (41.9) At a paragraph boundary (25.9)
Average Pause Time (s)	Pause Type	Formulation (12.48) Revision (11.36)
	Linguistic Location	Within sentence (9.59) At a sentence boundary (16.19) At a paragraph boundary (27.78)
Average Number of Revisions	By Level	Letter (14.1) Word (46.6) Phrase (36.7) Sentence (6.3) Paragraph (2.5) Layout (10.5) Punctuation (6.7)
	By Purpose	Content (94.7) Form (28.7)
	By Remoteness	Within Line (75.0) 2-10 Lines (36.3) 11-24 lines (5.3) 25+ lines (6.8)
Average Length of Compositions	Words	843 words
Average Number of Sentences	Sentences	33.9 sentences
Average Number of Paragraphs	Paragraphs	7.7 paragraphs
Keystroking Speed		163.32 ms per keystroke 73.66 ms standard deviation

All of the profiles have the same basis as per Table 3-2, but vary in the statistics presented in Table 3-1. The profile objects combine the basis with their unique values to produce student objects. Each student object is unique. That is, every student will have slightly different values for each of the statistics in Table 3-1. Per-student values are generated by using the average for the profile and assuming that the sample corresponds to a normal distribution. No standard deviations were given as part of the profiles in the study, so a standard deviation is assumed for each statistic that is proportionate to the size of its mean. As noted in the profile basis, all of the numbers are based on a composition of a certain length: 843 words. Because the ‘students’ will be simulating documents of different lengths, their final values for number of pauses of various types will be scaled linearly according to the document length. This is one of the fundamental assumptions of the system.

A completed composition is input into the simulator along with a generated student. The simulator will then output a number of timestamped records representing the composing process. The simulator works in the following steps:

1. Tokenize the entire document.
2. Loop through each of the individual tokens.
3. For each token, compute probabilities that the student will pause or deviate from the main text. These probabilities are based on the ratio of the number of pauses of a certain type the student is expected to make at certain locations to the length of the document. Test against the probabilities to determine whether an event occurs at this token.
    a. If a student is to make a deviation at this token, determine randomly whether it is to be a spelling error or a phrasal change.
        i. If a spelling error, change the characters in the current token randomly.
        ii. If a phrasal change, randomly pick a synonym from a WordNet synset for that token.
        iii. Add keystroking time for the replaced token.
        iv. Add the deviation to a collection of deviations.

   b. If a student is to pause at this token, determine the probability of the pause being for formulation or revision purposes.
      i. If a revision pause is generated, check if there are any currently outstanding deviations.
      ii. If so, return a random deviation to its original state.
      iii. Add a random time for the pause based on the profile, and then add keystroking timing equal to the number of characters in the original token.
      iv. If a revision pause is generated but there are no outstanding deviations, continue.
      v. If a formulation pause is generated, add time for the pause based on the profile.
   c. If there are no deviations or pauses, add keystroking time for the original token.
4. Create a record at each iteration of step 3 that contains the full concatenated text of all of the tokens processed so far, including deviations, and the timestamp generated by adding time for pausing and keystroking.
5. Once each token has been examined, if there are any outstanding deviations, correct them, adding the requisite pause and keystroke time for each correction as in 3b above. This constitutes the Stage II revisions.

As a note, keystroking time mentioned above is independent of profile and included in the profile basis. It is assumed that each word takes a certain amount of time to type depending on the subject’s typing speed, randomization from finger movements, and distance of the keys from each other. These items are generalized into an average keystroke speed with a standard deviation that gives an accurate representation of typing speed based on studies on the subject [100].

Importantly, the simulator does not do large-scale structural reconfigurations or order changes of phrases, sentences, or paragraphs. This limitation will have the effect of making the topic flow information smoother than it would be if real-world data were used. As students move and manipulate phrases, it is likely that metrics related to topic flow would tend to spike into local peaks and valleys more often throughout the process. Structure changes could be integrated into future iterations to make the simulator more comprehensive.

To tokenize the document, determine linguistic locations for pause calculations, and to determine parts-of-speech for tokens the OpenNLP suite is used in WriteSim as a lightweight natural language processing engine. WordNet synsets, which define interrelations and connections between words, were used to determine synonymic words or phrases as part of deviation generation [78].

Upon completing a simulation, WriteSim will have generated records that approximate word-level edits to the data. These records can be exported to a number of formats, including JSON. Specific to this research, a JSON exporter class is included to transfer the simulated data to files for processing by the SCALE competence analysis engine.

Smart Competence Analytics for Learning (SCALE)

The Smart Competence Analytics for Learning (SCALE) engine is a server-based information system designed to analyze learning traces and automatically produce measures of competence for students. SCALE operates using three levels of ontological sorting and three layers of processing. SCALE first parses learning trace data for a particular domain, extracting descriptive quantitative information from the traces. Secondly, SCALE uses the information from the parsing phase to construct a competence assessment using pattern-matching rules for the domain of interest. Finally, the competence information is combined into student profiles to provide them with feedback [56].

SCALE handles the domain of interest at its parsing layer by making use of a variety of natural language processing tools. These are used to break down each writing event in the learning trace and translate it to an ontological structure. SCALE's interaction ontology for composition includes words, phrases, sentences, and paragraphs in a tree structure. Further, 65 metrics are computed for each event based on the generated ontological structure. These metrics are used alongside production rules to determine values between 1 and 100 for competence in six different areas (see Table 3-3 and Table 3-4) [84, 70, 74, 80, 68, 78].

Table 3-3: SCALE Writing Metrics

Category (# of Metrics)	**Description**
Structure Counts (4)	Counts of the number of characters, words, sentences, and unique words.
Specific Word Counts (4)	Counts of the number of words longer than 5, 6, 7, or 8 characters.
Part-of Speech Count (14)	Counts of the number of pronouns, verbs, adjectives, connectors, tone words, content words, function words, prepositions, adverbs, determiners, conjunctions, articles, nouns, and unidentified part-of-speech words.
Word Qualities (17)	Counts of the number of words with specific qualities: positive, negative, space/location/place, afraid/annoyed, opposition/limitation/contradiction, examples/support/emphasis, happy/glad/joyful, famous/popular, mischievous/crazy, agreement/addition/similarity, sad/terrible, time/chronology/sequence, cause/condition/purpose, amazing/beautiful, conclusion/summary/restatement, effect/result/consequence, evil/mean.
Error Counts (2)	Number of grammatical and spelling errors in the event.
Ratios and Averages (8)	Average characters per word; the average words per sentence; ratios of content and function words to the total words in the event; ratios of misspellings and grammatical errors to the total words in the event; the ratio of the number of unique words in the events to the total words in the event (lexical diversity); and readability score [14].
Indexes (4)	Indexes of connectivity and topic flow, and the average connectivity and topic flow.
Composition Time (1)	A measure of the total time spent on the document, as of the given event.
Rubric and Essay Scores (5)	A compilation of several metrics to determine scores for four different rubric categories. These rubric categories combine to give an estimate of the grade of the essay.
Writing Competences (6)	A compilation of several metrics to determine the level of competence of the student in a particular area, as of the time of a particular event. Competence values range from 0 to 100. The competences measured are vocabulary use, vocabulary complexity, spelling, grammatical accuracy, transition, and topic flow. See Table 3-4.

Table 3-4: Competences

Competence	Description
Grammatical Accuracy	Grammatical accuracy measures the student's competence in writing correct grammatical structures. It is based on the ratio of instances of incorrect grammar (based on grammar rule violations) to the length of the document.
Spelling	Spelling measures the student's competence in spelling correctly. It is based on the ratio of misspelled words to correctly-spelled words.
Topic Flow	Topic flow is a measure of student competency in relating ideas to one another throughout the text. Topic flow is specifically based on the number of sentence-adjacent content words. Semantic distances are calculated for content words, and the competence is computed by comparing the content words to an ideal standard for topic flow.
Transition	Transition is similar to topic flow, but measures the overall connectedness of the student's composition. Transition competence is calculated by analyzing the number and distribution of connective words in the document.
Vocabulary Complexity	Vocabulary complexity is a formulaic manipulation of the Flesch-Kincaid readability index [62, 63] to measure the student's competence in creating a readable composition.
Vocabulary Usage	Vocabulary usage is a measure of the student's competence in using different kinds of words. It is based on the ratio of unique words to total words in the document (using word lemmas).

TETRAD V

TETRAD V is an application developed by researchers at Carnegie-Melon University that implements the algorithms developed and explained in Causation, Prediction, and Search [89]. These algorithms involve searching for causal interpretations from observed data, and a suite of tools for testing those interpretations. For a given data set, it is possible to derive several different possible causal models by enforcing different assumptions, constraints, and parameters on the data. Chapter 2 of this thesis discusses the background and major implications of this causal approach, and details how some of the algorithms work.

Data Flow

The purpose for using a simulator for writing data is because real-world human data is time-consuming to acquire and in distributed learning environments, especially in the fully-determined JSON format that SCALE requires. As a result, the data is simulated to begin the foray into this research on which factors and competences of writing relate to one another in a system of interventions.

WriteSim was used to simulate writing event records for the entire British Academic Written English (BAWE) corpus; a total of 2,761 essays [101]. These records, exported to JSON, produced 12GB of data. The essays were organized into batches by file size to be processed by SCALE.

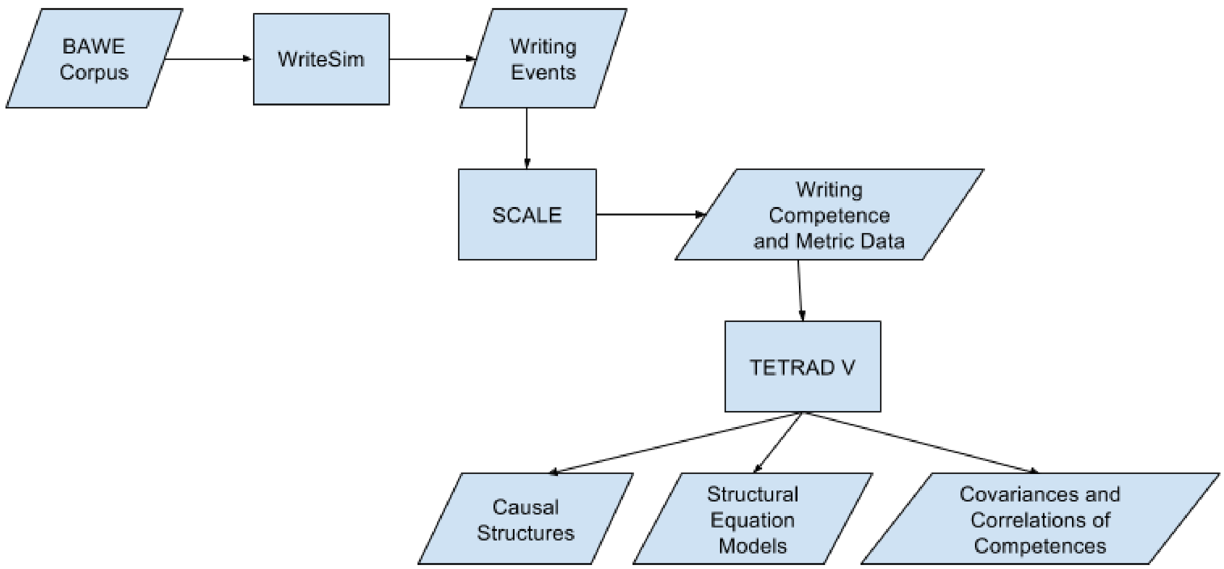
Over the course of three months, SCALE processed the first batch, a total of 390 essays. These essays were each in separate files and the JSON output contained the identifier for the original essay, for reference. Collectively, the essays represent 744,848 events. As mentioned above, the records contain the text of each edit, up to the point of 'completion' of the document. Thus, near the end of the composition, many events are hundreds of words long. Processing this batch of 390 essays took SCALE three months.

The essays were 1,825 words long on average, with a standard deviation of 141 words. The simulated students took an average of 2.51 hours to compose each essay, with a standard deviation of 44.86 minutes.

SCALE output the results in JSON form. For each event, the data consisted of values for the 53 metrics from SCALE, plus values for the six competences in Table 3-4, five rubric score categories, and a value for composition time; a total of 65 variables. This data was then sent to

TETRAD V where causal algorithms were then run on various sets and subsets of the data. Figure 3-1 shows an overview of the data flow between systems.

Figure 3-1: Data Flow



Causal Analysis

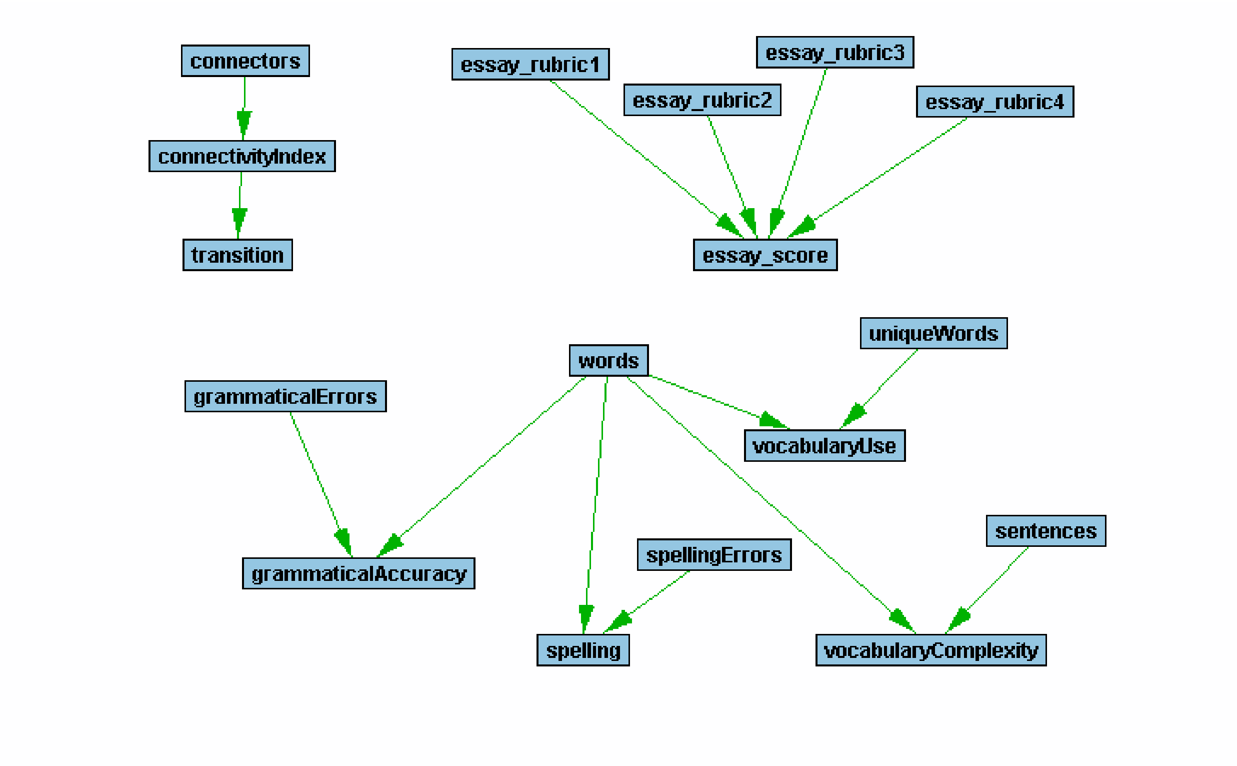
Prior to running causal searches, two sets of constraints were placed on the data set. The first was to trim twelve variables from the data which were direct linear determinants of one another. These variables caused the covariance matrices of any models run on the data to become singular, and so had to be removed to obtain any estimates of causal strength and model fit. The variables removed from the final causal models are shown in Table 3-5.

Table 3-5: Variables Excluded from the Models

Variable
Average Number of Characters per Word
Average Number of Words per Sentence
Content Word Ratio
Function Word Ratio
Topic Flow Index
Average Connectivity Index
Unknown Words
Grammar Error Ratio
Spelling Error Ratio
Average Topic Flow Index
Readability Score
Lexical Diversity

Secondly, the data was constrained on a set of background knowledge for the competences. The SCALE system uses particular sets of variables to determine competences, and those that were known to represent causal relationships were directly specified prior to running models in TETRAD. Figure 3-1 shows the relationships required within the models.

Figure 3-2: Background Knowledge for Causal Structure



The data was segmented into six different sets for causal analysis: all combinations of two horizontal subsets (modifying the number of records) and three vertical subsets (modifying the number of variables). The first horizontal subset was the full event data, with all 744,848 events included. For comparison, and to examine the data in a more simplistic fashion, the final records of each of the 390 simulated students formed the second horizontal segment. The simple case of the 390 final records will be examined first before moving on to the full data set.

The first vertical set contained all of the measured variables (less the twelve above); as complete a model as could be determined from the data. The second set contained only the six competences plus the essay score variable, to measure the interactions between them in isolation, and determine any causal effects between them in the absence of the other metrics. Finally, the competences plus their directly-adjacent variables from the results of the first sets were isolated to produce a less-dense model that could identify some of the latent paths between competences identified from models over the second variable set.

Table 3-6 summarizes the six sets of variables and the searches that were conducted on each. In the cases of the data subsets using only the final versions of each essay, the Fast Greedy Equivalence Search (FGS) algorithm was used in place of the PC algorithm as a search including causal sufficiency. As a scoring algorithm, FGS has more reliable results on smaller sample sizes [102]. For the larger sets of records, the PCPattern search algorithm was used, which correctly assumes causal sufficiency and only outputs patterns with no bidirected edges. In all cases, a DAG was selected from each causal sufficiency search algorithm result to serve as the basis for a structural equation model. Each such model was then estimated to the data to provide goodness-of-fit measures.

Table 3-6: Data Subsets for Causal Evaluation

Data Included	Variables Included	Description	Causal Searches
Final record of each simulated student (390 events).	All available.	All variables examined on a small sample size: the analysis expected from an automatic essay scoring system for traditional settings.	FGS search assuming causal sufficiency to obtain causal equivalence class. SEM constructed from a DAG within the FGS pattern. Estimation of SEM to determine fit to data. FCI search to examine potential latent common causes.
Final record of each simulated student (390 events).	Competences and essay score only.	Competence variables examined on a small sample size to use for comparison on the denser models.	FGS search assuming causal sufficiency to obtain causal equivalence class. SEM constructed from a DAG within the FGS pattern. Estimation of SEM to determine fit to data. FCI search to examine potential latent common causes.
Final record of each simulated student (390 events).	Competences, essay score, and first-order adjacencies as identified in row 1.	Expanded version of the competences set to include some clues about latent variables between competences.	FGS search assuming causal sufficiency to obtain causal equivalence class. SEM constructed from a DAG within the FGS pattern. Estimation of SEM to determine fit to data. FCI search to examine potential latent common causes.
All event records (744,848 events).	All available.	The fully-determined model, expected to be dense.	PCPattern search with $\alpha = 0.001$ to obtain causal equivalence class assuming causal sufficiency. SEM constructed from a DAG within the PC pattern. Estimation of SEM to determine fit to data. FCI search to examine potential latent common causes. Regression tables for each competency and its direct adjacencies.
All event records (744,848 events).	Competences and essay score only.	Competence variables examined in isolation to determine causal effects.	PCPattern search with $\alpha = 0.001$ to obtain causal equivalence class assuming causal sufficiency. SEM constructed from a DAG within the PC pattern. Estimation of SEM to determine fit to data. FCI search to examine potential latent common causes.
All event records (744,848 events).	Competences, essay score, and first-order adjacencies as identified in row 1.	Expanded version of the competences set to include some clues about latent variables between competences.	PCPattern search with $\alpha = 0.001$ to obtain causal equivalence class assuming causal sufficiency. SEM constructed from a DAG within the PC pattern. Estimation of SEM to determine fit to data. FCI search to examine potential latent common causes.

The results of the causal analysis above are presented in Chapter 4.

Chapter 4 - RESULTS

The question posed in Chapter 1 finds part of its answer here. A reasonable facsimile of writing trace information was generated for a number of essays, producing a large quantity of events which were then analyzed. The analysis and production of metrics and competences was the longest part of the research. SCALE took nearly three months to process the 744,848 events, and generated 12GB of data from them.

Causal inference was applied to the resulting analytical metrics and competence calculations, in a variety of configurations. The resulting causal models, and their relevant features, are described below.

These results will be interpreted bearing in mind the assumptions of causal inference, namely that the axioms hold and that the interactions of the metrics and competences can be described as linear systems of one another. Proper goodness-of-fit statistics for the structural equation models are necessary to produce any valid claims about the results.

It is expected that the causal models will provide a framework by which educators can determine how to improve various competences by influencing others. These general principles of competence interaction can be used alongside manual assessments to understand the key areas in which a struggling student may need to improve.

This chapter is divided into sections to organize the results of the causal searches and structural equation model estimations similarly to the data segmentation discussed in Chapter 3.

Each section deals with a subset of variables, and each addresses both cases of horizontal segmentation. That is, each section explores the models based on both the 390 final essay values, and the full 744,848 events. Thus there are two sets of models per section.

The first section contains details of the interactions between competences only, with the subset of seven variables including essay score. The second section addresses the full data set containing all 53 measured variables. The third section includes the competence variables plus their immediate adjacencies as identified in the fully-qualified model. Within each section, exogenous and endogenous variables will be listed along with the number of nodes and edges in the causal graph. The causal structures will be presented visually when it is useful to include them. The models which assume causal sufficiency will be compared to the models which allow for latent common causes. Finally, the covariance and correlation matrices for each structural equation model will be presented. As some matrices will be very large, some of these will be deferred to the Appendix A, which is a narrow-margin landscape section of the document that fits the tables more reliably, to ensure that the main text does not become cluttered.

A final section will describe the model fit of each structural equation model to its data, and the most significant models will be determined.

Discussions about the conceptual implications of these models will be deferred to Chapter 5.

For the purposes of simplicity, the metric and competence variable names have been abbreviated for presentation in the correlation and covariance matrices presented below and in Appendix A. These abbreviations are shown in Table 4-1.

Table 4-1: Variable Abbreviations

Variable	Abbr.	Variable	Abbr.	Variable	Abbr.
compositionTime	CT	essay_rubric2	ER2	prepositions	Pre
adjectives	Adj	essay_rubric3	ER3	pronouns	Pro
adverbs	Adv	essay_rubric4	ER4	sadTerrible	ST
afraidAnnoyed	AA	essay_score	ES	sentences	Sent
agreementAddition Similarity	AAS	evilMean	EM	spaceLocation Place	SLP
amazingBeautiful	AB	examplesSupportEmphasis	ESE	spellingErrors	SE
articles	Art	famousPopular	FP	Time Chronology Sequence	TCS
causeConditionPurpos e	CCP	functionWords	FW	toneWords	TW
characters	C	grammaticalErrors	GE	uniqueWords	UW
conclusionSummary Restatement	CSR	happyGladJoyful	HGJ	verbs	V
conjunctions	Conj	mischievousCrazy	MC	words	W
connectivityIndex	CI	negativeWords	NW	Grammatical Accuracy	GA
connectors	Conn	nouns	N	spelling	S
contentWords	CW	numberOfWordsLonger Than5Characters	WL5	topicFlow	TF
determiners	D	numberOfWordsLonger Than6Characters	WL6	transition	T
effectResult Consequence	ERC	numberOfWordsLonger Than7Characters	WL7	Vocabulary Complexity	VC
essay_rubric1	ER1	numberOfWordsLonger Than8Characters	WL8	vocabularyUs e	VU
oppositionLimitation Contradiction	OLC	positiveWords	PW		

Causal Interactions of Competences

The first set of variables under consideration contains the six competence variables, plus the essay score variable. This is the minimal determination of the causal structure of writing competence from this research, and provides a concise high-level understanding of how the competences are related.

Figure 4-1 shows the structural equation model resulting from a fast greedy equivalence search (FGS) over the data. Vocabulary use and grammatical accuracy are the exogenous variables, meaning that there is no other competence that influences them directly. Essay score and vocabulary complexity are endogenous, meaning that they are only effects of other competences. There is a single undirected edge in the pattern, which indicates that either spelling or topic flow could also be exogenous variables. There are 8 directed edges in the pattern. Table 4-2 and 4-3 show the covariance and correlation matrices for this model.

The model shows that intervening to increase vocabulary use has a directly positive effect on essay score, and a very small, but negative effect on transition. Grammatical accuracy has a large effect on essay score, and a moderate effect on vocabulary complexity, both positive. Essay score is fully determined by other competences, which matches intuitive expectations of what essay score is. A less-intuitive result is that vocabulary complexity is also only an effect. It doesn't cause any other competences to increase or decrease. The model indicates then that intervening to increase vocabulary complexity would have no net effect on essay score overall, which goes against intuitive expectations.

Figure 4-1: SEM from FGS Search over Competences – Final Essays

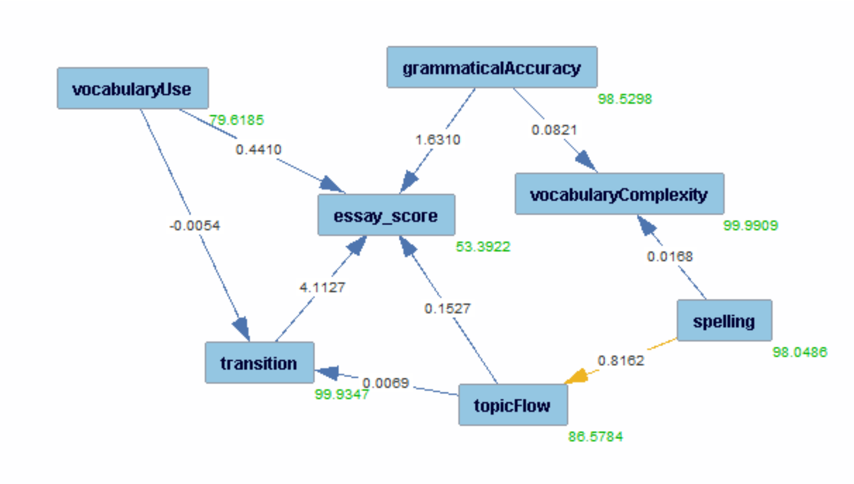


Table 4-2: Covariance Matrix for FGS Search over Competences – Final Essays

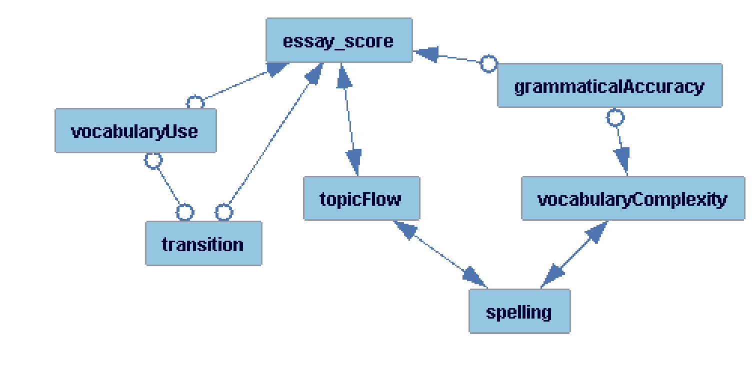
	ES	GA	S	TF	T	VC	VU
ES	28.3590						
GA	0.8740	0.5359					
S	0.2041	0.0000	1.3819				
TF	9.0343	0.0000	1.1278	49.9283			
T	0.2798	0.0000	0.0077	0.3423	0.0939		
VC	0.0752	0.0440	0.0232	0.0189	0.0001	0.0252	
VU	27.8313	0.0000	0.0000	0.0000	-0.3598	0.0000	66.4726

Table 4-3: Correlation Matrix for FGS Search over Competences – Final Essays

	ES	GA	S	TF	T	VC	VU
ES	1.0000						
GA	0.2242	1.0000					
S	0.0326	0.0000	1.0000				
TF	0.2401	0.0000	0.1358	1.0000			
T	0.1715	0.0000	0.0215	0.1581	1.0000		
VC	0.0890	0.3788	0.1242	0.0169	0.0027	1.0000	
VU	0.6410	0.0000	0.0000	0.0000	-0.1440	0.0000	1.0000

An FCI search over the competence variables (figure 4-2) indicates that, as expected, the relationships between competences are not direct in reality. Three bidirected edges in the FCI search indicate latent common causes between essay score and topic flow, topic flow and spelling, and spelling and vocabulary complexity. There is complete uncertainty in causal direction for vocabulary use and transition. All of the other variables that are into essay score potentially have latent systems between them.

Figure 4-2: FCI Search over Competences – Final Essays



Moving to the case where all of the events are used in the causal analysis provides a picture of the developmental nature of the writing process. Where the previous set of variables only examined competences the way a traditional assessment would, the full event set will provide insight into the interaction of competences within the writing process itself.

An FGS search over all of the events reveals a larger number of edges between variables, and therefore a larger number of interactions between competence variables. Figure 4-2 shows 16 total edges, 12 directed and 4 undirected.

The exogeneity of variables in this structure is uncertain because of the undirected edges in the pattern. Potentially, grammatical accuracy, vocabulary complexity, spelling, and vocabulary use could all be independent, acting only as causes of other competences. Essay score remains fully determined by other competences, which is consistent with our earlier result and strengthens the expectation that it is only a result of other competences. The transition variable has a variety of causes in this developmental structure. Interestingly, vocabulary use has a large negative effect on topic flow, and a smaller negative effect on transition, indicating that too much lexical diversity throughout the writing process could be a detriment to the composition's connectivity and flow. Many of the causal directions between competences are flipped in this data set, such as that between vocabulary complexity and spelling, indicating that competences develop in different directions during the writing process.

Notably, the correlation matrix in Table 4-5 is more completely defined. Moreover, adding the complete data set changes correlations and causal strengths between variables significantly.

Figure 4-3: SEM from FGS Search over Competences – All Events

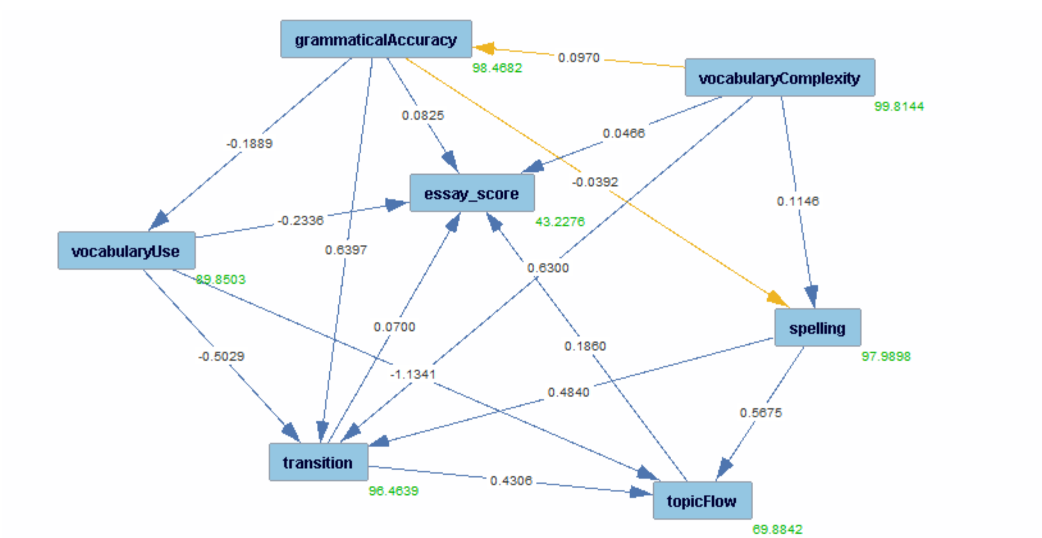


Table 4-4: Covariance Matrix for PC Search over Competences – All Events

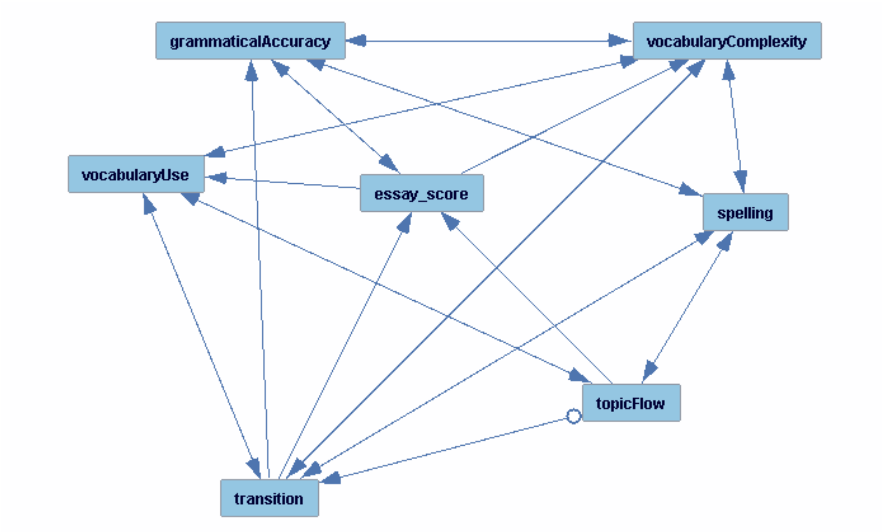
	ES	GA	S	TF	T	VC	VU
ES	69.8318						
GA	0.7220	2.0340					
S	0.7206	0.0431	3.0058				
TF	149.9804	1.4035	2.6874	639.6864			
T	52.5394	2.1906	2.2583	144.9962	246.9221		
VC	2.0751	1.0720	1.2250	4.5174	8.3431	11.0518	
VU	-33.8052	-0.3843	-0.0081	-87.6983	-32.9687	-0.2025	64.8053

Table 4-5: Correlation Matrix for PC Search over Competences – All Events

	ES	GA	S	TF	T	VC	VU
ES	1.0000						
GA	0.0606	1.0000					
S	0.0497	0.0174	1.0000				
TF	0.7096	0.0389	0.0613	1.0000			
T	0.4001	0.0977	0.0829	0.3648	1.0000		
VC	0.0747	0.2261	0.2125	0.0537	0.1597	1.0000	
VU	-0.5025	-0.0335	-0.0006	-0.4307	-0.2606	-0.0076	1.0000

The FCI search over the full event data gives stronger implications of the presence of latent common causes. Again, this is expected. There are 5 directed edges in the structure, and 10 bidirected edges.

Figure 4-4: FCI Search over Competences – All Events



The degree of uncertainty in the FCI searches informs on the true nature of the interactions between competences, that they are systems of many other metrics. The next two sections will examine larger systems of variables in detail.

Causal Interactions of All Data

The full data set contains 53 variables. Because of the large number of interconnections, it is difficult for an image to describe the causal structure reliably. This is particularly true in the case of the structural equation models, which contain error terms and coefficient values throughout. Instead, images of the causal models will be included for reference. Further, because of their size, the covariance and correlation matrices for the SEMs have been relegated to Appendix A. These will be referenced here as they are discussed.

Importantly, because of the large number of connections between variables, there will be a large number of latent systems between any two competence variables. Because of this, the discussion will focus primarily on first- or second-order adjacencies between variables.

Beginning with the simple case using only the final essay results, there are 165 edges in the causal structure. Of these, only 1 is undirected in the FGS search (figure 4-5). The exogenous variables in this case include some of the word-quality metrics (sad/terrible, afraid/annoyed, happy/glad/joyful, evil/mean) along with essay rubric 3, grammatical errors, and spelling errors. The presence of grammatical and spelling errors as independent variables indicates some similarity between the full model and the competence models above, as spelling and grammatical accuracy were exogenous variables in those models as well. The fact that essay rubric 3 is exogenous is unexpected. Essay rubric metrics are generally determined by other metrics. It is likely that this is different in the full event data set. The number of nouns and number of verbs serve as the only two endogenous variables in the structure, mostly determined by other types of count metrics. Specifically, the number of content words in the document is a strong positive common cause of both the number of nouns, and the number of verbs.

Essay score retains its status of being fully determined by other variables. In fact, most of the other competence variables become indirect causes of essay score through the essay rubric measures. Spelling also remains determined only by the number of words and spelling errors in the document, which was part of the background knowledge assumed over the graph. The number of spelling errors, however, has an effect on a number of other variables. Many of the competences will have words as a latent common cause in the previous models because of the background knowledge.

The covariance matrix for the interactions of all 53 variables on the final essay records can be found in figure A.1, and the correlation matrix in figure A.2 in Appendix A.

Figure 4-5: Causal Structure from FGS Search over All Variables – Final Essays

The FCI graph over the same variable and data set (figure 4-6) is interestingly much sparser, with only 86 edges in total. Of these, only 23 edges are directed, the remainder being bidirected, or having uncertain causal directions and structures between many variables. This indicates that there a host of unmeasured latent causes in the structure. Therefore, even with the large number of measurements taken on the final essays, there are still more to be found and tested.

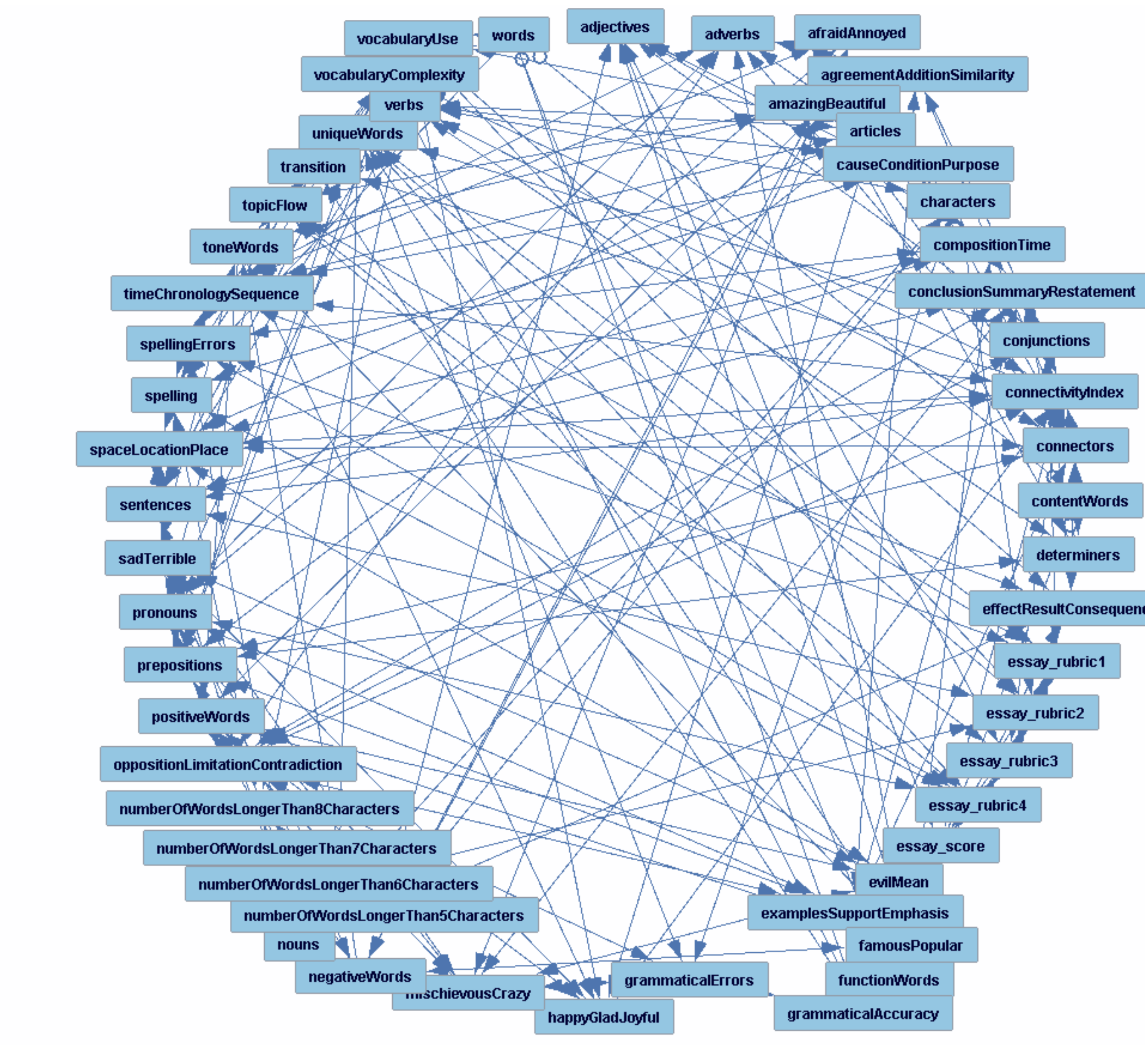
Figure 4-6: FCI Search over All Variables – Final Essays

A PC search over all of the variables with the complete event data intact (figure 4-7) gives a total of 156 total edges, 2 of which are undirected and the rest directed. There is a smaller maximum degree for variables in this structure (only 11) compared to the previous data segment (which had a maximum degree of 17). Overall, therefore, there are fewer potential relations in the developmental data, indicating that each individual variable will have stronger overall effects. There are a larger number of exogenous and potentially-exogenous variables in this structure as well. Essay rubric 3, spelling errors, negative words, unique words, function and content words, and the number of words longer than 6 or 8 characters all appear to be independent in this structure. Notably, essay rubric 3 remains independent, perhaps indicating that the metrics that determine it do not do so strongly enough to be dependent in the structure. In addition, four sets of word quality measures are exogenous (sad/terrible, time/chronology/sequence, opposition/limitation/contradiction, and cause/condition/purpose). The number of articles, adverbs, and nouns in the document, the transition competency, composition time, and the afraid/annoyed quality variable are considered endogenous. In this data set, essay score loses its endogenous nature, having an effect on connectivity index.

Figure 4-7: Causal Structure from PC Search over All Variables – All Events

In contrast to the FCI search over the final essays, the FCI search over the full data (figure 4-8) is the densest model of them all, at 170 nodes. Only 21 of these nodes are directed, indicating as in the previous model that there are a large number of variables external to this system that are latent common causes to many of these metrics.

Figure 4-8: FCI Search over All Variables – All Events



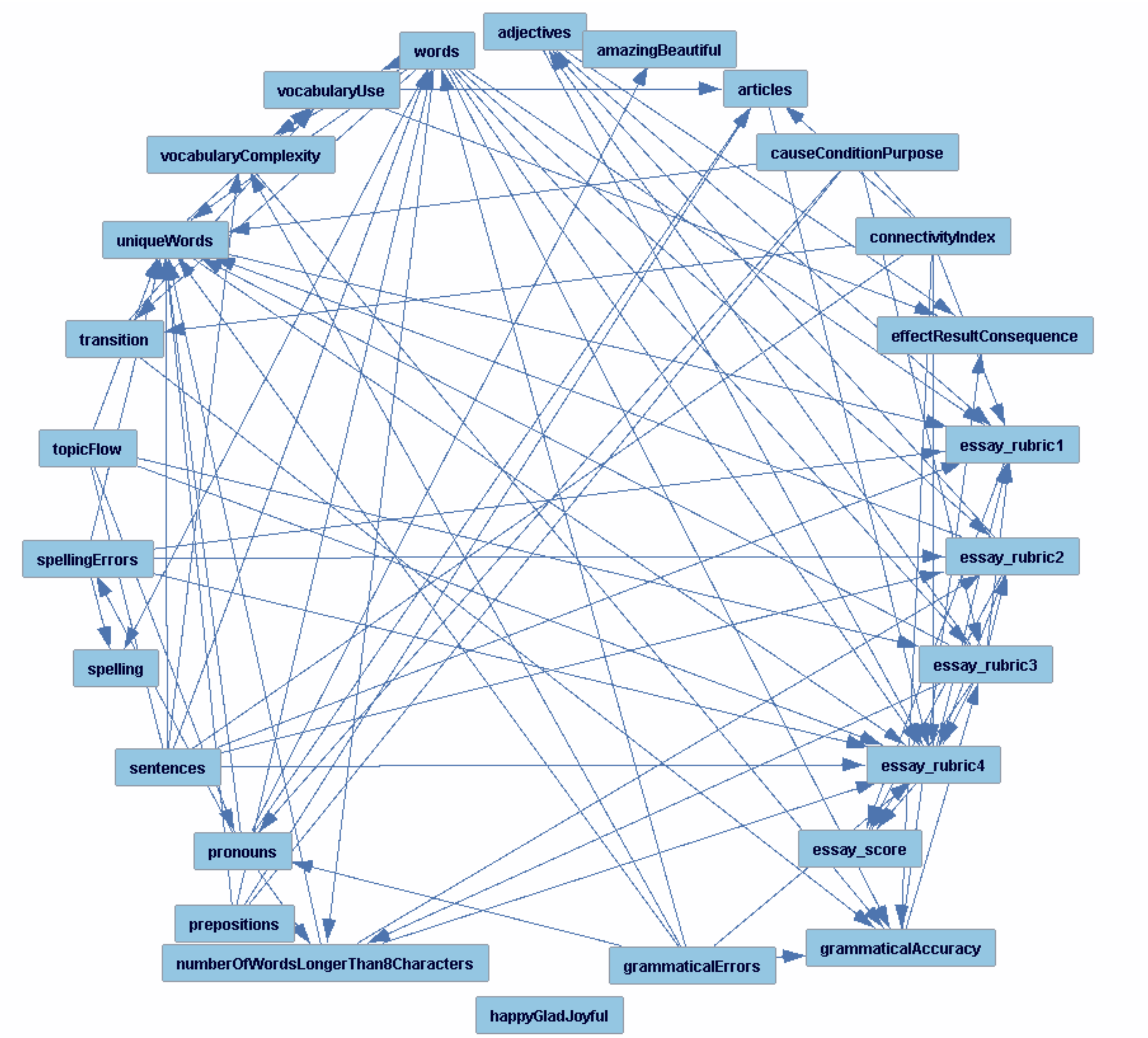
Causal Interactions of Competences and First-Order Adjacencies

From the causal adjacencies determined in the last section, a final set of models can be determined. 26 variables from the above models constitute the seven competences plus all of their immediate adjacencies, whether upstream or downstream causally. This smaller determination of variables may provide a simpler approach to identifying the relevant latent factors between competencies while constituting a simpler structural equation model that is more likely to fit the data.

As with the previous section, only the causal patterns and PAGs will be given here, as the full SEM specifications are too dense to be displayed visually. The covariance and correlation matrices of the SEMs are given in Appendix A.

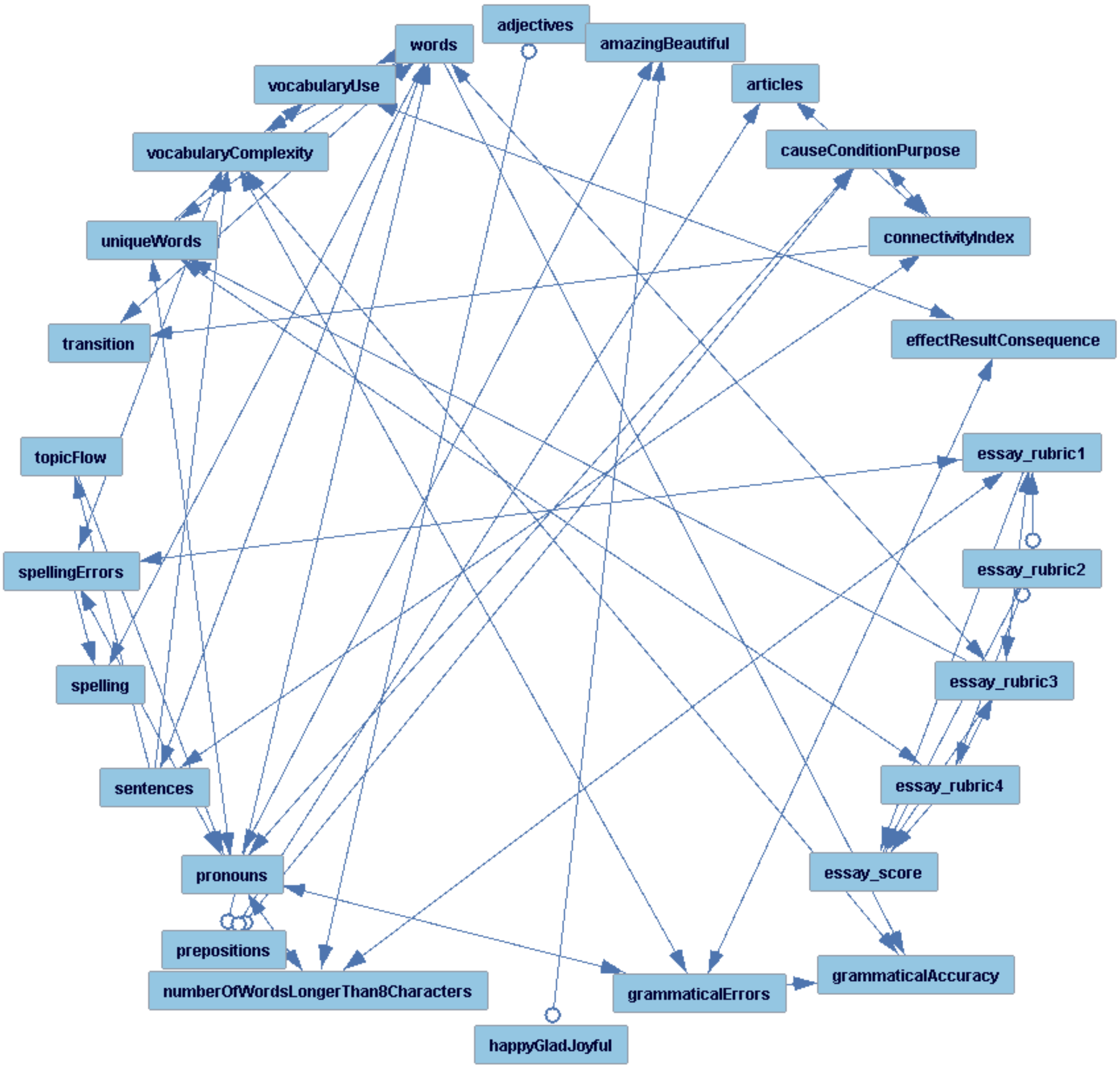
The FGS search shown in figure 4-9 shows the causal structure over the 390 final essays only. The graph has 83 edges, 79 directed and 4 undirected. The exogenous variables on this result are grammatical errors, sentences, topic flow, cause/condition/purpose, connectivity index, and prepositions. This is a decidedly different set of independent variables from previous searches. Most notably, essay rubric 3 has some causes in the form of topic flow, grammatical accuracy, number of words, and cause/condition/purpose. This seems to match intuition more closely. Endogenous variables include spelling (as in previous versions of the models, spelling seems primarily to be an effect), essay score, effect/result/consequence, and amazing/beautiful.

Figure 4-9: Causal Structure from FGS Search over Competences and Adjacencies – Final Essays



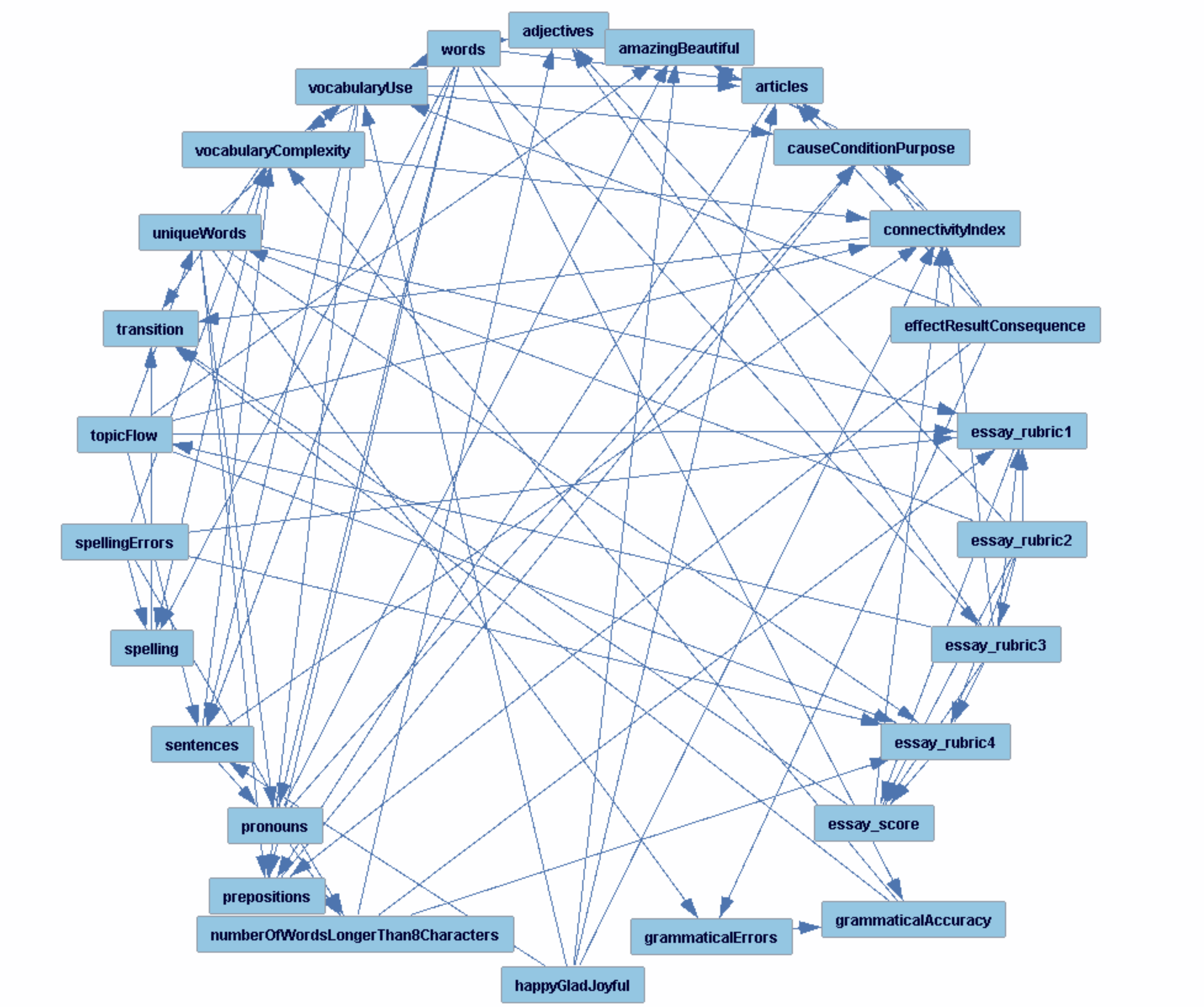
As in the previous data set, the FCI graph (Figure 4-10) is considerably sparser than the FGS search for the adjacencies over the completed essays. The results of the FCI search include 46 edges, 17 of which are directed, 22 of which are bidirected. The remaining nodes have uncertain latent structures in between, some of which are qualified by the complete graph in the previous section.

Figure 4-10: FCI Search over Competences and Adjacencies – Final Essays



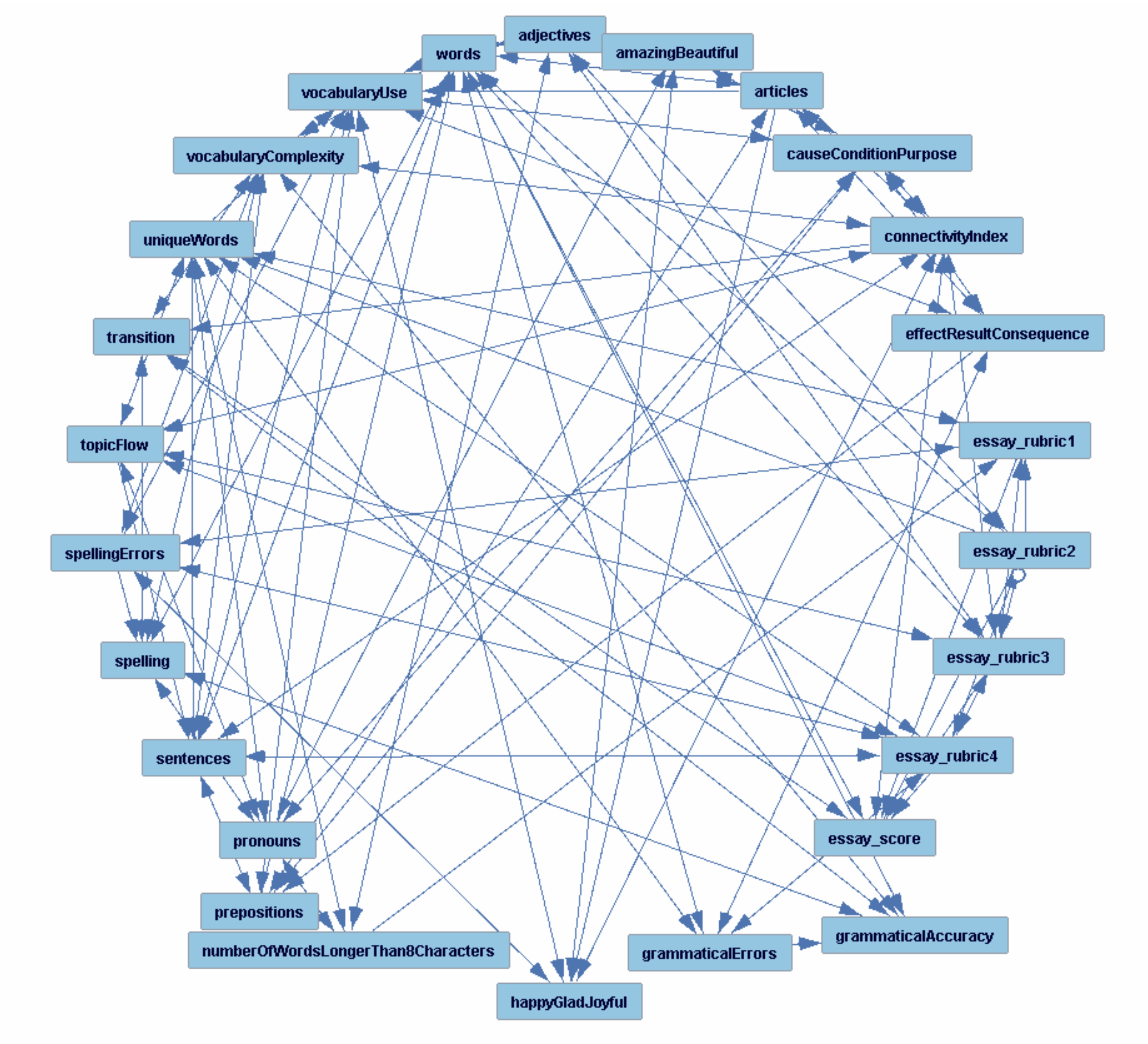
Utilizing the full set of event records and running a PC search yields the results in figure 4-11. There are 80 edges in the structure, all of them directed. The exogenous variables are similar to the analogous search over all of the variables: words, essay rubric 2, happy/glad/joyful, spelling errors, and effect/result/consequence. The endogenous variables are the transition competence and the number of adjectives.

Figure 4-11: Causal Structure from PC Search over Competences and Adjacencies – Final Essays



Finally, the FCI search over the adjacencies for the full event set is given in Figure 4-12. This set consists of 85 edges: 21 directed and 63 bidirected. There are latent common causes between nearly all of the variables, a result that is consistent with the full-variable set.

Figure 4-12: FCI Search over Competences and Adjacencies – Final Essays



Reliability of Models

The goodness-of-fit statistics for the structural equation models generated from choosing DAGs from the FGS and PC searches described in the previous sections are shown in Table 4-6 below. The minimally-determined dataset, of competence variables over only the 390 final essays, is the most reliable model for causation. Its p-value is greater than 0.05, indicating that the model cannot be rejected. The CFI and RMSEA scores are also very strong.

The p-values for all of the remaining models give cause to reject them, however with such a large sample size and large number of variables, the p-values are less meaningful compared to the other goodness-of-fit statistics. The competence model over all of the events is the next-strongest. Its CFI and RMSEA values are within acceptable parameters to validate this model as well.

The models involving all of the variables are uncertain in every respect, and cannot be used as statistically accurate models. When making conclusions, we can use these models only to speculate what some of the latent common causes to the competences might be within the structure. These are not appropriate causal models in their own rights. This phenomenon has been shown in the results above, through the indication that many causal directions change between the models, and are inconsistent between the full-data and final-essay pairs. In summary, the interventions on various individual metrics have very complex interactions, and it's impossible to determine with certainty how manipulating them will influence competences.

Table 4-6: Model Fit Statistics

	Competences	**Competences**	**All Variables**	**All Variables**	**Adjacencies**	**Adjacencies**
	Final Essays	**All Events**	**Final Essays**	**All Events**	**Final Essays**	**All Events**
Degrees of Freedom	12	5	1213	1220	242	245
Chi Square	5.0814	7568.5300	2.2966E+30	3.7767E+32	4.5784E+05	5.1515E+09
P Value	0.2370	0.0000	0.0000	0.0000	0.0000	0.0000
BIC Score	-56.5124	7500.9253	2.2966E+30	3.7767E+32	4.5640E+05	5.1515E+09
CFI	0.9986	0.9972	0.9994	0.9273	1.0000	1.0000
RMSEA	0.0257	0.0451	2.2062E+12	6.4467E+11	2.2048	5.3131

Chapter 5 will comment on the implications of these results for individual competences, and make recommendations for pedagogy and future research.

Chapter 5 – CONCLUSIONS AND FURTHER RESEARCH

Having the statistical results of the experiment in hand, it is possible to answer the research question in full. Because the question's goal is to connect writing competence with causality, the answer will be presented by individual competence factor. The causes and effects of each will be discussed in isolation, and the most significant ones will be covered in the recommendations.

Conclusions

In the sections below, each competence is examined to determine how it arises from the factors measured by the systems in this experiment. The validity of models will be kept in mind; the less-valid ones used only for speculation.

Essay Score

When evaluated using only the final essay results, essay score is endogenous. It is only an effect, never a cause. The truncated data has vocabulary use, grammatical accuracy, transition, and topic flow as common causes.

FCI on the truncated data reveals that there is a latent common cause between essay score and topic flow, and potential latent systems between essay score and its other three variables, representing systems of potential causes common to both.

A search over all of the events using FGS reveals that vocabulary complexity is an additional cause of essay score. The variable remains endogenous in this case, however in the equivalent FCI search the causal direction differs, with essay score having an effect on both vocabulary variables. Notably, the ambiguous relations from the other FCI search are gone in the complete data, leaving only one latent common cause between essay score and grammatical accuracy. Notably, the complete data show much more modest coefficients of the determiners on essay score.

In the correlation matrices, the most interesting differences are in vocabulary use, where the correlation directions are nearly flipped. The effects of grammatical accuracy in the truncated data is much larger than in the full data, and conversely, the effects of topic flow and transition on essay score are very pronounced in the full data. The effect of vocabulary complexity on essay score remains about the same.

The full data reveals that the essay score variable is completely determined by the essay rubrics. This is part of the background knowledge. When these variables are included, their coefficients in the model are 1 (for rubrics 1, 2, and 3) or 2 (for rubric 4), indicating the direct linear relationship present in the calculation of essay score from the rubric values. Considering all variables reveals that essay score seems to have an effect on connectivity index. In the adjacency set, this causal effect is on transition instead. Because these two are related, there is an indication that essay score is actually a determiner of transition in some way.

Much of the determination of essay score lies in determination of the rubrics. Interestingly, in the complete data set, rubric 3 is taken to be an exogenous variable. As rubrics are themselves determined by other elements of the system, this does not seem like a valid assessment. In the adjacency data, the rubrics are nearly endogenous, which more accurately represents them. Combining the common effects of final essay data and the full data, rubric 1 has spelling errors, and word count as causes. The search therefore reveals it as a measure of mechanical correctness, particularly in regard to spelling. Rubric 2 is exogenous in full event set, and mostly determined by other rubrics, words and sentences in the final essay data. It is therefore impossible to tell what role rubric 2 plays in determining essay score from the data. Rubric 3 has topic flow as an adjacency in both the full event data and the final essay data. The arrowheads are reversed in each, but the consistency suggests a relationship. Essay rubric 4 is also determined by number of words of various sorts, but also adds spelling errors, topic flow, and grammatical accuracy as causes.

In conclusion, essay score is caused by a combination of other competences. The mechanisms by which this occurs is a series of rubrics, which themselves are determined by other competences or components thereof. One consistent latent common cause for many competences to the various rubrics is the number of words in the document, which is used to determine overall ratios as part of calculating the rubrics themselves. Intervening to improve other competences will therefore improve overall essay score.

Vocabulary Use

In the significant model using only the final essays, vocabulary use is an exogenous common cause of essay score and transition. The effect on essay score is fairly strong, and the effect on transition is negative. There are, however, latent variables in both sets. In fact, the relation between vocabulary use and transition is very ambiguous; all that is truly known about it is that the two are not d-separated and therefore dependent in some way.

In the full event data, vocabulary use is negatively impacted by grammatical accuracy. It is unclear why this might be so. The effect of vocabulary use on essay score becomes negative in the full data. All of its immediate adjacencies have negative coefficients. This can potentially be explained by the nature of the composing process: as the essay is being developed in the early stages, the number of unique words will be high, but the essay will otherwise be incomplete due to lack of topic flow, transition, and other items. In the FCI search, the arrowheads between vocabulary use and essay score are reversed, and relationships between vocabulary use and all the other variables are revealed to have latent common causes.

Examining all the variables reveals the relations from the background knowledge: that words and unique words contribute to vocabulary use. The number of words has a negative effect, and the number of unique words has a positive effect, which is expected given the calculations.

Vocabulary use also has an effect on the number of articles in the document. This effect has a negative coefficient in the SEM, which indicates that better vocabulary use causes fewer articles in the document. There are a number of relationships from vocabulary use to different word-quality metrics. Interestingly, vocabulary use has a negative effect on some of these. The opposite is expected, that with an increased vocabulary range, more different types of words matching these qualities would be produced.

In the adjacencies variable set with the final essays, transition is a direct cause of vocabulary use, having a negative effect on it. In the full event data, this relationship is more complex, having potential common causes of words, vocabulary complexity, connectivity index, and happy/glad/joyful words being a system of effect. The conclusion is that the relationship between transition and vocabulary use exists, but is unclear in its exact mechanisms.

The causal connection of vocabulary use to words means that it will have some connection to the other competence variables, namely grammatical accuracy and vocabulary complexity, as the number of words represents a cause common to both. This even relates to topic flow, as the number of words is a common cause of vocabulary use and an essay rubric, which is further a cause of topic flow.

In conclusion, vocabulary use has a positive effect on essay score, and an interaction with different types of unique words that is unexpected. This may be an artifact of the data, but requires further investigation to prove. Vocabulary interacts with other competences because of a latent common cause in the number of words in the document.

Vocabulary Complexity

The relationship of vocabulary complexity to other competences is uncertain. In the final essay data set, the competence is endogenous, being affected by both grammatical accuracy and spelling. In the full event data set, vocabulary complexity is nearly exogenous, affecting essay score, spelling, and transition, and potentially grammatical accuracy. FCI in both cases relates latent causal systems between all of these connections. These factors all indicate that the causal directions into and out of vocabulary complexity are uncertain, the most that can be determined from these models are the adjacencies.

Examining all of the variables, it is apparent from both data segments that there is an relationship between vocabulary complexity and grammatical accuracy. Vocabulary complexity has interactions directly either with grammatical accuracy itself, or the number of grammatical errors in the document. This is distinct from the existence of words as a latent common cause. While the causal direction of this relationship remains uncertain, grammatical accuracy and vocabulary complexity clearly have some effect on one another.

Vocabulary complexity has a direct effect on transition in the complete data. In the final essay data, this relationship contains the connectivity index and number of sentences as a system of common causes between them. The number of sentences is shown to be a common cause between vocabulary complexity and topic flow.

In conclusion, it is difficult to state exactly what the effects of intervention to increase vocabulary complexity are. However, it is clear that it is linked to grammatical accuracy in some fashion.

Transition

In the final essay data, transition is affected by vocabulary use and has an effect on essay score. When the full event data is added, transition maintains its initial causal relationships, but adds an effect on topic flow, and becomes an effect of vocabulary complexity, grammatical accuracy, and spelling. It is the only competence out of the seven that is fully connected to all of the other competences. The FCI results reveal that there are latent common causes in almost every case between these variables.

As discussed in the previous sections, transition maintains its connections to other competences directly, even in the full data. While some causal directions become inverted when examining the full data, grammatical accuracy, vocabulary use, vocabulary complexity, and spelling all have relations directly to transition. The FCI searches over these more complete data sets indicate other unmeasured common causes.

There are several common-cause structures between topic flow and transition, most of them going between connectivity index and either sentences or adverbs.

Topic Flow

In the competence variable set over the final essays, it is possible that the topic flow variable may be exogenous. There is an undirected edge in the underlying pattern between topic flow and spelling. Otherwise, topic flow has an effect on both transition and essay score. Adding the complete event data reverses many of these relationships. The relation between topic flow and essay score remains consistent, however, topic flow becomes an effect of spelling, vocabulary use (with a negative coefficient) and transition. The FCI searches of these variable sets indicate latent common causes between all of these relationships. The relationship between transition and topic flow is particularly uncertain.

Examining the full data set, it becomes apparent that topic flow is linked to essay score through the essay rubric variables. Both the final essay and the full event data sets indicate relations from topic flow to three of these variables, though the causal direction remains uncertain. The relationship of transition to vocabulary complexity is related through the spelling competence, or through the number of spelling errors. This again relates back to topic flow being used in the rubric system, connecting it to other competences in the system.

In conclusion, topic flow contributes to essay score via the rubric variables, and determines or is determined by some of the other competences.

Spelling

Spelling is potentially exogenous in the system of variables with competences and final essay data. An undirected edge between spelling and topic flow indicates that the causal direction is uncertain. Spelling also has an effect on vocabulary complexity. Adding the full event data puts spelling into topic flow as a cause. Further, it reverses the causal direction of the spelling-vocabulary complexity edge, which indicates some uncertainty in this direction. Spelling also becomes a cause of transition in the full event data. Another undirected edge is added between grammatical accuracy and spelling. The FCI searches over the data indicate that there are latent common causes between all of the adjacencies that connect into the spelling competence.

The full variable set has different results depending on the data set used. The final essay results show only the relations specified in the background knowledge, that words and spelling errors are causes of spelling. However, the number of spelling errors is an exogenous variable in both of the complete data sets, and has a causal effect on essay rubrics, making it an indirect cause of essay score. Further, the number of words constitute a common cause of some other variables, including grammatical accuracy, vocabulary use, and vocabulary complexity. In the full data, the relationship with transition is still direct.

In conclusion, spelling is related to essay score through rubrics, through other competences through words, and directly to transition.

Grammatical Accuracy

Grammatical accuracy is exogenous in the simplest case, affecting only essay score and vocabulary complexity. In the full event data over the competences, the edge between grammatical accuracy and vocabulary complexity is undirected. Grammatical accuracy becomes a cause of vocabulary use, transition, and possibly spelling (via another undirected edge). The FCI algorithm makes an interesting switch of the causal direction between grammatical accuracy and transition. All of the other variables adjacent to grammatical accuracy have latent common causes.

The complete variable set shows that there is some ambiguity in the causal effects grammatical accuracy has on transition and vocabulary complexity. The directions are reversed for these variables in the final essay case and in the full event case. However, it is significant that these two competences seem to be directly related to grammatical accuracy. The number of words represents an immediate common cause between grammatical accuracy and vocabulary use.

In conclusion, grammatical accuracy is a very important competence as it touches many of the others directly.

Recommendations

The implications of this research are primarily directed toward educators in English composition. As causal analysis is a system of interventions, the results above provide a roadmap of which competence variables affect others, given the data set and assumptions we have used. The development of learning analytics systems designed to provide feedback to students based on learning traces (like SCALE) could also benefit from the recommendations below. As per the definition of learning analytics, these results will foster “understanding and optimization of learning and the environments in which it occurs,” particularly within the specific domain of writing.

It is important to recognize that these recommendations are in the context of making single interventions on competences. In cases where an instructor wishes to improve a competence, these recommendations will inform which single other competence to intervene upon to improve. Targeting multiple improvements at once may yield different results.

Pedagogically speaking, this set of models indicates that the most important competence to intervene upon to see improvements in other competences is grammatical accuracy. In the most reliable models, grammatical accuracy has positive effects on essay score directly, and increases vocabulary complexity modestly. It is also related positively to good transition within a composition, which will also increase overall essay score.

Vocabulary complexity, or readability, is also an important competence. Developmentally, keeping the document at a readable level, appropriate to the level of readership, is important to improve essay score and transition.

The transition variable is well-correlated with essay score. The more connectivity there is in a document, the better the essay will be overall. When developing essays, maintaining good transition and connectivity translates into good topic flow, and improve essay score. Transition can be improved by keeping other competences high during development, but it is likely that more research about the transition competence should be conducted to determine more reliable ways of inducing or intervening to increase it.

Vocabulary use, while appearing to negatively affect other competences in development, has a large effect on essay score in the finished products of the essays. This could be a product of the simulation process, which replaces certain words and phrases with others, increasing lexical diversity, but decreasing other competences like topic flow and grammatical accuracy until they are revised. A potential recommendation, then, is to encourage increasing vocabulary use and diversity of language during Stage II revisions.

The methods of this research make several assumptions, and these assumptions carry through to the potential recommendations. The writing events are simulations, not actual student traces. While care has been taken to mimic the behaviors of real students, and over such a large sample size of events, there may be little discernable difference between a developmental analysis of real and simulated data, it is important to bear in mind that the real world may differ from the results here. For this reason, the section below on further research in the area makes the suggestion that real-world data be obtained.

Future Research

Similar studies should be conducted with real-world low-level writing data when available. While the technology for sending this large volume of data over a distributed environment are challenging, it is likely that local studies would be more successful in recording granular writing traces. Moreover, a more efficient data storage mechanism could be devised where only keystroke streams are sent over the network, and are then correctly parsed into writing events. Relaxing some of the assumptions in the causal analysis models could also yield more accurate results. The assumption that competences are linear determinants of one another could be challenged, and generalized SEMs created to better estimate the interactions as nonlinear functions.

Exact interactions of the competences and metrics of interest could be studied in more depth using randomized experiments to test for more specific and precise outcomes.

Several relationships between competences were identified in this study that represent unmeasured common causes of several variables, even with the large suite of metrics included. Further research into writing analytics to discover more measurable variables and find creative uses of natural language processing tools would be valuable in a continued attempt to fully determine the systems that produce writing competence. Additionally, the computation of competence values could itself be refined and experimented upon, further validating the measurement methods.

Finally, this sort of analysis could be carried out upon a variety of other applications within learning analytics where learning traces are obtained.

REFERENCES

[1] "English Language Arts (K-9)," Alberta, 2000.

[2] L. Adler-Kassner and P. O'Neill, Reframing writing assessment to improve teaching and learning, Utah State University Press, 2010.

[3] S. A. Dansieh, "SMS texting and its potential impacts on students' written communication skills," *International Journal of English Linguistics,* vol. 1, p. 222, 2011.

[4] D. Murray, "Teach writing as a process not product," *The Leaflet,* vol. 71, pp. 11-14, 1972.

[5] A. Goldberg, M. Russell and A. Cook, "The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002," *The Journal of Technology, Learning and Assessment,* vol. 2, 2003.

[6] Dictionary.com, "competence," *Dictionary.com Unabridged,* Oct 2015.

[7] D. Sampson and D. Fytros, "Competence models in technology-enhanced competence-based learning," in *Handbook on Information Technologies for Education and Training*, Springer, 2008, pp. 155-177.

[8] F. W. Taylor, The principles of scientific management, Harper, 1914.

[9] D. C. McClelland, "Testing for competence rather than for" intelligence."," *American psychologist,* vol. 28, p. 1, 1973.

[10] R. M. Stogdill, Handbook of leadership: A survey of theory and research., Free Press, 1974.

[11] S. E. Dreyfus and H. L. Dreyfus, "A five-stage model of the mental activities involved in directed skill acquisition," 1980.

[12] M. A. Hoge, J. Tondora and A. F. Marrelli, "The fundamentals of workforce competency: Implications for behavioral health," *Administration and Policy in Mental Health and Mental Health Services Research,* vol. 32, pp. 509-531, 2005.

[13] J. Intagliata, D. Ulrich and N. Smallwoood, "Leveraging leadership competencies to produce leadership brand: Creating distinctiveness by focusing on strategy and results," *People and Strategy,* vol. 23, p. 12, 2000.

[14] G. Hamel and C. K. Prahalad, Competing for the Future, Harvard Business Press, 2013.

[15] C. K. Prahalad and G. Hamel, "The core competence of the corporation," in *Strategische unternehmungsplanung—strategische unternehmungsführung*, Springer, 2006, pp. 275-292.

[16] I. M. S. G. L. Consortium, "IMS reusable definition of competency or educational objective specification," IMS Global Learning Consortium,. Retrieved from http://www. imsglobal. org/competencies/index. html, 2005.

[17] H. Consortium, "HR-XML 3.3 Assessments Specifications," 2013.

[18] D. H. J. M. Dolmans, W. De Grave, I. H. A. P. Wolfhagen and V. D. a. C. P. M. Vleuten, "Problem-based learning: Future challenges for educational practice and research," *Medical education,* vol. 39, pp. 732-741, 2005.

[19] L. Barnard, W. Y. Lan, Y. M. To, V. O. Paton and S.-L. Lai, "Measuring self-regulation in online and blended learning environments," *The Internet and Higher Education,* vol. 12, pp. 1-6, 2009.

[20] L. Barnard-Brak, V. O. Paton and W. Y. Lan, "Profiles in self-regulated learning in the online learning environment," *The International Review of Research in Open and*

Distributed Learning, vol. 11, pp. 61-80, 2010.

[21] A. Bandura and R. H. Walters, "Social learning theory," 1977.

[22] J. Richardson and K. Swan, "Examing social presence in online courses in relation to students' perceived learning and satisfaction," 2003.

[23] G. Cheetham and G. E. Chivers, Professions, competence and informal learning, Edward Elgar Publishing, 2005.

[24] T. Anderson, The theory and practice of online learning, Athabasca University Press, 2008.

[25] G. M. Ferguson, "Knowledge Representation and Reasoning for Mixed-Initiative Planning.," 1995.

[26] A. Gaur, A history of writing, British library, 1992.

[27] S. Scribner, M. Cole and M. Cole, The psychology of literacy, vol. 198, Harvard University Press Cambridge, MA, 1981.

[28] "Knolwedge and Employability English Language Arts Grades 8 and 9," Alberta, 2000.

[29] N. Verhelst, P. Van Avermaet, S. Takala, N. Figueras and B. North, Common European Framework of Reference for Languages: learning, teaching, assessment, Cambridge University Press, 2009.

[30] J. A. Langer, "Beating the odds: Teaching middle and high school students to read and write well," *American Educational Research Journal,* vol. 38, pp. 837-880, 2001.

[31] S. Perl, "The composing processes of unskilled college writers," *Research in the Teaching of English,* pp. 317-336, 1979.

[32] A. Raimes, "What unskilled ESL students do as they write: A classroom study of composing," *Tesol Quarterly,* vol. 19, pp. 229-258, 1985.

[33] Dictionary.com, "analysis," *Dictionary.com Unabridged*, Jun 2014.

[34] M. Beaney, "Analysis," in *The Stanford Encyclopedia of Philosophy*, Summer 2016 ed., E. N. Zalta, Ed., Metaphysics Research Lab, Stanford University, 2016.

[35] A. Cooper, "A brief history of analytics," *Analytics Series*, vol. 1, 2012.

[36] B. Cohen, "Florence Nightingale," *Scientific American*, no. 250, pp. 128-137, 1984.

[37] A. G. Picciano, "The Evolution of Big Data and Learning Analytics in American Higher Education.," *Journal of Asynchronous Learning Networks*, vol. 16, pp. 9-20, 2012.

[38] "LAK '11: Proceedings of the 1st International Conference on Learning Analytics and Knowledge," New York, NY, USA, 2011.

[39] G. Siemens and S. J. d Baker and Ryan, "Learning analytics and educational data mining: towards communication and collaboration," in *Proceedings of the 2nd international conference on learning analytics and knowledge*, 2012.

[40] E. Duval, "Attention please!: learning analytics for visualization and recommendation," in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, 2011.

[41] D. Clow, "MOOCs and the funnel of participation," in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, 2013.

[42] R. Ferguson, "Learning analytics: drivers, developments and challenges," *International Journal of Technology Enhanced Learning*, vol. 4, pp. 304-317, 2012.

[43] K. E. Arnold and M. D. Pistilli, "Course signals at Purdue: using learning analytics to increase student success," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 2012.

[44] D. Boulanger, J. Seanosky, M. Baddeley, V. Kumar and others, "Learning Analytics in the Energy Industry: Measuring Competences in Emergency Procedures," in *Technology for Education (T4E), 2014 IEEE Sixth International Conference on*, 2014.

[45] P. M. Johnson, H. Kou, J. Agustin, C. Chan, C. Moore, J. Miglani, S. Zhen and W. E. J. Doane, "Beyond the personal software process: Metrics collection and analysis for the differently disciplined," in *Proceedings of the 25th international Conference on Software Engineering*, 2003.

[46] V. Kumar, T. Somasundaram, S. Harris, D. Boulanger, J. Seanosky, G. Paulmani, K. Panneerselvam and others, "An Approach to Measure Coding Competency Evolution," in *Smart Learning Environments*, Springer, 2015, pp. 27-43.

[47] C. Guillot, R. Guillot, V. Kumar and others, "MUSIX: Learning Analytics in Music Teaching," in *State-of-the-Art and Future Directions of Smart Learning*, Springer, 2016, pp. 269-273.

[48] S. Graf, K. B. Kinshuk, F. A. Khan, P. Maguire, A. Mahmoud, T. Rambharose and Q. Zhang, "Adaptivity and personalization in learning systems based on students' characteristics and context," in *The 1st international symposium on smart learning environment*, 2012.

[49] L. Zheng, M. M. El-Bishouty, C. Pinnell, J. Bell, V. Kumar and others, "A Framework to Automatically Analyze Regulation," in *Emerging Issues in Smart Learning*, Springer, 2015, pp. 23-30.

[50] G. P. Wiggins, Assessing student performance: Exploring the purpose and limits of testing., Jossey-Bass, 1993.

[51] L. P. Macfadyen and S. Dawson, "Numbers are not enough. Why e-learning analytics failed to inform an institutional strategic plan," *Journal of Educational Technology \& Society,* vol. 15, pp. 149-163, 2012.

[52] K. Govindarajan, T. S. Somasundaram, D. Boulanger, V. S. Kumar and others, "A framework for scheduling and managing big data applications in a distributed infrastructure," in *Advanced Computing (ICoAC), 2015 Seventh International Conference on*, 2015.

[53] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein and C. Welton, "MAD skills: new analysis practices for big data," *Proceedings of the VLDB Endowment,* vol. 2, pp. 1481-1492, 2009.

[54] K. Govindarajan, D. Boulanger, V. S. Kumar and others, "Parallel Particle Swarm Optimization (PPSO) clustering for learning analytics," in *Big Data (Big Data), 2015 IEEE International Conference on*, 2015.

[55] V. Kumar, D. Boulanger, J. Seanosky, K. Panneerselvam, T. S. Somasundaram and others, "Competence analytics," *Journal of Computers in Education,* vol. 1, pp. 251-270, 2014.

[56] D. Boulanger, J. Seanosky, C. Pinnell, J. Bell, V. Kumar and others, "SCALE: a competence analytics framework," in *State-of-the-Art and Future Directions of Smart Learning*, Springer, 2016, pp. 19-30.

[57] A. M. Turing, "Computing machinery and intelligence," *Mind,* pp. 433-460, 1950.

[58] J. R. Bellegarda, "Spoken language understanding for natural interaction: The siri experience," in *Natural Interaction with Robots, Knowbots and Smartphones*, Springer, 2014, pp. 3-14.

[59] N. Chomsky, Syntactic structures, Walter de Gruyter, 2002.

[60] J. Hutchins, "From first conception to first demonstration: the nascent years of machine translation, 1947--1954. a chronology," *Machine Translation,* vol. 12, pp. 195-252, 1997.

[61] T. Winograd, "Procedures as a representation for data in a computer program for understanding natural language," 1971.

[62] R. Flesch, "A new readability yardstick.," *Journal of applied psychology,* vol. 32, p. 221, 1948.

[63] J. P. Kincaid, F. a. R. P. Jr, R. L. Rogers and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," 1975.

[64] P. Brown, J. Cocke, S. D. Pietra, V. D. Pietra, F. Jelinek, R. Mercer and P. Roossin, "A statistical approach to language translation," in *Proceedings of the 12th conference on Computational linguistics-Volume 1*, 1988.

[65] M. P. Marcus, M. A. Marcinkiewicz and B. Santorini, "Building a large annotated corpus of English: The Penn Treebank," *Computational linguistics,* vol. 19, pp. 313-330, 1993.

[66] X. Carreras and L. Marquez, "Introduction to the CoNLL-2005 shared task: Semantic role labeling," in *Proceedings of the Ninth Conference on Computational Natural Language Learning*, 2005.

[67] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the conference on empirical methods in natural language processing (EMNLP),* 2013.

[68] K. Toutanova, D. Klein, C. D. Manning and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 2003.

[69] C. D. Manning, "Part-of-speech tagging from 97% to 100 is it time for some linguistics?," in *Computational Linguistics and Intelligent Text Processing*, Springer, 2011, pp. 171-189.

[70] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 2003.

[71] Y. Schabes, "Mathematical and computational aspects of lexicalized grammars," 1990.

[72] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu and D. Jurafsky, "Stanfordś multi-pass sieve coreference resolution system at the CoNLL-2011 shared task," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 2011.

[73] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky and C. Manning, "A multi-pass sieve for coreference resolution," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

[74] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the 2nd international conference on Knowledge capture*, 2003.

[75] J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, 2003.

[76] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval,* vol. 2, pp. 1-135, 2008.

[77] K. J. Gilhooly and R. H. Logie, "Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words," *Behavior Research Methods \& Instrumentation,* vol. 12, pp. 395-427, 1980.

[78] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM,* vol. 38, pp. 39-41, 1995.

[79] T. K. Landauer, P. W. Foltz and D. Laham, "An introduction to latent semantic analysis," *Discourse processes,* vol. 25, pp. 259-284, 1998.

[80] S. T. OŔourke and R. A. Calvo, "Analysing Semantic Flow in Academic Writing.," in *AIED*, 2009.

[81] S. T. OŔourke, R. A. Calvo and D. S. McNamara, "Visualizing Topic Flow in Students ́Essays.," *Educational Technology \& Society,* vol. 14, pp. 4-15, 2011.

[82] K. Kyle, "Objective measures of writing quality," Colorado State University. Libraries, 2011.

[83] M. D. Shermis, J. Burstein, D. Higgins and K. Zechner, "Automated essay scoring: Writing assessment and instruction," *International encyclopedia of education,* vol. 4, pp. 20-26, 2010.

[84] D. Boulanger, J. Seanosky, C. Clemens, V. Kumar and others, "SCALE: A Smart Competence Analytics Solution for English Writing," in *Advanced Learning Technologies (ICALT), 2016 IEEE 16th International Conference on*, 2016.

[85] C. Clemens, V. Kumar and D. Mitchnick, "Writing-Based Learning Analytics for

Education," in *2013 IEEE 13th International Conference on Advanced Learning Technologies*, 2013.

[86] W.-k. Ahn, C. W. Kalish, D. L. Medin and S. A. Gelman, "The role of covariation versus mechanism information in causal attribution," *Cognition,* vol. 54, pp. 299-352, 1995.

[87] B. Russell, "On the notion of cause," in *Proceedings of the Aristotelian society*, 1912.

[88] C. S. Peirce and J. Jastrow, "On small differences in sensation," 1884.

[89] P. Spirtes, C. Glymour and R. Scheines, Causation, prediction, and search, vol. 81, The MIT Press, 2000.

[90] E. H. Simpson, "The interpretation of interaction in contingency tables," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 238-241, 1951.

[91] D. Geiger and J. Pearl, "Axioms and algorithms for inferences involving conditional independence," 1989.

[92] P. Spirtes, "Variable definition and causal inference," 2008.

[93] H. Cramér, *Mathematical methods of statistics,* JSTOR, 1947.

[94] R. A. Fisher, "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," *Biometrika,* vol. 10, pp. 507-521, 1915.

[95] M. Drton, "Likelihood ratio tests and singularities," *The Annals of Statistics,* pp. 979-1012, 2009.

[96] G. Schwarz and others, "Estimating the dimension of a model," *The annals of statistics,* vol. 6, pp. 461-464, 1978.

[97] P. M. Bentler, "Comparative fit indexes in structural models.," *Psychological bulletin,* vol. 107, p. 238, 1990.

[98] R. C. MacCallum, M. W. Browne and H. M. Sugawara, "Power analysis and determination of sample size for covariance structure modeling.," *Psychological methods*, vol. 1, p. 130, 1996.

[99] L. Van Waes and P. J. Schellens, "Writing profiles: The effect of the writing mode on pausing and revision patterns of experienced writers," *Journal of pragmatics*, vol. 35, pp. 829-853, 2003.

[100] D. X. Song, D. Wagner and X. Tian, "Timing Analysis of Keystrokes and Timing Attacks on SSH.," in *USENIX Security Symposium*, 2001.

[101] H. Nesi, G. Sharpling and L. Ganobcsik-Williams, "Student papers across the curriculum: Designing and developing a corpus of British student writing," *Computers and Composition*, vol. 21, pp. 439-450, 2004.

[102] D. M. Chickering, "Learning equivalence classes of Bayesian-network structures," *Journal of machine learning research*, vol. 2, pp. 445-498, 2002.

APPENDIX A – Additional Tables

Table A-1: Covariance Matrix for FGS Search over All Variables – Final Essays

	CT	Adj	Adv	AA	AAS	AB	Art	CCP	C	CSR	Conj	CI	Conn	CW	D	ERC	ER1	ER2	ER3	ER4	ES	EM	ESE	FP	FW	GE
CT	7.2601E+12																									
Adj	0.0000	1115.7211																								
Adv	0.0000	190.1139	530.8073																							
AA	0.0000	0.0000	0.0000	1.2465																						
AAS	0.0000	268.7837	216.1654	0.0000	635.4212																					
AB	0.0000	7.4228	9.3884	0.3177	11.8442	12.4570																				
Art	0.0000	225.6624	45.7415	-0.2358	14.8148	4.2123	1792.2444																			
CCP	0.0000	37.5189	96.3147	0.0000	104.6164	3.0644	63.1179	90.6749																		
C	0.0000	21295.0718	5702.6344	0.0000	12819.3012	361.3008	17945.1671	1759.6477	1.1071E+06																	
CSR	0.0000	1.6871	2.1789	0.0000	2.9410	0.0868	-0.8844	0.7155	80.1730	2.1434																
Conj	0.0000	149.0419	64.9893	1.0104	275.9258	6.0215	-38.6991	7.6248	7134.7164	1.1702	259.0541															
CI	0.0000	0.6309	1.3568	0.0000	-0.4357	0.0107	2.1085	0.4319	20.0860	0.0004	-0.3690	0.0379														
Conn	0.0000	620.6068	784.9436	0.0000	990.2700	30.5561	378.0636	256.2041	30207.5904	7.2570	392.5214	0.8923	2554.7319													
CW	0.0000	2327.4797	705.4553	0.0000	1402.2967	40.8075	2041.9018	234.0122	1.1700E+05	9.0605	744.2285	3.5005	3411.8280	14775.6194												
D	0.0000	21.2869	26.9237	0.0000	33.9664	1.0481	12.9676	8.7878	1036.1239	0.2489	13.4635	0.0306	87.6276	117.0261	26.5674											
ERC	0.0000	45.9200	19.6879	0.0000	29.5335	1.3472	98.3987	14.6356	2412.6139	0.2874	5.7651	0.0930	112.6351	323.1787	3.8634	74.2382										
ER1	0.0000	17.2271	2.9636	0.0000	13.2428	0.2506	2.7663	-0.8144	827.1733	0.0566	9.8362	0.0016	20.9524	82.9456	0.7187	-0.5811	1.3988									
ER2	0.0000	16.2488	3.1029	0.0000	12.3342	0.2422	2.6550	-0.4170	777.9201	0.0551	8.7887	0.0064	20.2457	78.0653	0.6944	-0.4425	1.2115	1.1510								
ER3	0.0000	10.8842	1.3923	0.0000	9.0998	0.1423	-1.8190	-1.1442	510.8002	0.0324	7.3392	0.0032	11.8976	51.1404	0.4081	-1.2670	1.0949	1.0129	0.9658							
ER4	0.0000	13.0262	1.2732	0.0000	10.7451	0.1664	-3.5714	-1.0391	564.6222	0.0396	8.1367	-0.0009	13.9158	56.1941	0.4773	-1.1019	1.1350	1.1022	1.0267	1.1661						
ES	0.0000	70.4124	10.0051	0.0000	56.1670	0.9679	-3.5404	-4.4538	3245.1381	0.2233	42.2376	0.0095	80.9273	324.5394	2.7758	-4.4944	5.9753	5.5797	5.1271	5.5961	27.8742					
EM	0.0000	0.0000	0.0000	0.0000	0.0000	0.7470	-0.5543	0.0000	0.0000	0.0000	2.3753	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	2.9129				
ESE	0.0000	23.6295	30.5173	0.0000	41.1910	1.2157	-12.3867	10.0209	1122.8774	1.1223	16.3888	0.0050	101.6396	126.8979	3.4862	4.0249	0.7929	0.7721	0.4532	0.5543	3.1270	0.0000	15.7185			
FP	0.0000	21.2805	27.9452	0.0000	25.4009	0.7339	1.3963	3.9660	951.6956	0.1707	13.6958	0.0489	61.3577	100.4286	2.1046	-0.6337	1.5429	1.4239	1.2682	1.2550	6.7450	0.0000	2.3903	19.5652		
FW	0.0000	1061.7400	1018.9156	0.0000	1179.4471	43.5046	2390.9296	334.2567	85354.9390	9.5305	513.6839	2.5837	3637.3308	8565.0367	124.7609	210.0204	47.8228	45.3693	26.0287	18.6312	156.4833	0.0000	133.4808	106.8210	13625.9443	
GE	0.0000	-31.3734	-2.7331	0.0000	7.4578	-0.0942	-37.4042	-0.4783	0.0000	0.0068	2.4851	-0.0611	-7.8731	-128.4926	-0.2700	-16.4273	-0.4167	0.2768	0.0000	-1.4683	-3.0766	0.0000	0.0954	5.2649	289.9713	251.1954
HGJ	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MC	0.0000	0.0000	0.0000	0.0386	0.0000	0.0507	-0.0376	0.0000	0.0000	0.0000	0.1613	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0908	0.0000	0.0000	0.0000	0.0000
NW	0.0000	0.0000	0.0000	1.2299	0.0000	1.6150	-1.1983	0.0000	0.0000	0.0000	5.1355	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	2.8914	0.0000	0.0000	0.0000	0.0000
N	0.0000	1297.0216	-447.0017	0.0642	391.3218	8.0035	1074.5725	-43.5358	64471.3899	1.8202	323.9976	-0.2941	662.1030	8356.0883	22.7102	175.4285	48.4679	44.6738	31.0359	33.1517	190.4810	0.1509	25.4928	17.9104	3822.7976	-81.6077
WL5	0.0000	1836.5076	177.0940	0.0000	959.3347	22.1652	967.6361	41.7860	86982.7787	4.9746	607.7722	0.5222	1853.1819	8708.5662	63.5644	146.6856	76.4128	71.0079	49.7827	57.2979	311.7992	0.0000	69.6724	68.9691	4432.3631	0.0000
WL6	0.0000	1549.1749	106.5615	0.0000	767.7616	17.0278	732.5301	18.3298	72323.1566	3.8279	499.4703	0.2016	1423.6547	6782.9588	48.8315	111.3881	63.8499	59.5550	41.3926	48.6425	262.0825	0.0000	53.6129	53.8864	3366.3235	0.0000
WL7	0.0000	1194.7387	42.2863	0.0000	553.4638	11.5767	486.9486	-1.6249	54797.5653	2.6090	372.8881	-0.0671	967.9012	4706.5120	33.1991	74.4007	48.6755	45.6096	31.3622	37.8017	201.2507	0.0000	36.5410	37.5594	2249.0634	0.0000
WL8	0.0000	897.1514	12.1930	0.0000	394.7785	7.8686	322.2939	-8.1743	40332.7824	1.7792	271.8668	-0.1014	657.8730	3371.9907	22.5651	52.4057	35.9024	33.7590	23.0836	28.2963	149.3376	0.0000	24.9189	25.9736	1436.2177	0.0000
OLC	0.0000	53.2023	118.5750	0.1876	51.0597	2.9053	-31.1317	16.9956	2030.8783	0.6782	41.2497	0.2325	222.3097	227.9879	7.6252	-3.6250	1.6801	1.6105	1.0737	1.0850	6.5343	0.4411	9.4987	7.5542	311.8012	1.1735
PW	0.0000	1.3530	6.1401	0.1148	5.1728	2.5578	-0.5471	1.6375	76.9227	0.0406	2.2285	0.0098	13.4198	5.8884	0.4603	-0.1867	0.0393	0.0428	0.0201	0.0205	0.1431	0.2700	0.5689	0.3706	22.4945	0.5396
Pre	0.0000	377.0194	329.6077	0.0000	306.9564	14.0071	717.8174	100.9081	23166.2813	2.9187	136.0450	0.1256	1171.1078	2465.6694	40.1691	73.4208	15.6227	13.7518	8.2151	6.9449	51.4795	0.0000	40.8785	31.6817	2998.2573	-50.8228
Pro	0.0000	-7.2280	415.7855	0.0000	277.6346	8.6527	-125.2451	100.5357	709.8282	2.2944	81.1297	0.7465	723.4375	-188.4508	24.8140	-42.7851	-0.8593	-0.3690	-0.7011	-1.1065	-4.1423	0.0000	32.1353	22.0116	1355.0598	53.5096
ST	0.0000	0.0000	0.0000	0.0000	0.0000	0.5549	-0.4117	0.0000	0.0000	0.0000	1.7644	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Sent	0.0000	125.1730	37.7285	0.0000	69.9140	2.0857	167.2745	22.8396	6814.0034	0.4434	30.4830	1.1804	174.3806	965.6169	5.9813	22.9998	3.3371	3.6495	2.4353	4.1813	17.7845	0.0000	6.2095	2.1653	588.1736	-11.8939
SLP	0.0000	43.4656	46.9272	0.0000	24.8624	1.4733	123.0413	14.5499	2463.6509	-0.0922	9.0689	0.4322	123.1781	277.3242	4.2250	11.2536	1.5292	1.4014	0.8305	0.6620	5.0852	0.0000	-1.2916	3.6073	318.6780	-5.5939
SE	0.0000	62.4136	-60.3008	0.0000	-31.8123	-1.3633	109.8738	-28.4771	3867.6823	-0.3876	13.8928	-0.0094	-113.9867	387.2257	-3.9098	4.5299	6.4777	-0.3865	0.0000	-3.1157	-0.1402	0.0000	-5.4290	-1.7942	197.0847	0.0000
TCS	0.0000	29.2133	54.9339	-0.0090	20.7696	1.2544	115.9112	19.3408	1649.9241	0.2023	0.1645	0.7077	105.8624	199.7769	3.6311	21.4635	-0.0806	0.0545	-0.3987	-0.5101	-1.4449	-0.0211	2.8339	2.1002	275.3153	-2.9381
TW	0.0000	1.3633	6.1869	1.3239	5.2122	4.1638	-1.7285	1.6500	77.5091	0.0409	7.2904	0.0099	13.5221	5.9333	0.4638	-0.1881	0.0396	0.0431	0.0202	0.0207	0.1442	3.1124	0.5732	0.3735	22.6660	0.5437
UW	0.0000	867.0633	448.4049	0.0000	1061.3585	22.9487	-24.7781	12.1281	45649.1096	5.3474	698.8579	-0.2634	1918.6896	4562.6429	65.8113	-88.0478	89.4294	81.6270	76.0696	75.5322	398.1904	0.0000	74.8946	132.6841	4193.7673	336.6962
V	0.0000	323.5028	422.4751	-0.0875	572.6600	16.3761	710.9814	149.5254	25638.4377	3.3926	222.9980	1.9018	1378.7825	3397.1936	47.2924	81.6051	14.6286	14.4740	8.0276	9.8760	56.8821	-0.2058	47.5154	32.0577	2348.1542	-30.9354
W	0.0000	3389.2197	1724.3709	0.0000	2581.7438	84.3121	4432.8313	568.2689	2.0236E+05	18.5909	1257.9124	6.0843	7049.1587	23340.6561	241.7870	533.1991	130.7684	123.4346	77.1692	74.8253	481.0227	0.0000	260.3787	207.2496	22190.9810	161.4788
GA	0.0000	3.0674	0.8672	0.0000	0.9099	0.0437	3.6992	0.2600	96.4167	0.0085	0.5145	0.0035	3.6495	16.7396	0.1252	1.0037	0.0837	0.0478	0.0381	0.1042	0.3780	0.0000	0.1184	-0.1434	-3.0683	-11.4371
S	0.0000	-1.0193	4.4242	0.0000	3.4845	0.1312	-2.8496	1.9200	-68.4416	0.0336	0.1195	0.0047	10.9706	-4.7323	0.3763	0.1236	-0.2578	0.1057	0.0531	0.2188	0.3385	0.0000	0.4707	0.2389	4.6862	0.1111
TF	0.0000	43.3177	34.1393	0.0000	32.7503	0.9806	29.3639	7.3511	2023.7152	0.2202	16.5493	0.2604	81.9888	254.6707	2.8122	3.8328	1.9148	1.9170	1.6873	1.9839	9.4869	0.0000	3.0844	3.0341	177.8335	-2.0404
T	0.0000	2.3497	1.4005	0.0000	0.8832	0.0398	3.1811	0.4416	113.9030	0.0075	0.4083	0.0264	3.3313	12.9318	0.1143	0.3334	0.0760	0.0750	0.0486	0.0509	0.3014	0.0000	0.1055	0.1173	9.4535	-0.0397
VC	0.0000	0.2274	0.1085	0.0000	0.1119	0.0049	0.2824	0.0218	8.8599	0.0010	0.0623	-0.0007	0.4090	1.1811	0.0140	0.0656	0.0083	0.0052	0.0036	0.0056	0.0282	0.0000	0.0143	-0.0006	0.3108	-0.7129
VU	0.0000	-11.8536	-5.6611	0.0000	37.8682	-0.1000	-149.3106	-17.0272	-1336.5319	0.0129	38.8687	-0.2734	-8.3606	-232.4879	-0.2868	-28.0505	5.8487	5.2012	6.0545	6.0613	29.2270	0.0000	0.1800	8.2620	-231.0408	32.8739

	HGJ	MC	NW	N	WL5	WL6	WL7	WL8	OLC	PW	Pre	Pro	ST	Sent	SLP	SE	TCS	TW	UW	V	W	GA	S	TF	T	VC	VU
HGJ	6.7126																										
MC	0.0000	0.1658																									
NW	0.0000	0.1963	6.2514																								
N	0.0000	0.0102	0.3263	7363.0497																							
WL5	0.0000	0.0000	0.0000	5148.0283	8477.3686																						
WL6	0.0000	0.0000	0.0000	3970.0041	7048.6373	6540.4737																					
WL7	0.0000	0.0000	0.0000	2714.6143	5340.5877	5115.1210	4513.1876																				
WL8	0.0000	0.0000	0.0000	1913.1990	3930.8455	3764.8947	3321.8522	2786.0945																			
OLC	0.0000	0.0300	0.9537	-47.5120	97.9295	71.1693	44.2305	27.1887	65.6184																		
PW	1.5072	0.2403	0.5837	-9.4085	0.6226	0.1141	-0.2950	-0.6379	1.6850	4.0928																	
Pre	0.0000	0.0000	0.0000	1106.8753	1398.9286	1085.8142	749.5951	505.0197	105.0924	5.0177	1323.0877																
Pro	0.0000	0.0000	0.0000	-1051.8539	-349.5012	-303.4368	-242.0410	-206.4773	104.0351	12.9667	222.8231	1084.3585															
ST	0.0000	0.0675	2.1478	0.1121	0.0000	0.0000	0.0000	0.0000	0.3276	0.2005	0.0000	0.0000	2.1402														
Sent	0.0000	0.0000	0.0000	520.5944	414.6938	302.4169	189.0741	130.6343	9.2886	0.9785	47.6830	55.3731	0.0000	247.1760													
SLP	0.0000	0.0000	0.0000	109.5279	141.1127	107.2692	71.7665	47.7283	2.3720	0.2010	129.3787	-7.9119	0.0000	17.7005	84.9576												
SE	0.0000	0.0000	0.0000	379.0702	376.9455	313.4171	237.4688	174.7847	-13.4347	-1.1404	62.2033	-84.2497	0.0000	18.4393	5.9129	445.0797											
TCS	0.0000	-0.0014	-0.0456	10.5376	58.7084	39.0632	20.3816	10.3315	8.3456	0.9565	63.3455	68.8029	-0.0156	33.3168	13.3656	-0.5172	77.2553										
TW	1.5187	0.3574	6.7293	-9.1596	0.6273	0.1150	-0.2972	-0.6427	2.6347	4.5897	5.0560	13.0655	2.3120	0.9859	0.2026	-1.1491	0.9190	11.2052									
UW	0.0000	0.0000	0.0000	2219.9137	3887.9682	3115.7958	2250.2683	1597.9917	182.3778	8.1564	1176.5237	358.8786	0.0000	43.0933	110.8880	35.5802	-3.2288	8.2185	7215.3307								
V	0.0000	-0.0140	-0.4450	681.5904	1666.0033	1237.2775	797.6304	548.5209	97.6964	8.2522	576.8563	491.2901	-0.1529	294.5329	69.0908	-3.1527	109.3868	7.8780	972.1069	2040.3152							
W	0.0000	0.0000	0.0000	12178.8859	13140.9292	10149.2823	6955.5754	4808.2085	539.7892	28.3829	5463.9267	1166.6090	0.0000	1553.7905	596.0022	584.3104	475.0922	28.5993	8756.4102	5745.3477	45531.6372						
GA	0.0000	0.0000	0.0000	9.5115	6.4842	5.0861	3.5651	2.4953	0.1893	-0.0124	4.9477	-1.9979	0.0000	1.1162	0.5175	0.2896	0.3165	-0.0125	-11.1145	3.9868	13.6713	0.5491					
S	0.0000	0.0000	0.0000	-11.9742	-11.1982	-9.8455	-7.9649	-6.0766	1.0928	0.0808	0.4195	5.3264	0.0000	0.0790	0.0926	-23.4961	0.3547	0.0814	4.1144	4.1224	-0.0461	-0.0061	1.3526				
TF	0.0000	0.0000	0.0000	100.0489	138.5087	106.3590	72.2630	50.8828	8.5928	0.4843	34.4316	28.3587	0.0000	44.5535	6.2184	-0.0377	7.8402	0.4880	106.3562	79.1842	432.5043	0.2702	0.2996	49.4882			
T	0.0000	0.0000	0.0000	5.6343	8.2633	6.7196	4.9485	3.6085	0.3274	0.0134	2.1920	0.5488	0.0000	1.3872	0.5051	0.3462	0.6304	0.0135	3.9124	3.6360	22.3853	0.0196	-0.0032	0.3535	0.0939		
VC	0.0000	0.0000	0.0000	0.6388	0.6094	0.4926	0.3598	0.2488	0.0311	-0.0004	0.5561	-0.1265	0.0000	-0.1355	0.0455	0.0271	0.0083	-0.0004	-0.3018	0.2191	1.4919	0.0439	-0.0004	-0.0118	0.0003	0.0251	
VU	0.0000	0.0000	0.0000	-135.9240	14.8493	23.3995	28.1755	23.8276	3.3553	0.0147	-40.4612	3.2560	0.0000	-45.8325	-6.9451	-14.6839	-16.7259	0.0148	531.7271	-76.4991	-463.5288	-1.7076	0.4695	-2.1098	-0.4515	-0.0803	77.9603

Table A-2: Correlation Matrix for FGS Search over All Variables – Final Essays

	CT	Adj	Adv	AA	AAS	AB	Art	CCP	C	CSR	Conj	CI	Conn	CW	D	ERC	ER1	ER2	ER3	ER4	ES	EM	ESE	FP	FW	GE
CT	1.0000																									
Adj	0.0000	1.0000																								
Adv	0.0000	0.2470	1.0000																							
AA	0.0000	0.0000	0.0000	1.0000																						
AAS	0.0000	0.3192	0.3722	0.0000	1.0000																					
AB	0.0000	0.0630	0.1155	0.0806	0.1331	1.0000																				
Art	0.0000	0.1596	0.0469	-0.0050	0.0139	0.0282	1.0000																			
CCP	0.0000	0.1180	0.4390	0.0000	0.4358	0.0912	0.1566	1.0000																		
C	0.0000	0.6059	0.2352	0.0000	0.4833	0.0973	0.4029	0.1756	1.0000																	
CSR	0.0000	0.0345	0.0646	0.0000	0.0797	0.0168	-0.0143	0.0513	0.0520	1.0000																
Conj	0.0000	0.2772	0.1753	0.0562	0.6801	0.1060	-0.0568	0.0497	0.4213	0.0497	1.0000															
CI	0.0000	0.0970	0.3025	0.0000	-0.0888	0.0155	0.2559	0.2330	0.0981	0.0013	-0.1178	1.0000														
Conn	0.0000	0.3676	0.6741	0.0000	0.7772	0.1713	0.1767	0.5323	0.5680	0.0981	0.4825	0.0907	1.0000													
CW	0.0000	0.5732	0.2519	0.0000	0.4577	0.0951	0.3968	0.2022	0.9148	0.0509	0.3804	0.1479	0.5553	1.0000												
D	0.0000	0.1236	0.2267	0.0000	0.2614	0.0576	0.0594	0.1790	0.1910	0.0330	0.1623	0.0305	0.3364	0.1868	1.0000											
ERC	0.0000	0.1596	0.0992	0.0000	0.1360	0.0443	0.2698	0.1784	0.2661	0.0228	0.0416	0.0554	0.2586	0.3086	0.0870	1.0000										
ER1	0.0000	0.4361	0.1088	0.0000	0.4442	0.0600	0.0552	-0.0723	0.6647	0.0327	0.5167	0.0070	0.3505	0.5769	0.1179	-0.0570	1.0000									
ER2	0.0000	0.4534	0.1255	0.0000	0.4561	0.0640	0.0585	-0.0408	0.6891	0.0351	0.5090	0.0307	0.3734	0.5986	0.1256	-0.0479	0.9548	1.0000								
ER3	0.0000	0.3316	0.0615	0.0000	0.3673	0.0410	-0.0437	-0.1223	0.4940	0.0225	0.4640	0.0169	0.2395	0.4281	0.0806	-0.1496	0.9420	0.9607	1.0000							
ER4	0.0000	0.3611	0.0512	0.0000	0.3947	0.0437	-0.0781	-0.1011	0.4969	0.0250	0.4682	-0.0042	0.2550	0.4281	0.0858	-0.1184	0.8887	0.9514	0.9675	1.0000						
ES	0.0000	0.3993	0.0823	0.0000	0.4220	0.0519	-0.0158	-0.0886	0.5842	0.0289	0.4971	0.0092	0.3033	0.5057	0.1020	-0.0988	0.9569	0.9851	0.9881	0.9816	1.0000					
EM	0.0000	0.0000	0.0000	0.0000	0.0000	0.1240	-0.0077	0.0000	0.0000	0.0000	0.0865	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000				
ESE	0.0000	0.1784	0.3341	0.0000	0.4122	0.0869	-0.0738	0.2654	0.2692	0.1934	0.2568	0.0065	0.5072	0.2633	0.1706	0.1178	0.1691	0.1815	0.1163	0.1295	0.1494	0.0000	1.0000			
FP	0.0000	0.1440	0.2742	0.0000	0.2278	0.0470	0.0075	0.0942	0.2045	0.0264	0.1924	0.0568	0.2744	0.1868	0.0923	-0.0166	0.2949	0.3001	0.2917	0.2627	0.2888	0.0000	0.1363	1.0000		
FW	0.0000	0.2723	0.3789	0.0000	0.4008	0.1056	0.4838	0.3007	0.6949	0.0558	0.2734	0.1137	0.6165	0.6036	0.2074	0.2088	0.3464	0.3623	0.2269	0.1478	0.2539	0.0000	0.2884	0.2069	1.0000	
GE	0.0000	-0.0593	-0.0075	0.0000	0.0187	-0.0017	-0.0557	-0.0032	0.0000	0.0003	0.0097	-0.0198	-0.0098	-0.0667	-0.0033	-0.1203	-0.0222	0.0163	0.0000	-0.0858	-0.0368	0.0000	0.0015	0.0751	0.1567	1.0000
HGJ	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MC	0.0000	0.0000	0.0000	0.0850	0.0000	0.0353	-0.0022	0.0000	0.0000	0.0000	0.0246	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1307	0.0000	0.0000	0.0000	0.0000
NW	0.0000	0.0000	0.0000	0.4406	0.0000	0.1830	-0.0113	0.0000	0.0000	0.0000	0.1276	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6776	0.0000	0.0000	0.0000	0.0000
N	0.0000	0.4525	-0.2261	0.0007	0.1809	0.0264	0.2958	-0.0533	0.7141	0.0145	0.2346	-0.0176	0.1527	0.8011	0.0513	0.2373	0.4776	0.4853	0.3680	0.3578	0.4205	0.0010	0.0749	0.0472	0.3817	-0.0600
WL5	0.0000	0.5972	0.0835	0.0000	0.4133	0.0682	0.2482	0.0477	0.8978	0.0369	0.4101	0.0291	0.3982	0.7781	0.1339	0.1849	0.7017	0.7189	0.5502	0.5763	0.6414	0.0000	0.1909	0.1693	0.4124	0.0000
WL6	0.0000	0.5735	0.0572	0.0000	0.3766	0.0597	0.2140	0.0238	0.8499	0.0323	0.3837	0.0128	0.3483	0.6900	0.1171	0.1599	0.6675	0.6864	0.5208	0.5570	0.6138	0.0000	0.1672	0.1506	0.3566	0.0000
WL7	0.0000	0.5324	0.0273	0.0000	0.3268	0.0488	0.1712	-0.0025	0.7752	0.0265	0.3449	-0.0051	0.2850	0.5763	0.0959	0.1285	0.6126	0.6328	0.4750	0.5211	0.5674	0.0000	0.1372	0.1264	0.2868	0.0000
WL8	0.0000	0.5089	0.0100	0.0000	0.2967	0.0422	0.1442	-0.0163	0.7262	0.0230	0.3200	-0.0099	0.2466	0.5256	0.0829	0.1152	0.5751	0.5962	0.4450	0.4964	0.5359	0.0000	0.1191	0.1112	0.2331	0.0000
OLC	0.0000	0.1966	0.6353	0.0207	0.2501	0.1016	-0.0908	0.2203	0.2383	0.0572	0.3164	0.1474	0.5430	0.2315	0.1826	-0.0519	0.1754	0.1853	0.1349	0.1240	0.1528	0.0319	0.2958	0.2108	0.3297	0.0091
PW	0.0000	0.0200	0.1317	0.0508	0.1014	0.3582	-0.0064	0.0850	0.0361	0.0137	0.0684	0.0250	0.1312	0.0239	0.0441	-0.0107	0.0164	0.0197	0.0101	0.0094	0.0134	0.0782	0.0709	0.0414	0.0953	0.0168
Pre	0.0000	0.3103	0.3933	0.0000	0.3348	0.1091	0.4661	0.2913	0.6053	0.0548	0.2324	0.0177	0.6370	0.5577	0.2143	0.2343	0.3631	0.3524	0.2298	0.1768	0.2681	0.0000	0.2835	0.1969	0.7061	-0.0882
Pro	0.0000	-0.0066	0.5480	0.0000	0.3345	0.0744	-0.0898	0.3206	0.0205	0.0476	0.1531	0.1165	0.4347	-0.0471	0.1462	-0.1508	-0.0221	-0.0104	-0.0217	-0.0311	-0.0238	0.0000	0.2461	0.1511	0.3525	0.1025
ST	0.0000	0.0000	0.0000	0.0000	0.0000	0.1075	-0.0066	0.0000	0.0000	0.0000	0.0749	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Sent	0.0000	0.2384	0.1042	0.0000	0.1764	0.0376	0.2513	0.1526	0.4119	0.0193	0.1205	0.3857	0.2194	0.5053	0.0738	0.1698	0.1795	0.2164	0.1576	0.2463	0.2143	0.0000	0.0996	0.0311	0.3205	-0.0477
SLP	0.0000	0.1412	0.2210	0.0000	0.1070	0.0453	0.3153	0.1658	0.2540	-0.0068	0.0611	0.2409	0.2644	0.2475	0.0889	0.1417	0.1403	0.1417	0.0917	0.0665	0.1045	0.0000	-0.0353	0.0885	0.2962	-0.0383
SE	0.0000	0.0886	-0.1241	0.0000	-0.0598	-0.0183	0.1230	-0.1418	0.1742	-0.0126	0.0409	-0.0023	-0.1069	0.1510	-0.0360	0.0249	0.2596	-0.0171	0.0000	-0.1368	-0.0013	0.0000	-0.0649	-0.0192	0.0800	0.0000
TCS	0.0000	0.0995	0.2713	-0.0009	0.0937	0.0404	0.3115	0.2311	0.1784	0.0157	0.0012	0.4137	0.2383	0.1870	0.0801	0.2834	-0.0078	0.0058	-0.0462	-0.0537	-0.0311	-0.0014	0.0813	0.0540	0.2683	-0.0211
TW	0.0000	0.0122	0.0802	0.3543	0.0618	0.3524	-0.0122	0.0518	0.0220	0.0084	0.1353	0.0152	0.0799	0.0146	0.0269	-0.0065	0.0100	0.0120	0.0061	0.0057	0.0082	0.5448	0.0432	0.0252	0.0580	0.0102
UW	0.0000	0.3056	0.2291	0.0000	0.4957	0.0765	-0.0069	0.0150	0.5107	0.0430	0.5112	-0.0159	0.4469	0.4419	0.1503	-0.1203	0.8902	0.8957	0.9112	0.8235	0.8879	0.0000	0.2224	0.3531	0.4230	0.2501
V	0.0000	0.2144	0.4060	-0.0017	0.5029	0.1027	0.3718	0.3476	0.5394	0.0513	0.3067	0.2163	0.6039	0.6187	0.2031	0.2097	0.2738	0.2987	0.1808	0.2025	0.2385	-0.0027	0.2653	0.1605	0.4453	-0.0432
W	0.0000	0.4755	0.3508	0.0000	0.4800	0.1120	0.4907	0.2797	0.9013	0.0595	0.3663	0.1465	0.6536	0.8999	0.2198	0.2900	0.5182	0.5392	0.3680	0.3247	0.4270	0.0000	0.3078	0.2196	0.8909	0.0477
GA	0.0000	0.1239	0.0508	0.0000	0.0487	0.0167	0.1179	0.0368	0.1237	0.0078	0.0431	0.0243	0.0974	0.1858	0.0328	0.1572	0.0955	0.0602	0.0523	0.1302	0.0966	0.0000	0.0403	-0.0437	-0.0355	-0.9738
S	0.0000	-0.0262	0.1651	0.0000	0.1189	0.0320	-0.0579	0.1734	-0.0559	0.0197	0.0064	0.0207	0.1866	-0.0335	0.0628	0.0123	-0.1874	0.0847	0.0465	0.1742	0.0551	0.0000	0.1021	0.0464	0.0345	0.0060
TF	0.0000	0.1843	0.2106	0.0000	0.1847	0.0395	0.0986	0.1097	0.2734	0.0214	0.1462	0.1902	0.2306	0.2978	0.0776	0.0632	0.2301	0.2540	0.2441	0.2612	0.2554	0.0000	0.1106	0.0975	0.2166	-0.0183
T	0.0000	0.2295	0.1983	0.0000	0.1143	0.0368	0.2452	0.1513	0.3532	0.0168	0.0828	0.4431	0.2150	0.3471	0.0723	0.1262	0.2098	0.2282	0.1615	0.1537	0.1863	0.0000	0.0869	0.0865	0.2642	-0.0082
VC	0.0000	0.0430	0.0297	0.0000	0.0280	0.0088	0.0421	0.0144	0.0532	0.0044	0.0244	-0.0243	0.0511	0.0614	0.0172	0.0481	0.0441	0.0307	0.0230	0.0326	0.0337	0.0000	0.0227	-0.0008	0.0168	-0.2841
VU	0.0000	-0.0402	-0.0278	0.0000	0.1701	-0.0032	-0.3994	-0.2025	-0.1439	0.0010	0.2735	-0.1591	-0.0187	-0.2166	-0.0063	-0.3687	0.5601	0.5491	0.6977	0.6357	0.6270	0.0000	0.0051	0.2115	-0.2242	0.2349

	HGJ	MC	NW	N	WL5	WL6	WL7	WL8	OLC	PW	Pre	Pro	ST	Sent	SLP	SE	TCS	TW	UW	V	W	GA	S	TF	T	VC	VU
HGJ	1.0000																										
MC	0.0000	1.0000																									
NW	0.0000	0.1928	1.0000																								
N	0.0000	0.0003	0.0015	1.0000																							
WL5	0.0000	0.0000	0.0000	0.6516	1.0000																						
WL6	0.0000	0.0000	0.0000	0.5721	0.9466	1.0000																					
WL7	0.0000	0.0000	0.0000	0.4709	0.8634	0.9415	1.0000																				
WL8	0.0000	0.0000	0.0000	0.4224	0.8088	0.8820	0.9368	1.0000																			
OLC	0.0000	0.0091	0.0471	-0.0684	0.1313	0.1086	0.0813	0.0636	1.0000																		
PW	0.2875	0.2916	0.1154	-0.0542	0.0033	0.0007	-0.0022	-0.0060	0.1028	1.0000																	
Pre	0.0000	0.0000	0.0000	0.3546	0.4177	0.3691	0.3068	0.2630	0.3567	0.0682	1.0000																
Pro	0.0000	0.0000	0.0000	-0.3723	-0.1153	-0.1139	-0.1094	-0.1188	0.3900	0.1946	0.1860	1.0000															
ST	0.0000	0.1132	0.5872	0.0009	0.0000	0.0000	0.0000	0.0000	0.0276	0.0678	0.0000	0.0000	1.0000														
Sent	0.0000	0.0000	0.0000	0.3859	0.2865	0.2378	0.1790	0.1574	0.0729	0.0308	0.0834	0.1070	0.0000	1.0000													
SLP	0.0000	0.0000	0.0000	0.1385	0.1663	0.1439	0.1159	0.0981	0.0318	0.0108	0.3859	-0.0261	0.0000	0.1221	1.0000												
SE	0.0000	0.0000	0.0000	0.2094	0.1941	0.1837	0.1676	0.1570	-0.0786	-0.0267	0.0811	-0.1213	0.0000	0.0556	0.0304	1.0000											
TCS	0.0000	-0.0004	-0.0021	0.0140	0.0725	0.0550	0.0345	0.0223	0.1172	0.0538	0.1981	0.2377	-0.0012	0.2411	0.1650	-0.0028	1.0000										
TW	0.1751	0.2622	0.8040	-0.0319	0.0020	0.0004	-0.0013	-0.0036	0.0972	0.6777	0.0415	0.1185	0.4721	0.0187	0.0066	-0.0163	0.0312	1.0000									
UW	0.0000	0.0000	0.0000	0.3046	0.4971	0.4536	0.3943	0.3564	0.2651	0.0475	0.3808	0.1283	0.0000	0.0323	0.1416	0.0199	-0.0043	0.0289	1.0000								
V	0.0000	-0.0008	-0.0039	0.1759	0.4006	0.3387	0.2629	0.2301	0.2670	0.0903	0.3511	0.3303	-0.0023	0.4147	0.1659	-0.0033	0.2755	0.0521	0.2534	1.0000							
W	0.0000	0.0000	0.0000	0.6652	0.6689	0.5881	0.4852	0.4269	0.3123	0.0657	0.7040	0.1660	0.0000	0.4632	0.3030	0.1298	0.2533	0.0400	0.4831	0.5961	1.0000						
GA	0.0000	0.0000	0.0000	0.1496	0.0950	0.0849	0.0716	0.0638	0.0315	-0.0083	0.1836	-0.0819	0.0000	0.0958	0.0758	0.0185	0.0486	-0.0050	-0.1766	0.1191	0.0865	1.0000					
S	0.0000	0.0000	0.0000	-0.1200	-0.1046	-0.1047	-0.1019	-0.0990	0.1160	0.0343	0.0099	0.1391	0.0000	0.0043	0.0086	-0.9576	0.0347	0.0209	0.0416	0.0785	-0.0002	-0.0071	1.0000				
TF	0.0000	0.0000	0.0000	0.1657	0.2138	0.1869	0.1529	0.1370	0.1508	0.0340	0.1346	0.1224	0.0000	0.4028	0.0959	-0.0003	0.1268	0.0207	0.1780	0.2492	0.2881	0.0518	0.0366	1.0000			
T	0.0000	0.0000	0.0000	0.2142	0.2928	0.2711	0.2403	0.2231	0.1319	0.0216	0.1966	0.0544	0.0000	0.2879	0.1788	0.0535	0.2340	0.0131	0.1503	0.2626	0.3423	0.0862	-0.0089	0.1639	1.0000		
VC	0.0000	0.0000	0.0000	0.0470	0.0418	0.0385	0.0338	0.0298	0.0242	-0.0012	0.0966	-0.0243	0.0000	-0.0544	0.0312	0.0081	0.0059	-0.0007	-0.0224	0.0306	0.0442	0.3739	-0.0023	-0.0106	0.0066	1.0000	
VU	0.0000	0.0000	0.0000	-0.1794	0.0183	0.0328	0.0475	0.0511	0.0469	0.0008	-0.1260	0.0112	0.0000	-0.3302	-0.0853	-0.0788	-0.2155	0.0005	0.7090	-0.1918	-0.2460	-0.2610	0.0457	-0.0340	-0.1668	-0.0574	1.0000

Table A-3: Covariance Matrix for PC Search over All Variables – All Events

	CT	Adj	Adv	AA	AAS	AB	Art	CCP	C	CSR	Conj	CI	Conn	CW	D	ERC	ER1	ER2	ER3	ER4	ES	EM	ESE	FP	FW	GE
CT	3.3057E+12																									
Adj	2.2069E+07	735.2698																								
Adv	7.3072E+06	125.4810	254.2163																							
AA	35774.8076	0.5544	0.8764	0.5310																						
AAS	1.4336E+07	407.4159	114.7012	0.4212	622.8685																					
AB	1.5434E+05	3.4296	2.1605	0.1791	2.6084	4.8247																				
Art	1.8166E+07	163.5599	197.5496	0.8148	108.3774	2.0730	1646.9407																			
CCP	3.1556E+06	35.6173	90.3443	0.4570	0.0000	0.4891	91.0996	117.1329																		
C	9.9510E+08	24461.1864	7339.0110	32.4339	17603.3885	154.0177	12992.2037	3068.2738	1.6882E+06																	
CSR	1.5668E+05	2.7407	2.0479	0.0145	2.4980	0.0478	3.2765	0.6149	176.5569	0.8921																
Conj	5.6033E+06	141.3965	59.1386	0.2097	177.1254	1.1422	58.0545	0.0000	9762.4756	1.3557	159.7127															
CI	1.7258E+05	3.6007	1.3780	0.0116	3.0414	0.0022	0.6977	0.9597	271.6293	0.0491	1.9487	0.1901														
Conn	3.0100E+07	645.3549	407.1144	1.3785	791.6416	7.1209	757.7056	140.1959	32327.8219	6.1697	252.4502	4.2050	1781.2637													
CW	7.4032E+07	661.2956	546.6698	0.2739	364.1713	9.1888	6247.7092	0.0000	53003.8685	5.1358	328.3708	-20.3237	2974.0610	83641.5974												
D	1.0489E+06	22.0012	15.9624	0.0467	18.1129	0.7387	26.0266	6.0331	1296.1616	0.3101	7.4600	0.0624	62.3585	95.3762	13.6240											
ERC	2.8204E+06	31.7589	20.5634	0.1232	42.6183	-0.3664	188.2041	13.0006	682.7729	0.1517	-6.8022	0.2817	111.2896	983.1077	-0.3294	79.0149										
ER1	6.6319E+05	17.4616	3.3875	0.0107	11.7380	0.1284	7.4525	0.0000	1289.1739	0.1186	10.5841	0.2517	19.7001	52.1143	0.8062	-1.0603	1.6797									
ER2	6.1849E+05	16.7692	3.3491	0.0110	10.9488	0.1187	9.7220	0.0000	1247.9888	0.1153	9.8725	0.2453	19.2589	67.0935	0.7645	-0.4760	1.5668	1.5261								
ER3	6.1403E+05	17.0359	1.7181	0.0147	11.7998	0.0361	4.4429	0.0000	1334.4041	0.1247	10.6398	0.4410	13.6284	0.0000	0.2723	-0.0002	1.6886	1.6297	2.1818							
ER4	6.7850E+05	17.1103	1.2029	0.0168	12.1264	-0.0152	9.9658	0.0000	1393.0397	0.1272	10.9343	0.4960	12.9884	20.8219	0.0263	1.9159	1.7353	1.7056	2.4200	2.8311						
ES	3.2527E+06	85.4873	10.8604	0.0699	58.7395	0.2528	41.5490	0.0000	6657.6461	0.6129	52.9650	1.9300	78.5641	160.8516	1.8955	2.2953	8.4058	8.1339	10.3401	11.5233	49.9264					
EM	12645.5431	0.2933	0.1699	0.0487	0.0878	0.2229	-0.0446	0.0000	9.1780	0.0085	0.0791	-0.0015	0.2942	0.0000	0.0252	-0.1171	0.0126	0.0111	0.0000	-0.0084	0.0067	1.2921				
ESE	4.1130E+05	8.9561	16.0011	0.0605	9.4455	0.2335	10.4920	1.3928	501.0474	0.4737	7.3040	0.0411	40.8507	45.0236	1.2351	-1.1938	0.4022	0.3855	0.1047	-0.0469	0.7985	0.0232	7.4697			
FP	7.8212E+05	20.4950	11.3974	0.0755	6.9151	0.9533	-13.3169	0.0000	723.0526	0.1108	6.2353	-0.1902	22.7727	0.0000	1.9875	-9.1653	0.9896	0.8713	0.0000	-0.6655	0.5299	0.1804	1.6435	11.5184		
FW	5.7007E+07	509.2210	420.9551	0.2109	280.4248	7.0757	4810.9574	0.0000	40814.8563	3.9548	252.8571	-15.6500	2290.1323	0.0000	73.4431	757.0277	40.1298	51.6644	0.0000	16.0336	123.8615	0.0000	34.6698	0.0000	64406.9927	
GE	4.6264E+05	26.5873	20.3216	0.0601	19.3719	0.7940	-23.8718	0.0510	1435.0421	0.2994	17.3664	-0.3147	49.7817	0.1628	3.7866	-18.4513	1.8101	1.5947	0.0087	-1.2030	1.0075	0.1328	3.5676	7.6242	0.1254	100.9442
HGJ	1.6244E+05	-0.6656	0.2780	0.0707	-0.2877	0.4221	17.9953	0.0001	-16.2530	0.0099	-0.2596	0.0101	2.5101	92.6694	-0.1509	3.6095	-0.0412	-0.0166	0.0000	0.1551	0.2525	0.1052	-0.0930	-0.4432	71.3588	-1.2826
MC	10373.5559	0.2371	0.2219	0.0194	0.1933	0.0631	0.0838	0.0332	11.9329	0.0240	0.1083	0.0017	0.4014	0.1062	0.0173	-0.0260	0.0084	0.0080	0.0057	0.0036	0.0294	0.0242	0.0263	0.0556	0.0818	0.1561
NW	33184.4525	0.4049	0.3086	0.5406	0.0000	0.9683	0.4136	0.0000	0.0000	0.0119	0.0000	0.0031	0.0667	0.0000	0.0000	-0.0011	0.0000	0.0000	0.0000	0.0000	0.0000	0.2626	0.0281	0.4174	0.0000	0.0529
N	5.7309E+07	1477.8950	400.6625	1.4772	1409.6761	9.2545	690.6674	117.9253	61264.9174	8.4262	282.1687	8.2587	2307.4769	2443.4068	74.8842	192.5233	36.6906	34.1057	30.6748	29.5826	160.6363	0.4298	25.4504	33.8621	1881.5098	65.5577
WL5	6.9157E+07	2162.4718	419.0509	1.6831	2163.3843	9.9055	552.1273	89.9431	76838.8119	9.0273	294.6137	11.5270	2991.9134	1578.0962	76.4797	340.3425	39.0847	37.8258	40.4284	42.1506	201.6402	0.2814	20.2011	22.1720	1215.1900	44.3009
WL6	5.3529E+07	1910.9223	273.6700	0.9848	2226.3754	7.1733	229.6962	0.0000	35380.0137	5.1995	0.0000	4.6522	2759.0244	0.0000	51.3716	437.8138	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	7.0231	0.0000	0.0000	0.5854
WL7	3.1440E+07	999.9125	153.7727	0.6904	1056.7357	3.9157	145.6333	-8.6373	23230.5415	3.5750	-28.2920	2.4661	1402.9260	-151.5457	37.1004	215.9728	-3.7533	-3.6324	-3.8824	-4.0478	-19.3637	-0.0270	4.1362	-2.1292	-116.6956	-3.7478
WL8	1.3902E+07	221.3450	69.8536	0.5299	0.0000	1.4353	123.4735	0.0000	19038.0334	2.6956	0.0000	1.6854	223.1180	0.0000	27.6431	-0.0064	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	3.7791	0.0000	0.0000	0.3150
OLC	4.7919E+05	10.5530	49.7650	0.1806	34.7032	0.1399	4.3530	0.0000	551.4798	0.7083	31.2917	0.3319	57.6307	0.0000	0.8007	-0.2057	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	7.6805	0.0000	0.0000	10.1171
PW	82054.3488	2.0691	1.1650	0.1142	0.6743	1.2724	-0.6468	0.0000	70.5033	0.0353	0.6080	-0.0138	2.2310	0.0000	0.1938	-0.8962	0.0965	0.0850	0.0000	-0.0649	0.0517	0.2709	0.1652	1.1960	0.0000	0.8684
Pre	1.2841E+07	264.2238	248.4016	0.3781	129.7989	7.9101	621.3116	54.7543	15995.4451	2.7674	117.0388	-2.6750	940.1826	3017.8536	59.9482	-62.7068	18.5747	17.7600	0.0000	-8.7955	18.7436	0.8051	31.3030	63.4279	2323.8542	115.6345
Pro	5.7499E+06	123.8483	149.0496	1.3805	144.5622	0.6673	79.5935	29.7955	7700.3551	2.2112	71.2358	2.1188	265.8585	95.1885	7.2998	16.6762	3.4382	3.5775	5.1025	6.0123	24.1427	-0.0225	16.4918	-1.7711	73.2985	9.3216
ST	10906.5127	0.1331	0.1014	0.1782	0.0000	0.3182	0.1996	0.0000	0.0000	0.0048	0.0000	0.0015	0.0219	0.0000	0.0000	-0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.1699	0.0092	0.1372	0.0000	0.0219
Sent	1.1471E+07	174.7048	75.6905	0.3656	148.0207	1.0893	154.0150	41.1023	11438.3854	2.1222	60.4479	2.4362	301.1263	507.1047	11.3972	22.0846	6.4841	6.3622	6.9407	7.6984	35.1838	0.0196	3.7937	1.5467	390.4886	4.1884
SLP	4.1250E+06	-14.6491	37.4428	0.2038	-11.0250	-0.5712	339.5806	26.5488	22.2849	0.2994	-9.9412	0.7055	126.5741	1463.2697	0.7910	67.6101	-1.5777	-0.7076	0.0000	2.8528	3.4203	-0.1732	-0.5769	-13.6431	1126.7695	-24.8480
SE	3.1681E+06	7.6349	0.0568	-0.0375	25.1072	0.0746	4.2916	0.0000	398.9856	-0.0398	22.6390	-0.5465	27.8241	0.0000	0.5793	0.0005	3.5929	0.0000	0.0000	-0.5396	2.5137	0.0000	-0.2677	0.0000	0.0000	-0.0223
TCS	2.0675E+06	18.0282	61.7092	0.4216	0.0000	0.3205	82.5752	59.3822	1555.5055	0.4015	0.0000	0.8183	93.5552	0.0000	3.8652	13.2367	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.1067	0.0000	0.0000	0.0259
TW	1.1240E+05	2.4557	1.4548	0.6459	0.7021	2.2455	-0.2341	0.0053	72.1362	0.0467	0.6201	-0.0109	2.3645	0.0990	0.2017	-0.9008	0.0978	0.0862	0.0004	-0.0650	0.0544	0.5252	0.1919	1.5670	0.0762	0.9253
UW	1.3023E+07	884.9587	394.4710	-0.1271	458.4657	24.5586	-905.7999	0.0000	47937.8598	4.8659	413.3955	-12.8086	1431.2003	0.0000	131.7715	-607.5714	65.6081	57.7656	0.0000	-44.1198	35.1341	3.4933	76.1159	275.2095	0.0000	501.5788
V	3.2723E+07	693.4593	348.4426	1.7253	692.7741	4.9328	385.1914	214.7483	40257.7999	6.0371	197.0045	6.4749	1313.9407	405.0429	53.0643	71.2314	13.0554	12.1777	13.1241	13.4874	65.3319	0.0976	16.0473	7.6912	311.8973	21.8885
W	1.3104E+08	1170.5165	967.6249	0.4848	644.5961	16.2645	11058.6665	0.0000	93818.7248	9.0906	581.2279	-35.9737	5264.1933	83641.5974	168.8193	1740.1354	92.2441	118.7579	0.0000	36.8556	284.7131	0.0000	79.6934	0.0000	64406.9927	0.2882
GA	54399.7221	-0.9923	-0.6607	-0.0036	-0.8579	-0.0412	8.7341	-0.0034	-33.7178	-0.0128	-0.7669	0.0060	0.3884	54.1528	-0.1400	2.8568	-0.0594	-0.0281	-0.0006	0.1031	0.1180	-0.0087	-0.1833	-0.5019	41.6996	-6.6452
S	-92895.8273	0.8498	1.4310	0.0079	-1.2310	0.0145	13.5431	0.0000	85.4219	0.0150	-1.1100	-0.0146	4.4229	105.0677	0.1650	2.1858	-0.1762	0.1492	0.0000	0.1420	0.2571	0.0000	0.1519	0.0000	80.9058	0.0047
TF	1.8842E+07	382.8383	-7.0151	0.6648	312.7678	-3.7797	229.7738	20.9331	33294.7648	3.6804	244.8314	18.1746	170.8582	-968.1050	-16.4120	111.9903	37.2725	36.9129	67.8172	82.8465	307.6957	-0.7210	-11.9282	-56.7978	-745.4751	-102.5499
T	5.0395E+06	104.5270	40.2865	0.3391	87.6894	0.0434	31.5859	27.8553	7928.3604	1.4298	56.3645	5.5781	124.5301	-513.4747	1.7766	10.8206	7.3128	7.2243	12.9010	14.6181	56.6744	-0.0500	1.1534	-5.8612	-395.3937	-12.7254
VC	-1.1043E+05	-3.4629	-0.7115	-0.0080	-3.7574	-0.0271	18.4478	-1.1028	-180.3013	-0.0365	-1.4285	-0.0228	0.2524	150.8196	-0.1209	3.7965	-0.1086	0.0043	-0.1862	-0.0503	-0.3910	-0.0054	-0.0543	-0.3222	116.1364	-3.8267
VU	-3.4427E+06	44.9241	3.9812	-0.0294	21.8964	1.8316	-487.9274	0.0000	1336.9332	0.1516	19.7439	0.0363	-48.8731	-2906.1586	6.8866	-122.5251	3.1335	1.4053	0.0000	-5.6659	-6.7930	0.3439	4.6245	27.0961	-2237.8450	49.3736

	HGJ	MC	NW	N	WL5	WL6	WL7	WL8	OLC	PW	Pre	Pro	ST	Sent	SLP	SE	TCS	TW	UW	V	W	GA	S	TF	T	VC	VU
HGJ	3.0401																										
MC	0.0278	0.0729																									
NW	0.3819	0.0345	3.1418																								
N	-0.1613	0.5464	0.0000	6945.0611																							
WL5	-0.5970	0.6387	0.0000	8431.4334	13779.7386																						
WL6	0.0013	0.3819	0.0000	8833.1513	15329.1749	19431.3706																					
WL7	0.0582	0.2691	0.0000	4534.1725	7693.1061	9564.1247	6248.4518																				
WL8	0.0003	0.2055	0.0000	592.0358	561.3808	0.0000	2633.1534	4865.0523																			
OLC	0.0000	0.0700	0.0000	17.1497	16.2617	0.0000	-1.5616	0.0000	53.3059																		
PW	0.5517	0.0870	0.5890	3.3018	2.1619	0.0000	-0.2076	0.0000	0.0000	1.8135																	
Pre	-5.0074	0.2857	0.0000	977.1233	483.4125	0.0000	-46.4224	0.0000	0.0000	6.1847	1903.9359																
Pro	0.4022	0.4909	0.0000	349.4915	480.5760	342.2845	249.9814	184.1838	62.7633	-0.1727	-6.6058	447.3190															
ST	0.1854	0.0143	1.0326	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1936	0.0000	0.0000	0.8034														
Sent	0.7957	0.0925	0.0000	594.5583	644.2029	422.8048	303.9312	227.5119	6.5904	0.1508	93.2145	76.4678	0.0000	169.4514													
SLP	5.3762	-0.0455	0.0000	35.1931	-2.7485	0.0000	0.2639	0.0000	0.0000	-1.3303	1.2367	15.7394	0.0000	28.6542	146.5198												
SE	1.7448	-0.0146	0.0000	110.1075	11.7650	0.0000	-1.1298	0.0000	0.0000	0.0000	0.0000	-13.0459	0.0000	4.7680	0.0000	235.0436											
TCS	0.0001	0.0169	0.0000	71.2898	45.3260	0.0000	-4.3527	0.0000	0.0000	0.0000	55.7463	15.1622	0.0000	23.1045	27.0298	0.0000	60.4581										
TW	0.9167	0.1194	3.6801	3.4003	2.2678	0.0773	-0.1654	0.0155	0.0015	2.3602	6.2676	-0.1654	1.2107	0.1625	-1.3360	0.0008	0.0035	5.9813									
UW	-48.3767	1.1407	0.0000	2245.0349	1469.9848	0.0000	-141.1637	0.0000	0.0000	26.8351	4205.2228	-117.4197	0.0000	102.5450	-904.5259	0.0000	0.0000	27.0897	18246.1928								
V	-0.3196	0.4384	0.0000	2252.2131	3014.9704	2476.2454	1852.8225	1332.4710	38.5980	0.7500	244.7514	361.0479	0.0000	419.8839	36.4114	27.9250	108.8697	0.8028	509.9201	2217.8302							
W	164.0282	0.1880	0.0000	4324.9166	2793.2862	0.0000	-268.2413	0.0000	0.0000	0.0000	5341.7077	168.4870	0.0000	897.5932	2590.0392	0.0000	0.0000	0.1752	0.0000	716.9402	1.4805E+05						
GA	0.1907	-0.0102	-0.0035	-1.5151	-1.1076	-0.0385	0.0730	-0.0207	-0.6660	-0.0572	-4.1533	-0.5046	-0.0014	0.3055	3.3130	0.0015	-0.0017	-0.0608	-33.0199	-0.9767	95.8523	2.1682					
S	0.0642	0.0031	0.0000	-3.5168	2.5526	0.0000	-0.2451	0.0000	0.0000	0.0000	6.7101	2.7337	0.0000	0.7400	3.2535	-19.1045	0.0000	0.0002	0.0000	-1.3692	185.9735	0.1201	3.6236				
TF	7.9223	-0.0163	0.0000	655.4488	1157.4015	215.3314	75.1504	115.8701	3.3564	-5.5382	-913.4678	220.5867	0.0000	253.5446	164.5698	-13.3972	11.7670	-5.5769	-3765.6551	477.8675	-1713.5801	5.6414	-2.8691	3150.6747			
T	0.4748	0.0444	0.0882	238.0047	334.5279	133.1317	69.7800	47.6024	9.3533	-0.4378	-77.7123	62.2443	0.0428	70.8789	24.0636	-20.3638	23.8516	-0.3539	-394.3600	186.2395	-908.8685	1.3656	0.7717	533.2239	288.7514		
VC	0.3243	-0.0067	-0.0019	-11.3572	-13.5574	-11.3460	-8.5142	-6.1053	-0.5490	-0.0360	3.7453	-0.9859	-0.0008	-2.9943	5.2385	-3.3711	-0.6198	-0.0381	-21.2153	-11.0284	266.9560	1.3581	1.7331	-7.3028	4.1214	12.8472	
VU	-10.6774	0.1057	0.0000	65.0930	44.0109	0.0000	-4.2264	0.0000	0.0000	2.6421	221.4231	-17.4418	0.0000	-22.2686	-182.4462	0.0000	0.0000	2.6608	1796.4534	24.3539	-5144.0036	-6.5814	-6.4617	-309.2053	-5.8436	-11.3327	362.3513

Table A-4: Correlation Matrix for PC Search over All Variables – All Events

	CT	Adj	Adv	AA	AAS	AB	Art	CCP	C	CSR	Conj	CI	Conn	CW	D	ERC	ER1	ER2	ER3	ER4	ES	EM	ESE	FP	FW	GE
CT	1.0000																									
Adj	0.4476	1.0000																								
Adv	0.2521	0.2902	1.0000																							
AA	0.0270	0.0281	0.0754	1.0000																						
AAS	0.3159	0.6020	0.2882	0.0232	1.0000																					
AB	0.0386	0.0576	0.0617	0.1119	0.0476	1.0000																				
Art	0.2462	0.1486	0.3053	0.0276	0.1070	0.0233	1.0000																			
CCP	0.1604	0.1214	0.5236	0.0579	0.0000	0.0206	0.2074	1.0000																		
C	0.4212	0.6943	0.3543	0.0343	0.5429	0.0540	0.2464	0.2182	1.0000																	
CSR	0.0912	0.1070	0.1360	0.0210	0.1060	0.0230	0.0855	0.0602	0.1439	1.0000																
Conj	0.2439	0.4126	0.2935	0.0228	0.5616	0.0411	0.1132	0.0000	0.5945	0.1136	1.0000															
CI	0.2177	0.3046	0.1982	0.0366	0.2795	0.0023	0.0394	0.2034	0.4795	0.1193	0.3537	1.0000														
Conn	0.3923	0.5639	0.6050	0.0448	0.7516	0.0768	0.4424	0.3069	0.5895	0.1548	0.4733	0.2285	1.0000													
CW	0.1408	0.0843	0.1186	0.0013	0.0505	0.0145	0.5323	0.0000	0.1411	0.0188	0.0898	-0.1612	0.2437	1.0000												
D	0.1563	0.2198	0.2712	0.0174	0.1966	0.0911	0.1738	0.1510	0.2703	0.0890	0.1599	0.0388	0.4003	0.0893	1.0000											
ERC	0.1745	0.1318	0.1451	0.0190	0.1921	-0.0188	0.5217	0.1351	0.0591	0.0181	-0.0606	0.0727	0.2966	0.3824	-0.0100	1.0000										
ER1	0.2814	0.4969	0.1639	0.0113	0.3629	0.0451	0.1417	0.0000	0.7656	0.0969	0.6462	0.4455	0.3601	0.1390	0.1685	-0.0920	1.0000									
ER2	0.2754	0.5006	0.1700	0.0122	0.3551	0.0437	0.1939	0.0000	0.7775	0.0988	0.6324	0.4554	0.3694	0.1878	0.1677	-0.0433	0.9786	1.0000								
ER3	0.2286	0.4253	0.0730	0.0136	0.3201	0.0111	0.0741	0.0000	0.6953	0.0894	0.5700	0.6848	0.2186	0.0000	0.0499	0.0000	0.8820	0.8931	1.0000							
ER4	0.2218	0.3750	0.0448	0.0137	0.2888	-0.0041	0.1459	0.0000	0.6372	0.0800	0.5142	0.6761	0.1829	0.0428	0.0042	0.1281	0.7958	0.8206	0.9737	1.0000						
ES	0.2532	0.4462	0.0964	0.0136	0.3331	0.0163	0.1449	0.0000	0.7252	0.0918	0.5931	0.6265	0.2634	0.0787	0.0727	0.0365	0.9179	0.9318	0.9907	0.9692	1.0000					
EM	0.0061	0.0095	0.0094	0.0588	0.0031	0.0893	-0.0010	0.0000	0.0062	0.0079	0.0055	-0.0031	0.0061	0.0000	0.0060	-0.0116	0.0085	0.0079	0.0000	-0.0044	0.0008	1.0000				
ESE	0.0828	0.1208	0.3672	0.0304	0.1385	0.0389	0.0946	0.0471	0.1411	0.1835	0.2115	0.0345	0.3541	0.0570	0.1224	-0.0491	0.1135	0.1142	0.0259	-0.0102	0.0413	0.0075	1.0000			
FP	0.1268	0.2227	0.2106	0.0305	0.0816	0.1279	-0.0967	0.0000	0.1640	0.0346	0.1454	-0.1285	0.1590	0.0000	0.1587	-0.3038	0.2250	0.2078	0.0000	-0.1165	0.0221	0.0468	0.1772	1.0000		
FW	0.1235	0.0740	0.1040	0.0011	0.0443	0.0127	0.4671	0.0000	0.1238	0.0165	0.0788	-0.1414	0.2138	0.0000	0.0784	0.3356	0.1220	0.1648	0.0000	0.0375	0.0691	0.0000	0.0500	0.0000	1.0000	
GE	0.0253	0.0976	0.1269	0.0082	0.0773	0.0360	-0.0585	0.0005	0.1099	0.0316	0.1368	-0.0719	0.1174	0.0001	0.1021	-0.2066	0.1390	0.1285	0.0006	-0.0712	0.0142	0.0116	0.1299	0.2236	0.0000	1.0000
HGJ	0.0512	-0.0141	0.0100	0.0557	-0.0066	0.1102	0.2543	0.0000	-0.0072	0.0060	-0.0118	0.0133	0.0341	0.1838	-0.0234	0.2329	-0.0182	-0.0077	0.0000	0.0529	0.0205	0.0531	-0.0195	-0.0749	0.1613	-0.0732
MC	0.0211	0.0324	0.0516	0.0985	0.0287	0.1063	0.0076	0.0114	0.0340	0.0939	0.0317	0.0143	0.0352	0.0014	0.0174	-0.0108	0.0240	0.0240	0.0143	0.0080	0.0154	0.0789	0.0356	0.0607	0.0012	0.0576
NW	0.0103	0.0084	0.0109	0.4186	0.0000	0.2487	0.0057	0.0000	0.0000	0.0071	0.0000	0.0040	0.0009	0.0000	0.0000	-0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.1303	0.0058	0.0694	0.0000	0.0030
N	0.3782	0.6540	0.3015	0.0243	0.6778	0.0506	0.2042	0.1307	0.5658	0.1071	0.2679	0.2273	0.6560	0.1014	0.2434	0.2599	0.3397	0.3313	0.2492	0.2110	0.2728	0.0045	0.1117	0.1197	0.0890	0.0783
WL5	0.3240	0.6794	0.2239	0.0197	0.7384	0.0384	0.1159	0.0708	0.5038	0.0814	0.1986	0.2252	0.6039	0.0465	0.1765	0.3262	0.2569	0.2608	0.2332	0.2134	0.2431	0.0021	0.0630	0.0557	0.0408	0.0376
WL6	0.2112	0.5056	0.1231	0.0097	0.6400	0.0234	0.0406	0.0000	0.1953	0.0395	0.0000	0.0765	0.4690	0.0000	0.0998	0.3533	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0184	0.0000	0.0000	0.0004
WL7	0.2188	0.4665	0.1220	0.0120	0.5357	0.0226	0.0454	-0.0101	0.2262	0.0479	-0.0283	0.0716	0.4205	-0.0066	0.1272	0.3074	-0.0366	-0.0372	-0.0333	-0.0304	-0.0347	-0.0003	0.0191	-0.0079	-0.0058	-0.0047
WL8	0.1096	0.1170	0.0628	0.0104	0.0000	0.0094	0.0436	0.0000	0.2101	0.0409	0.0000	0.0554	0.0758	0.0000	0.1074	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0198	0.0000	0.0000	0.0004
OLC	0.0361	0.0533	0.4275	0.0339	0.1905	0.0087	0.0147	0.0000	0.0581	0.1027	0.3391	0.1043	0.1870	0.0000	0.0297	-0.0032	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3849	0.0000	0.0000	0.1379
PW	0.0335	0.0567	0.0543	0.1164	0.0201	0.4302	-0.0118	0.0000	0.0403	0.0278	0.0357	-0.0236	0.0393	0.0000	0.0390	-0.0749	0.0553	0.0511	0.0000	-0.0286	0.0054	0.1770	0.0449	0.2617	0.0000	0.0642
Pre	0.1619	0.2233	0.3570	0.0119	0.1192	0.0825	0.3509	0.1159	0.2821	0.0671	0.2122	-0.1406	0.5105	0.2391	0.3722	-0.1617	0.3285	0.3295	0.0000	-0.1198	0.0608	0.0162	0.2625	0.4283	0.2099	0.2638
Pro	0.1495	0.2160	0.4420	0.0896	0.2739	0.0144	0.0927	0.1302	0.2802	0.1107	0.2665	0.2298	0.2978	0.0156	0.0935	0.0887	0.1254	0.1369	0.1633	0.1689	0.1616	-0.0009	0.2853	-0.0247	0.0137	0.0439
ST	0.0067	0.0055	0.0071	0.2728	0.0000	0.1616	0.0055	0.0000	0.0000	0.0057	0.0000	0.0038	0.0006	0.0000	0.0000	-0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.1667	0.0038	0.0451	0.0000	0.0024
Sent	0.4847	0.4949	0.3647	0.0385	0.4556	0.0381	0.2915	0.2917	0.6763	0.1726	0.3674	0.4293	0.5481	0.1347	0.2372	0.1909	0.3843	0.3956	0.3610	0.3515	0.3825	0.0013	0.1066	0.0350	0.1182	0.0320
SLP	0.1874	-0.0446	0.1940	0.0231	-0.0365	-0.0215	0.6913	0.2027	0.0014	0.0262	-0.0650	0.1337	0.2478	0.4180	0.0177	0.6284	-0.1006	-0.0473	0.0000	0.1401	0.0400	-0.0126	-0.0174	-0.3321	0.3668	-0.2043
SE	0.1137	0.0184	0.0002	-0.0034	0.0656	0.0022	0.0069	0.0000	0.0200	-0.0027	0.1168	-0.0818	0.0430	0.0000	0.0102	0.0000	0.1808	0.0000	0.0000	-0.0209	0.0232	0.0000	-0.0064	0.0000	0.0000	-0.0001
TCS	0.1462	0.0855	0.4978	0.0744	0.0000	0.0188	0.2617	0.7057	0.1540	0.0547	0.0000	0.2414	0.2851	0.0000	0.1347	0.1915	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0521	0.0000	0.0000	0.0003
TW	0.0253	0.0370	0.0373	0.3624	0.0115	0.4180	-0.0024	0.0002	0.0227	0.0202	0.0201	-0.0103	0.0229	0.0001	0.0223	-0.0414	0.0309	0.0285	0.0001	-0.0158	0.0031	0.1889	0.0287	0.1888	0.0001	0.0377
UW	0.0530	0.2416	0.1832	-0.0013	0.1360	0.0828	-0.1652	0.0000	0.2731	0.0381	0.2422	-0.2175	0.2510	0.0000	0.2643	-0.5060	0.3748	0.3462	0.0000	-0.1941	0.0368	0.0228	0.2062	0.6003	0.0000	0.3696
V	0.3822	0.5430	0.4641	0.0503	0.5894	0.0477	0.2015	0.4213	0.6579	0.1357	0.3310	0.3154	0.6611	0.0297	0.3053	0.1702	0.2139	0.2093	0.1887	0.1702	0.1963	0.0018	0.1247	0.0481	0.0261	0.0463
W	0.1873	0.1122	0.1577	0.0017	0.0671	0.0192	0.7082	0.0000	0.1877	0.0250	0.1195	-0.2144	0.3242	0.7516	0.1189	0.5088	0.1850	0.2498	0.0000	0.0569	0.1047	0.0000	0.0758	0.0000	0.6596	0.0001
GA	0.0203	-0.0249	-0.0281	-0.0034	-0.0233	-0.0128	0.1462	-0.0002	-0.0176	-0.0092	-0.0412	0.0093	0.0062	0.1272	-0.0258	0.2183	-0.0311	-0.0154	-0.0003	0.0416	0.0113	-0.0052	-0.0455	-0.1004	0.1116	-0.4492
S	-0.0268	0.0165	0.0471	0.0057	-0.0259	0.0035	0.1753	0.0000	0.0345	0.0083	-0.0461	-0.0176	0.0551	0.1908	0.0235	0.1292	-0.0714	0.0634	0.0000	0.0443	0.0191	0.0000	0.0292	0.0000	0.1675	0.0002
TF	0.1846	0.2515	-0.0078	0.0163	0.2233	-0.0307	0.1009	0.0345	0.4565	0.0694	0.3451	0.7427	0.0721	-0.0596	-0.0792	0.2245	0.5123	0.5323	0.8179	0.8772	0.7758	-0.0113	-0.0778	-0.2981	-0.0523	-0.1818
T	0.1631	0.2269	0.1487	0.0274	0.2068	0.0012	0.0458	0.1515	0.3591	0.0891	0.2625	0.7529	0.1736	-0.1045	0.0283	0.0716	0.3320	0.3441	0.5140	0.5113	0.4720	-0.0026	0.0248	-0.1016	-0.0917	-0.0745
VC	-0.0169	-0.0356	-0.0124	-0.0030	-0.0420	-0.0034	0.1268	-0.0284	-0.0387	-0.0108	-0.0315	-0.0146	0.0017	0.1455	-0.0091	0.1192	-0.0234	0.0010	-0.0352	-0.0083	-0.0154	-0.0013	-0.0055	-0.0265	0.1277	-0.1063
VU	-0.0995	0.0870	0.0131	-0.0021	0.0461	0.0438	-0.6316	0.0000	0.0541	0.0084	0.0821	0.0044	-0.0608	-0.5279	0.0980	-0.7241	0.1270	0.0598	0.0000	-0.1769	-0.0505	0.0159	0.0889	0.4194	-0.4632	0.2582

	HGJ	MC	NW	N	WL5	WL6	WL7	WL8	OLC	PW	Pre	Pro	ST	Sent	SLP	SE	TCS	TW	UW	V	W	GA	S	TF	T	VC	VU
HGJ	1.0000																										
MC	0.0590	1.0000																									
NW	0.1236	0.0721	1.0000																								
N	-0.0011	0.0243	0.0000	1.0000																							
WL5	-0.0029	0.0202	0.0000	0.8619	1.0000																						
WL6	0.0000	0.0101	0.0000	0.7604	0.9368	1.0000																					
WL7	0.0004	0.0126	0.0000	0.6883	0.8291	0.8680	1.0000																				
WL8	0.0000	0.0109	0.0000	0.1019	0.0686	0.0000	0.4776	1.0000																			
OLC	0.0000	0.0355	0.0000	0.0282	0.0190	0.0000	-0.0027	0.0000	1.0000																		
PW	0.2350	0.2392	0.2468	0.0294	0.0137	0.0000	-0.0020	0.0000	0.0000	1.0000																	
Pre	-0.0658	0.0243	0.0000	0.2687	0.0944	0.0000	-0.0135	0.0000	0.0000	0.1053	1.0000																
Pro	0.0109	0.0860	0.0000	0.1983	0.1936	0.1161	0.1495	0.1249	0.4065	-0.0061	-0.0072	1.0000															
ST	0.1187	0.0591	0.6500	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1604	0.0000	0.0000	1.0000														
Sent	0.0351	0.0263	0.0000	0.5481	0.4216	0.2330	0.2954	0.2506	0.0693	0.0086	0.1641	0.2777	0.0000	1.0000													
SLP	0.2547	-0.0139	0.0000	0.0349	-0.0019	0.0000	0.0003	0.0000	0.0000	-0.0816	0.0023	0.0615	0.0000	0.1819	1.0000												
SE	0.0653	-0.0035	0.0000	0.0862	0.0065	0.0000	-0.0009	0.0000	0.0000	0.0000	0.0000	-0.0402	0.0000	0.0239	0.0000	1.0000											
TCS	0.0000	0.0081	0.0000	0.1100	0.0497	0.0000	-0.0071	0.0000	0.0000	0.0000	0.1643	0.0922	0.0000	0.2283	0.2872	0.0000	1.0000										
TW	0.2150	0.1808	0.8489	0.0167	0.0079	0.0002	-0.0009	0.0001	0.0001	0.7166	0.0587	-0.0032	0.5523	0.0051	-0.0451	0.0000	0.0002	1.0000									
UW	-0.2054	0.0313	0.0000	0.1994	0.0927	0.0000	-0.0132	0.0000	0.0000	0.1475	0.7135	-0.0411	0.0000	0.0583	-0.5532	0.0000	0.0000	0.0820	1.0000								
V	-0.0039	0.0345	0.0000	0.5739	0.5454	0.3772	0.4977	0.4056	0.1123	0.0118	0.1191	0.3625	0.0000	0.6849	0.0639	0.0387	0.2973	0.0070	0.0802	1.0000							
W	0.2445	0.0018	0.0000	0.1349	0.0618	0.0000	-0.0088	0.0000	0.0000	0.0000	0.3182	0.0207	0.0000	0.1792	0.5561	0.0000	0.0000	0.0002	0.0000	0.0396	1.0000						
GA	0.0743	-0.0255	-0.0013	-0.0123	-0.0064	-0.0002	0.0006	-0.0002	-0.0620	-0.0288	-0.0646	-0.0162	-0.0011	0.0159	0.1859	0.0001	-0.0001	-0.0169	-0.1660	-0.0141	0.1692	1.0000					
S	0.0194	0.0059	0.0000	-0.0222	0.0114	0.0000	-0.0016	0.0000	0.0000	0.0000	0.0808	0.0679	0.0000	0.0299	0.1412	-0.6546	0.0000	0.0000	0.0000	-0.0153	0.2539	0.0429	1.0000				
TF	0.0809	-0.0011	0.0000	0.1401	0.1757	0.0275	0.0169	0.0296	0.0082	-0.0733	-0.3730	0.1858	0.0000	0.3470	0.2422	-0.0156	0.0270	-0.0406	-0.4967	0.1808	-0.0793	0.0683	-0.0269	1.0000			
T	0.0160	0.0097	0.0029	0.1681	0.1677	0.0562	0.0519	0.0402	0.0754	-0.0191	-0.1048	0.1732	0.0028	0.3204	0.1170	-0.0782	0.1805	-0.0085	-0.1718	0.2327	-0.1390	0.0546	0.0239	0.5590	1.0000		
VC	0.0519	-0.0069	-0.0003	-0.0380	-0.0322	-0.0227	-0.0301	-0.0244	-0.0210	-0.0075	0.0239	-0.0130	-0.0003	-0.0642	0.1207	-0.0613	-0.0222	-0.0043	-0.0438	-0.0653	0.1936	0.2573	0.2540	-0.0363	0.0677	1.0000	
VU	-0.3217	0.0206	0.0000	0.0410	0.0197	0.0000	-0.0028	0.0000	0.0000	0.1031	0.2666	-0.0433	0.0000	-0.0899	-0.7918	0.0000	0.0000	0.0572	0.6987	0.0272	-0.7023	-0.2348	-0.1783	-0.2894	-0.0181	-0.1661	1.0000

Table A-5: Covariance Matrix for FGS Search over Competences and Adjacencies – Final Essays

	Adj	AB	Art	CCP	CI	ERC	ER1	ER2	ER3	ER4	ES	GE	HGJ	WL8	Pre	Pro	Sent	SE	UW	W	GA	S	TF	T	VC	VU
Adj	1119.9872																									
AB	-0.3016	12.6205																								
Art	183.8859	-2.9222	1915.6004																							
CCP	7.3109	1.9978	105.5495	96.3863																						
CI	0.4458	0.0130	2.3347	0.4346	0.0380																					
ERC	50.5480	-0.2670	83.8550	10.5047	0.1503	76.2721																				
ER1	17.7515	-0.0588	-1.5216	-1.6711	0.0092	-1.0580	1.3573																			
ER2	18.3648	-0.0238	-0.0889	-1.4175	0.0152	-0.7871	1.1876	1.1494																		
ER3	13.2910	-0.0232	-5.3350	-2.0984	0.0076	-1.4199	1.0867	1.0123	0.9734																	
ER4	15.8186	-0.0295	-6.6802	-2.3874	0.0034	-1.1974	1.1557	1.1274	1.0597	1.2288																
ES	81.0444	-0.1647	-20.3059	-9.9618	0.0387	-5.6598	5.9430	5.6041	5.1919	5.8004	28.3398															
GE	2.4432	1.7052	-30.8963	0.0000	0.0000	-28.2040	-2.6743	-1.4761	-2.0385	-3.6982	-13.5854	251.1954														
HGJ	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	6.7126													
WL8	756.9291	-10.1708	317.6567	-55.4483	0.4422	33.1705	35.5578	34.7631	24.9024	31.2603	157.7439	-55.4727	0.0000	2857.7143												
Pre	360.2006	2.7794	816.2781	134.0964	0.6047	47.9014	11.7756	11.8732	5.6211	4.2378	37.7454	0.0000	0.0000	423.4621	1321.7733											
Pro	-15.9687	20.3640	-154.7175	105.7758	0.6888	-14.1365	-3.1126	-1.2597	-1.2275	-1.5599	-8.7195	90.2837	0.0000	-538.5019	147.1594	1078.1911										
Sent	92.2853	1.1593	104.9679	13.6392	1.1910	20.6492	3.9352	4.5983	3.5163	5.4367	22.9233	0.0000	0.0000	145.4050	18.9754	61.3821	249.4019									
SE	-57.1537	-3.2320	11.9524	-16.7879	-0.1093	-4.1750	6.0310	-1.0693	0.1948	-2.8597	-0.5629	-14.3291	0.0000	85.4667	-23.3560	-171.1219	-9.7421	445.2249								
UW	1099.9067	8.0060	-163.3696	-16.3518	0.5375	-131.1645	79.7001	74.4066	68.0579	69.2862	360.7370	190.1112	0.0000	1481.4460	1051.1069	423.8866	137.3569	0.9681	6021.7402							
W	3113.8783	21.0150	4020.4413	614.7177	9.2297	387.3829	109.0391	116.0766	66.5150	65.3404	422.3114	480.4556	0.0000	4050.3527	5399.0310	1112.6590	1501.9156	-176.5924	8378.7968	45912.6675						
GA	1.3898	-0.0691	3.1898	0.2739	0.0021	1.5003	0.1784	0.1259	0.1285	0.2034	0.8396	-11.2809	0.0000	4.5215	2.6089	-3.6594	0.5428	0.5808	-4.3669	-0.7767	0.5348					
S	5.2113	0.1880	2.1245	1.3244	0.0122	0.4907	-0.2488	0.1373	0.0353	0.1985	0.3208	1.1000	0.0000	-1.8022	4.9689	9.9538	1.5565	-24.0275	5.7130	41.0721	-0.0317	1.4094				
TF	25.5372	1.0040	-3.4028	2.5319	0.2211	0.7623	1.7605	1.8432	1.7344	2.2018	9.7416	0.0000	0.0000	25.4896	3.5225	53.1552	46.2982	-8.4364	97.3370	278.8107	0.1008	0.6448	49.9283			
T	1.6264	0.0169	3.3452	0.5264	0.0269	0.3330	0.0546	0.0610	0.0353	0.0324	0.2156	0.2060	0.0000	2.0474	2.6802	0.8936	1.3642	-0.1418	4.0763	25.2645	0.0095	0.0250	0.2533	0.0941		
VC	0.1709	-0.0047	0.2690	0.0356	-0.0007	0.1064	0.0151	0.0111	0.0100	0.0126	0.0615	-0.7004	0.0000	0.3825	0.4358	-0.2470	-0.1777	0.0392	0.1750	0.6466	0.0429	-0.0017	-0.0330	-0.0001	0.0250	
VU	24.2579	0.2163	-154.0846	-21.8279	-0.2811	-28.6826	5.4720	4.6478	5.5121	5.6930	27.0179	6.1271	0.0000	37.3895	-53.5860	11.4510	-33.6253	5.8322	409.1624	-523.7779	-0.4863	-0.6735	1.8104	-0.5218	0.0001	66.2220

Table A-6: Correlation Matrix for FGS Search over Competences and Adjacencies – Final Essays

	Adj	AB	Art	CCP	CI	ERC	ER1	ER2	ER3	ER4	ES	GE	HGJ	WL8	Pre	Pro	Sent	SE	UW	W	GA	S	TF	T	VC	VU
Adj	1.0000																									
AB	-0.0025	1.0000																								
Art	0.1255	-0.0188	1.0000																							
CCP	0.0223	0.0573	0.2456	1.0000																						
CI	0.0684	0.0188	0.2738	0.2272	1.0000																					
ERC	0.1729	-0.0086	0.2194	0.1225	0.0883	1.0000																				
ER1	0.4553	-0.0142	-0.0298	-0.1461	0.0404	-0.1040	1.0000																			
ER2	0.5119	-0.0062	-0.0019	-0.1347	0.0726	-0.0841	0.9508	1.0000																		
ER3	0.4025	-0.0066	-0.1235	-0.2166	0.0394	-0.1648	0.9454	0.9571	1.0000																	
ER4	0.4264	-0.0075	-0.1377	-0.2194	0.0158	-0.1237	0.8949	0.9487	0.9690	1.0000																
ES	0.4549	-0.0087	-0.0872	-0.1906	0.0374	-0.1217	0.9582	0.9819	0.9885	0.9829	1.0000															
GE	0.0046	0.0303	-0.0445	0.0000	0.0000	-0.2038	-0.1448	-0.0869	-0.1304	-0.2105	-0.1610	1.0000														
HGJ	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000													
WL8	0.4231	-0.0536	0.1358	-0.1057	0.0425	0.0710	0.5709	0.6066	0.4721	0.5275	0.5543	-0.0655	0.0000	1.0000												
Pre	0.2960	0.0215	0.5130	0.3757	0.0854	0.1509	0.2780	0.3046	0.1567	0.1052	0.1950	0.0000	0.0000	0.2179	1.0000											
Pro	-0.0145	0.1746	-0.1077	0.3281	0.1077	-0.0493	-0.0814	-0.0358	-0.0379	-0.0429	-0.0499	0.1735	0.0000	-0.3068	0.1233	1.0000										
Sent	0.1746	0.0207	0.1519	0.0880	0.3871	0.1497	0.2139	0.2716	0.2257	0.3106	0.2727	0.0000	0.0000	0.1722	0.0330	0.1184	1.0000									
SE	-0.0809	-0.0431	0.0129	-0.0810	-0.0266	-0.0227	0.2453	-0.0473	0.0094	-0.1223	-0.0050	-0.0428	0.0000	0.0758	-0.0304	-0.2470	-0.0292	1.0000								
UW	0.4235	0.0290	-0.0481	-0.0215	0.0356	-0.1935	0.8816	0.8944	0.8889	0.8055	0.8732	0.1546	0.0000	0.3571	0.3726	0.1664	0.1121	0.0006	1.0000							
W	0.4342	0.0276	0.4287	0.2922	0.2211	0.2070	0.4368	0.5053	0.3146	0.2751	0.3702	0.1415	0.0000	0.3536	0.6931	0.1581	0.4438	-0.0391	0.5039	1.0000						
GA	0.0568	-0.0266	0.0997	0.0381	0.0149	0.2349	0.2094	0.1605	0.1780	0.2509	0.2156	-0.9733	0.0000	0.1157	0.0981	-0.1524	0.0470	0.0376	-0.0769	-0.0050	1.0000					
S	0.1312	0.0446	0.0409	0.1136	0.0528	0.0473	-0.1799	0.1079	0.0301	0.1508	0.0508	0.0585	0.0000	-0.0284	0.1151	0.2553	0.0830	-0.9592	0.0620	0.1615	-0.0365	1.0000				
TF	0.1080	0.0400	-0.0110	0.0365	0.1606	0.0124	0.2139	0.2433	0.2488	0.2811	0.2590	0.0000	0.0000	0.0675	0.0137	0.2291	0.4149	-0.0566	0.1775	0.1841	0.0195	0.0769	1.0000			
T	0.1584	0.0155	0.2492	0.1748	0.4504	0.1243	0.1527	0.1856	0.1166	0.0952	0.1321	0.0424	0.0000	0.1249	0.2403	0.0887	0.2816	-0.0219	0.1712	0.3844	0.0425	0.0686	0.1168	1.0000		
VC	0.0323	-0.0083	0.0389	0.0229	-0.0219	0.0770	0.0821	0.0657	0.0641	0.0720	0.0731	-0.2794	0.0000	0.0452	0.0758	-0.0476	-0.0711	0.0117	0.0143	0.0191	0.3708	-0.0088	-0.0295	-0.0027	1.0000	
VU	0.0891	0.0075	-0.4326	-0.2732	-0.1773	-0.4036	0.5772	0.5327	0.6865	0.6311	0.6237	0.0475	0.0000	0.0859	-0.1811	0.0429	-0.2616	0.0340	0.6479	-0.3004	-0.0817	-0.0697	0.0315	-0.2090	0.0001	1.0000

Table A-7: Covariance Matrix for FGS Search over Competences and Adjacencies – All Events

	Adj	AB	Art	CCP	CI	ERC	ER1	ER2	ER3	ER4	ES	GE	HGJ	WL8	Pre	Pro	Sent	SE	UW	W	GA	S	TF	T	VC	VU
Adj	1066.5906																									
AB	14.0550	4.9593																								
Art	1162.4876	16.3941	4883.9998																							
CCP	169.9040	3.6834	554.5235	106.7496																						
CI	3.2678	0.0632	-6.7752	-0.4907	0.1619																					
ERC	0.2712	0.3501	29.2778	15.3699	0.0143	68.5518																				
ER1	30.5939	0.4163	-37.5407	-2.5673	0.6097	0.0000	4.2946																			
ER2	30.5459	0.4304	-26.8062	-1.3488	0.5698	0.0000	3.9654	3.7369																		
ER3	18.8939	0.2818	-65.2202	-6.1190	0.6334	0.0000	4.2003	3.8762	4.3832																	
ER4	18.3398	0.2750	-64.4924	-6.0549	0.6172	0.0001	4.1288	3.8301	4.3327	4.3166																
ES	116.7132	1.6785	-258.5518	-22.1448	3.0472	0.0002	20.7180	19.2385	21.1251	20.9247	102.9310															
GE	93.1794	1.2222	-80.2028	-6.4179	1.6580	-14.2261	11.7434	11.0600	11.3961	11.1994	56.5984	121.0428														
HGJ	0.3526	0.4724	2.9942	0.1134	0.0278	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	2.8870													
WL8	946.3680	11.2658	478.9335	107.1005	0.9674	0.1509	23.8708	15.4837	10.6631	8.6004	67.2182	47.9901	0.1106	4443.7139												
Pre	1259.0808	17.5978	3568.5894	467.7449	-0.9906	29.8769	3.8380	10.7835	-20.5481	-20.4995	-46.9256	33.1352	0.6207	551.6004	3287.4062											
Pro	376.9642	10.9669	472.5952	105.5339	1.9225	0.1464	14.8275	15.0222	10.3453	10.2428	60.6806	46.5599	0.1073	538.6875	536.9031	522.6343										
Sent	432.3659	6.2811	1280.6957	164.0982	-0.0278	0.7929	-1.8335	0.9290	-10.4770	-10.4502	-32.2819	4.1196	0.5810	199.4201	1082.8645	193.4772	460.3576									
SE	69.6259	-0.0077	-1.6712	-0.3423	-0.2072	0.0000	3.7159	0.0000	0.0000	-1.4242	0.8675	0.0000	0.0000	485.3624	-0.3666	0.0000	0.0000	235.0436								
UW	2125.0184	29.5381	-1683.8453	-72.4237	38.5559	0.0000	267.6525	252.0771	259.7358	255.2546	1289.9747	766.6270	0.0000	1094.4892	897.7554	1061.8727	97.6434	0.0000	17472.7351							
W	11548.5461	154.6113	34817.3318	4399.6726	-27.4451	0.0000	-75.1887	0.0000	-313.1938	-310.7168	-1009.8161	65.7856	0.0000	5006.5520	29454.0510	4857.3536	10645.4168	0.0000	1499.3653	290532.1110						
GA	1.3445	0.0200	27.9006	3.2868	-0.1177	0.9365	-0.8218	-0.7281	-0.9530	-0.9385	-4.3799	-7.9259	0.0000	0.0828	16.9084	0.0803	6.6224	0.0000	-49.4976	183.8081	2.3095					
S	8.8919	0.1947	45.4885	5.7463	-0.0163	0.0498	-0.4447	-0.0427	-0.4488	-0.3288	-1.5938	-0.3605	-0.0015	-33.2459	37.9746	6.0659	13.5562	-19.1275	-0.9674	375.1383	0.3590	3.8786				
TF	229.8429	4.8473	-772.2605	-71.6887	8.1821	0.0000	50.2726	46.9592	53.1024	51.7612	253.8566	133.2104	0.0000	128.6498	-252.3266	124.8160	-94.4482	0.0000	3036.0923	-3794.3115	-11.2266	-5.5039	923.1672			
T	86.5067	1.7790	-137.5787	-7.7886	4.5854	0.9661	15.0252	14.1818	15.6104	15.1609	75.1393	38.0226	0.8961	10.7451	-1.2137	51.9841	12.9768	-11.2671	962.7242	-465.5429	-1.5989	1.1650	205.8911	256.1299		
VC	6.7473	0.0951	38.6435	4.7531	-0.0002	0.5232	-0.6015	-0.4491	-0.7765	-0.7451	-3.3173	-4.6177	-0.0159	-3.9944	28.2441	2.6415	8.5455	-3.2528	-29.0381	287.5280	1.4611	1.8089	-10.1110	5.4270	12.6602	
VU	-181.4886	-2.3900	-1336.1010	-156.1848	4.6118	-4.5663	28.2654	24.2546	35.4563	34.9423	157.8608	72.5134	-0.0776	-62.2071	-898.5934	-60.3533	-347.5229	0.0000	1631.1063	-9563.3651	-10.9666	-12.6280	418.9095	107.6787	-12.4042	483.4233

Table A-8: Correlation Matrix for FGS Search over Competences and Adjacencies – All Events

	Adj	AB	Art	CCP	CI	ERC	ER1	ER2	ER3	ER4	ES	GE	HGJ	WL8	Pre	Pro	Sent	SE	UW	W	GA	S	TF	T	VC	VU
Adj	1.0000																									
AB	0.1933	1.0000																								
Art	0.5093	0.1053	1.0000																							
CCP	0.5035	0.1601	0.7680	1.0000																						
CI	0.2487	0.0705	-0.2410	-0.1180	1.0000																					
ERC	0.0010	0.0190	0.0506	0.1797	0.0043	1.0000																				
ER1	0.4520	0.0902	-0.2592	-0.1199	0.7313	0.0000	1.0000																			
ER2	0.4838	0.1000	-0.1984	-0.0675	0.7326	0.0000	0.9898	1.0000																		
ER3	0.2763	0.0604	-0.4458	-0.2829	0.7520	0.0000	0.9681	0.9577	1.0000																	
ER4	0.2703	0.0594	-0.4442	-0.2821	0.7384	0.0000	0.9590	0.9536	0.9961	1.0000																
ES	0.3522	0.0743	-0.3647	-0.2113	0.7466	0.0000	0.9854	0.9809	0.9946	0.9927	1.0000															
GE	0.2593	0.0499	-0.1043	-0.0565	0.3746	-0.1562	0.5151	0.5200	0.4948	0.4900	0.5071	1.0000														
HGJ	0.0064	0.1249	0.0252	0.0065	0.0407	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000													
WL8	0.4347	0.0759	0.1028	0.1555	0.0361	0.0003	0.1728	0.1202	0.0764	0.0621	0.0994	0.0654	0.0010	1.0000												
Pre	0.6724	0.1378	0.8906	0.7896	-0.0429	0.0629	0.0323	0.0973	-0.1712	-0.1721	-0.0807	0.0525	0.0064	0.1443	1.0000											
Pro	0.5049	0.2154	0.2958	0.4468	0.2090	0.0008	0.3130	0.3399	0.2161	0.2157	0.2616	0.1851	0.0028	0.3535	0.4096	1.0000										
Sent	0.6170	0.1315	0.8541	0.7402	-0.0032	0.0045	-0.0412	0.0224	-0.2332	-0.2344	-0.1483	0.0175	0.0159	0.1394	0.8802	0.3944	1.0000									
SE	0.1391	-0.0002	-0.0016	-0.0022	-0.0336	0.0000	0.1170	0.0000	0.0000	-0.0447	0.0056	0.0000	0.0000	0.4749	-0.0004	0.0000	0.0000	1.0000								
UW	0.4922	0.1003	-0.1823	-0.0530	0.7250	0.0000	0.9771	0.9865	0.9385	0.9294	0.9619	0.5271	0.0000	0.1242	0.1185	0.3514	0.0344	0.0000	1.0000							
W	0.6560	0.1288	0.9243	0.7900	-0.1266	0.0000	-0.0673	0.0000	-0.2775	-0.2775	-0.1847	0.0111	0.0000	0.1393	0.9531	0.3942	0.9205	0.0000	0.0210	1.0000						
GA	0.0271	0.0059	0.2627	0.2093	-0.1926	0.0744	-0.2609	-0.2478	-0.2995	-0.2972	-0.2841	-0.4740	0.0000	0.0008	0.1941	0.0023	0.2031	0.0000	-0.2464	0.2244	1.0000					
S	0.1382	0.0444	0.3305	0.2824	-0.0205	0.0031	-0.1090	-0.0112	-0.1088	-0.0804	-0.0798	-0.0166	-0.0005	-0.2532	0.3363	0.1347	0.3208	-0.6335	-0.0037	0.3534	0.1200	1.0000				
TF	0.2316	0.0716	-0.3637	-0.2284	0.6694	0.0000	0.7984	0.7995	0.8348	0.8200	0.8235	0.3985	0.0000	0.0635	-0.1448	0.1797	-0.1449	0.0000	0.7560	-0.2317	-0.2431	-0.0920	1.0000			
T	0.1655	0.0499	-0.1230	-0.0471	0.7122	0.0073	0.4530	0.4584	0.4659	0.4560	0.4628	0.2159	0.0330	0.0101	-0.0013	0.1421	0.0378	-0.0459	0.4551	-0.0540	-0.0657	0.0370	0.4234	1.0000		
VC	0.0581	0.0120	0.1554	0.1293	-0.0001	0.0178	-0.0816	-0.0653	-0.1042	-0.1008	-0.0919	-0.1180	-0.0026	-0.0168	0.1384	0.0325	0.1119	-0.0596	-0.0617	0.1499	0.2702	0.2581	-0.0935	0.0953	1.0000	
VU	-0.2527	-0.0488	-0.8695	-0.6875	0.5214	-0.0251	0.6203	0.5707	0.7703	0.7649	0.7077	0.2998	-0.0021	-0.0424	-0.7128	-0.1201	-0.7367	0.0000	0.5612	-0.8070	-0.3282	-0.2916	0.6271	0.3060	-0.1586	1.0000