THE DIAGNOSTICITY OF ARGUMENT DIAGRAMS

by

Collin F. Lynch

Bachelor of Arts in Artificial Intelligence,Hampshire College, 2000

Master of Science in Intelligent Systems, University of Pittsburgh, 2011

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Collin F. Lynch

It was defended on

January 30th 2014

and approved by

Professor Diane Litman, Intelligent Systems Program, Department of Computer Science, and Learning Research & Development Center, University of Pittsburgh

Professor Christian Schunn, Intelligent Systems Program, Department of Psychology, and Learning Research & Development Center, University of Pittsburgh

Professor Vincent Aleven, Human-Computer Interaction Institute, School of Computer Science, Carnegie Mellon University

DISSERTATION DIRECTOR: Professor Kevin D. Ashley, Intelligent Systems Program, School of Law, and Learning Research & Development Center, University of Pittsburgh

THE DIAGNOSTICITY OF ARGUMENT DIAGRAMS

Collin F. Lynch, PhD

University of Pittsburgh, 2014

Can argument diagrams be used to diagnose and predict argument performance?

Argumentation is a complex domain with robust and often contradictory theories about the structure and scope of valid arguments. Argumentation is central to advanced problem solving in many domains and is a core feature of day-to-day discourse. Argumentation is quite literally, all around us, and yet is rarely taught explicitly. Novices often have difficulty parsing and constructing arguments particularly in written and verbal form. Such formats obscure key argumentative moves and often mask the strengths and weaknesses of the argument structure with complicated phrasing or simple sophistry. Argument diagrams have a long history in the philosophy of argument and have been seen increased application as instructional tools. Argument diagrams reify important argument structures, avoid the serial limitations of text, and are amenable to automatic processing.

This thesis addresses the question posed above. In it I show that diagrammatic models of argument can be used to predict students' essay grades and that automatically-induced models can be competitive with human grades. In the course of this analysis I survey analytical tools such as Augmented Graph Grammars that can be applied to formalize argument analysis, and detail a novel Augmented Graph Grammar formalism and implementation used in the study. I also introduce novel machine learning algorithms for regression and tolerance reduction. This work makes contributions to research on Education, Intelligent Tutoring Systems, Machine Learning, Educational Datamining, Graph Analysis, and online grading.

Keywords: Argumentation; Essay Writing; Argument Diagrams; Graph Analysis; Machine Learning; Ill-Defined Domains; Intelligent Tutoring Systems; Educational Datamining; Multiple Representations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

1.0 INTRODUCTION

Can argument diagrams be used to diagnose and predict argument performance?

1.1 OVERVIEW

In this thesis I will answer the question stated above. As I will describe below Argumentation is an essential component of advanced problem solving and essential to many educational domains including STEM fields. Argumentation is also difficult both for novice arguers and for existing intelligent tutoring technologies due to its complex and often implicit structure. Previous researchers have approached instruction in argumentation through the use of argument diagrams, graphical models that reify the semantic structure of arguments in a visual form. While such methods have their adherents their performance has been mixed. Moreover, it has not yet been shown that the argument diagrams themselves are diagnostic or can be used to evaluate novice arguers. In order to address this crucial question I focus on two important sub-questions: whether expert human grades of argument diagrams are both *reliable* and *valid*, and whether those diagrams can be graded automatically via *empirically-valid* rules and predictive *models*.

These questions were addressed in a grading and machine-learning study conducted in the context of a Psychological Research Methods course at the University of Pittsburgh. Students in the course were given a novel argument diagramming ontology implemented in the LASAD diagramming toolkit that was designed to reify the key features in written research reports. They were then tasked with reading and diagramming existing arguments and planning their own written arguments for class assignments with the toolkit. The students' diagrams and

subsequent essays were then graded using a parallel grading rubric defined for this study. This rubric contains both specific structural questions focused on key aspects of the argument and more general *gestalt* questions that address the argument persuasiveness, coherence, and quality. In parallel with this grading process I defined a set of novel diagram rules using Augmented Graph Grammars. These rules were designed to detect important components of the argumentation structure as well as violations of argumentative norms. These rules were then used as the basis for a set of regression models designed to predict essay grades from the diagram features.

As I describe in my analysis some of the expert diagram and essay grades are reliable. Of these reliable grades all but one are valid predictors of subsequent student performance. Therefore argument diagrams can be graded by expert graders. I also show that the *a-priori* diagram rules do correlate with the reliable essay grades and are thus empirically valid. I then show that the induced regression models can be used to predict the reliable grades and that these models are competitive with the expert human graders. Therefore in answer to the primary research question argument diagrams can be used to predict argument performance.

In addition to the above conclusions this work makes contributions to research on education by demonstrating the utility of argument diagrams for undergraduate writing courses. It also contributes to the literature on Intelligent Tutoring Systems (ITS) by showing that it is possible to develop *empirically-valid* diagnostic rules for argument diagrams and that those rules can form the basis for predictive models. It also further demonstrates the utility of *weak-theory-scaffolding* in ill-defined domains. This work makes other contributions to the literature on Educational Datamining (EDM) by highlighting the potential for diagnostic regression for argument diagrams, and to the domain of graph analysis and machine learning through the development of a robust Augmented Graph Grammar engine AGG. Finally this work makes a further technical contribution in the development of the SNG online grading toolkit which was instrumental to the data collection process.

1.2 INTRODUCTION

Argumentation is central to advanced problem solving, particularly in ill-defined domains where problem solvers must frame or recharacterize the problems to make them solvable and then defend those characterizations [73, 125, 140]. Attorneys articulate legal rules that, if accepted, will advance their clients' interests and then defend said rules by citing relevant precedents. Empirical researchers identify hypotheses, support them by citing relevant literature, and test them by experiment. Policy makers, architects, theoreticians, and even artists use arguments in their work to frame ill-defined problems, reify open-textured concepts, and anticipate potential critiques [126]. Even traditionally well-defined domains such as mathematics and physics involve a form of argumentation (i.e. proofs [61]). According to G.H. Hardy the complexity of the proof is integral to what makes a theorem beautiful [43].

Argumentation is an exercise in structured persuasion. At a general level advocates make claims (e.g. "The sky changes outside the cave"), and support these claims through the application of *argument schema* such as a citation of sources (e.g. "Hippocrates dictates that our fingernails should be even with the length of our fingers" [119]) or of counterfactual conditionals [98] (e.g. "Suppose that the Federation supported the Rebellion; then the Enterprise would destroy the Death Star."). The validity of the argument is often domain specific based upon accepted standards of proof and debate. Hypothetical cases and normative principles, for example, are standard in legal argument but may be unacceptable in some theoretical work. Skill in argumentation has been included in the Common Core Standards [114]. Argument instruction appears in a number of sections within the proposed standards, a notable example is shown in Table 1.1 (pp. 4).

When addressing complex or, worse yet, *wicked* problems, solvers cannot simply state a solution, they must *justify* it [22]. They must use argument both to convince others to accept what they have provided as a solution to the problem at hand, and for many domains, convince them to implement it, thus ignoring competing proposals. In Voss' study of policy making, the researchers found that expert problem solvers not only justified their solutions but constructed their justifications during the problem-solving process [126]. The importance of justification has also been highlighted in well-defined domains by other authors including

Aleven et al. [3], Conati & VanLehn [21], and Chi et al. [15] who found that asking students to justify their solutions improves performance.

Despite its importance and structured nature, argumentation is not always taught explicitly, even in domains such as law where its centrality is widely acknowledged. Students typically receive instruction in argumentation through *authentic* practice such as writing or reading essays and in-class debate, "...that ritual of fire charitably known as the Socratic method..." [2]. As a consequence, students risk getting "lost in the text," distracted by lower level grammatical issues and writing or speaking styles, thus failing to identify the underlying argument structure. Similarly, students' actual knowledge of argumentation is often masked by their level of development of oral or writing skills, or lack thereof, which can limit the effectiveness of expert or peer review.

Table 1.1: A segment of the Common Core Standards on Argumentation for English Language Arts & Literacy drawn from the common core draft (see [114]).

Argumentation is listed in the proposed Common Core Standards under reading and writing skills for grades 6-12. 12^{th} graders are expected to:

Write arguments to support claims in an analysis of substantive topics or texts, using valid reasoning and relevant and sufficient evidence.

- a Introduce precise, knowledgeable claim(s), establish the significance of the claim(s), distinguish the claim(s) from alternate or opposing claims, and create an organization that logically sequences claim(s), counterclaims, reasons, and evidence.
- b Develop claim(s) and counterclaims fairly and thoroughly, supplying the most relevant evidence for each while pointing out the strengths and limitations of both in a manner that anticipates the audiences knowledge level, concerns, values, and possible biases.
- c Use words, phrases, and clauses as well as varied syntax to link the major sections of the text, create cohesion, and clarify the relationships between claim(s) and reasons, between reasons and evidence, and between claim(s) and counterclaims.
- d Establish and maintain a formal style and objective tone while attending to the norms and conventions of the discipline in which they are writing.
- e Provide a concluding statement or section that follows from and supports the argument presented.

Arguments are not always made explicitly but are implicit in the style of presentation, the order of claims made, or the tone of the presentation. The logic of an argument is often encoded in domain-specific ways which students must be taught [63], or are left implicit [2]. As a consequence, arguments are often difficult to detect and argumentation, as a skill, difficult to acquire. Novice arguers often face difficulty in processing and making arguments, sometimes missing or omitting crucial argument moves. Law students, for example, often fail to recognize the key legal criteria being applied in a case while science students often fail to identify or articulate clear and testable hypotheses.

In recent years diagrammatic models of argument have been growing in prominence as theoretical models, practical tools, and educational interventions. These models are designed to make argument schema explicit, reifying the essential claims and structured relationships among them. These models represent persuasive arguments as a graph consisting of content nodes linked by relational arcs representing logical connections, support, opposition, or argumentative schema. In Toulmin diagrams, for example, specialized nodes represent a *claim*, the *data* on which the claim rests, as well as the *warrant* that validates this connection. Additional nodes include the *backing* for a warrant, and the *rebuttal*. A classic Toulmin diagram as described by [118] is shown in Figure 1.1 (pp. 6).

Other diagrammatic models of argument include the box and line structures of Vorobej [124], the weighted proofs of Carneades [37], and the Test-Hypo-Fact structure used in LARGO (described below). Reed, Walton & Macagno [96] present a historical overview of argument diagrams describing their ongoing use in legal argument, evidentiary arguments, philosophical logic, and Artificial Intelligence (AI).

Proponents of these models point to their potential to scaffold students' comprehension. Diagrammatic models, they claim, reify essential argumentative concepts and relationships or other opaque features of dialogue, making them explicit to novice arguers [4]. They can also help to highlight interconnections between cases, a goal cited by Spiro [112]. Suthers [117] cites two specific benefits of representational notations generally that are applicable to diagrammatic models of argument: *constraints* which limit the objects that can be expressed; and *salience* which makes specific objects or relations explicit. Harrell and Wetzel [46] in

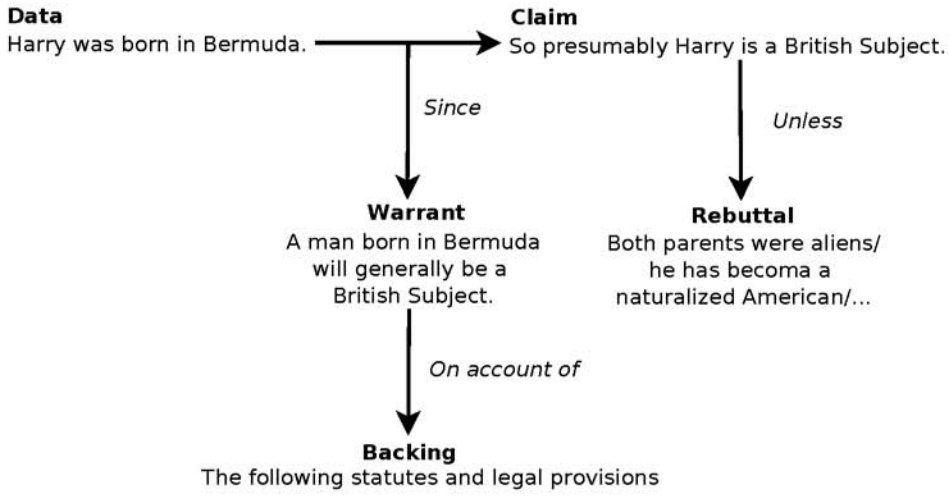


Figure 1.1: A Sample Toulmin diagram presenting an argument over the citizenship of a British subject born in Bermuda. The diagram includes an initial *datapoint* "Harry was born in Bermuda." This is in turn linked to a *claim* "So presumably Harry is a British Subject." This relatinship is supported by an existing *warrant* which provides a grounding for the relationship between the data and the claim and a *rebuttal* which includes counterarguments.

turn cite Schema theory (see [101]) in arguing that argument diagrams serve to reduce the cognitive load associated with processing individual argument schema and thus make them easier to acquire. Thus it is argued that novice diagrammers will focus on the key argumentation concepts and this will, in turn, improve their ability to both recognize argument structures and employ them (e.g. [69]).

Equally important, argument diagrams are amenable to computer processing. Textual documents have a number of basic features that can be used for automatic evaluation such as wordcount, sentence length, and more complex measures such as the automatic coherence measures used in Coh-Metrix [38]. While these structural features are often useful they

are not directly connected to the semantics of the argument being represented. Argument diagrams, by framing the structure in terms of argument components better link the simple syntactic features of the structure to the semantic features of the argument itself.

Scheuer et al. [104] survey a wider range of argument diagramming systems. They note the advantages cited by researchers including the ability of diagrams to make arguments more readable and to force arguers to make their contributions more explicit and comprehensible. On the other hand, they also note that diagrams, as stated by Buckingham Shum et al., risk increasing the students' cognitive load while feeling 'unnatural' or unintuitive [110].

It is not yet clear how well these promises and perils have been realized. Prior work on the use of argument diagrams has shown that they provide benefits for comprehension. Less work has been done on the effect of using argument diagrams, either to annotate existing arguments or to generate novel ones, has had on the students' ability to produce arguments. Similarly, little work has been done on the extent to which diagrammatic models of argument can be used to assess students' comprehension of argumentation and their ability to make novel arguments.

1.3 RESEARCH QUESTIONS

Do student-produced diagrams reliably encode pedagogically-relevant information that is of interest to domain experts? Can the diagrams be used to provide empirically valid assessment and guidance? And can we take advantage of the diagram structure to provide automatic assessment even on realistic open-ended tasks. In order for argument diagrams to be widely accepted in educational practice, these questions must be addressed. Novel educational interventions are justified by the extent to which they may be used to *diagnose* current and *improve* future performance of pedagogically relevant skills. And educators have no reason to incorporate new techniques unless they improve instructional outcomes or facilitate more opportunities for intervention and guidance.

While argument diagrams have been used in ITSs such as LARGO [91] or Belvedere [117], they have been promoted chiefly as pragmatic or *effective* interventions that improve

student performance not *diagnostic* ones. Much like a cricket player cross-training with a soccer game the practice is helpful but doesn't necessarily show off your bowling. This question of diagnosticity is important, however, for both theoretical and practical reasons. If one of the primary benefits of argument diagramming is the reification of argument structures then the diagram should reflect natural practice. If, however they are not diagnostic, then explicit scaffolding is not a useful explanation. Similarly, if the diagrams are not diagnostic then it will be difficult to convince often skeptical domain experts to use them in place of traditional representations. And finally, ITS designers have traditionally been focused on developing *transferable* interventions where the advice given by a system, such as showing the correct physics principle, transfers directly to offline practice. If the structure of argument diagrams cannot be connected to traditional essays then the motivation for hinting explicit process models is not as clear.

This we need to be able to convince educators, through scientific argument, that the quality of student-produced diagrams is indicative of students' understanding of real-world argumentation. We must also convince them that this quality is consistent with their own pedagogical goals. We must also convince educators that the act of drawing diagrams, and feedback that the students receive, will improve their subsequent ability to make well-reasoned essay arguments or engage in oral debate. In short we must convince skeptical domain experts that both the assessment of arguments and argument skills transfer out of the diagram context and into the "real world."

With that in mind this thesis addresses two general questions:

Q_h Can student-produced argument diagrams be assessed reliably by *human* graders and are those assessments valid predictors of future performance?

Q_a Can argument diagrams be analyzed *automatically* to diagnose students' argumentation skills and to predict future performance on "real-world" tasks?

As will be discussed in Chapter 2 (pp. 11) previous researchers have tested the educational utility of argument diagrams with mixed success. Most of this work has focused on the use of argument diagrams for reading or domain comprehension. This work has been driven by the a-priori assumption that diagrams encode pedagogically relevant information

or that (as with note-taking [122]) the act of constructing diagrams aids in problem solving. To date, little work has been done on the use of argument diagrams to *diagnose* student performance.

The present research makes two crucial extensions of prior work. First, rather than testing students' general understanding of argumentation, it focuses on their specific ability to compose their own arguments, in this case written research reports in which students describe hypotheses and empirical studies designed to test them (see Chapter 3 (pp. 27)). Second, rather than focusing solely on general graph features, such as node or edge counting, my research focuses on complex a-priori rules and induce predictive linear models.

This work makes contributions to the study of argument diagrams by expanding our understanding of their potential applications. It contributes to education by demonstrating the utility of argument diagramming as a graded instructional intervention. It also contributes to the literature on educational datamining by exploring and testing analytical methods for ill-defined domains. Finally it contributes to the literature on Intelligent Tutoring Systems (ITS) and AI in Education by identifying mechanisms for the automatic evaluation and diagnosis of student argumentation skills, methods which are suitable for incorporation into existing tutoring technologies.

1.4 OUTLINE

The discussion of this work will be divided into four parts. Chapters 2 and 3 survey relevant prior work and describe the study context.

In Chapter 4 I address question Q_h and describe studies of reliability and validity conducted with human graders. It will be demonstrated that both criteria are met. Student-produced argument diagrams can be graded *reliably* and those grades are *valid* predictors of students' argumentation abilities. While the research does not demonstrate a perfect correlation between students' diagram and essay grades, the observed relationships are strong enough for practical use.

Chapters 5 (pp. 61) - 7 (pp. 97) focus on question Q_a. They describe methods for

the automatic analysis of argument diagrams using augmented graph grammars and the automatic induction of complex predictive models from a-priori rules. They also discuss the individual empirical validity of the rules and the predictive quality of associated models. As the reported analyses show, some but not all of the rules are significant predictors of student performance while more complex models can provide strong overall predictions of student quality. These results are positive and are consistent given the low-level nature of the individual rules and the complex open-ended structure of the diagrams.

Finally Chapter 8 (pp. 131) presents overall conclusions and future applications of this work as well as an extended discussion of challenges for agreement.

2.0 BACKGROUND

This chapter will present an overview of relevant prior work. It will begin by discussing general studies of argumentation and use of argumentation systems in education. It will then summarize prior work on the use of argument diagrams as educational interventions (Section 2.2) and survey the ways in which prior researchers have analyzed student-produced diagrams (Section 2.3). The discussion will focus in detail on two diagramming tools: LARGO (Section 2.4), and LASAD (Section 2.5). The chapter will then conclude with a brief introduction to graph grammars (Section 2.6) and diagnostic rule induction (Section 2.7).

2.1 ARGUMENTATION

Argumentation is an essential aspect of many domains. While it may rarely be taught explicitly it has received a great deal of recent interest in education generally and in AI and Education in particular. Due to the complex and open-ended structure of argumentation much of the non-graphical work has been focused on scaffolding the process (e.g. [106]) and the use of expert systems (e.g. [39]).

In recent years a great deal of research has also been invested in automatic essay analysis (e.g. [30]) and guidance (e.g. [99, 28]). This work has focused on evaluating the essay text using coherence metrics and other Natural Language Processing (NLP) techniques. This work has chiefly grown out of large-scale courses and automated educational assessment. While these techniques have shown success, they are focused primarily on the textual and syntactic aspects of written arguments such as verbal style and tone rather than the deeper domain-specific semantic relationships that are the subject of this thesis. Therefore they are

limited in their ability to provide the deep semantic and structural guidance that students often require.

2.2 EDUCATIONAL IMPACT

The use of argument diagrams as educational tools has been approached by a number of authors including Ashley [4], Pinkwart et al. [91], Suthers [117], Carr [14], Easterday [33], Chryssafidou [19, 20], and Harrell [46]. In turning to diagrams the authors noted their benefits in reifying opaque arguments and scaffolding student behavior. Despite the promise of diagrams as educational tools, the results of prior research have been somewhat mixed.

In a survey by van den Braak et al [12] the authors examined four sets of studies by Carr [14], Suthers [117], Schank & Leake [102] and van Gelder [123]. The study domains ranged from legal argumentation [14] to scientific research [117]. The studies examined a variety of outcome measures, as well, ranging from improvements in critical thinking [123] to increased argument complexity [14, 20]. Despite promising trends reported in some of the studies, the authors remain skeptical, having identified a number of methodological limitations such as absent control conditions, self-selected conditions, or small sample sizes that prevented them from drawing definitive conclusions that argument diagrams contribute to learning.

In [14], for example, Carr sought to test the impact of diagramming on written legal arguments and legal comprehension. Students participating in the study were directed wither to diagram planned arguments in QuestMap a graphical tool or to plan them as normal in text. The students were permitted to self-select which tool they would use. The author evaluated the students using an in-class assignment grade and written essays. The essays were annotated using a Toulmin model and the resulting diagrams were evaluated for complexity and compared to existing expert diagrams. Ultimately, while Carr found qualitative improvement by the students he found no significant difference between the conditions. The absence of true random assignment or clear guidelines for the text group makes it difficult to draw general conclusions from this work.

In [104] Scheuer et al. drew similar conclusions to van Den Braak et al. Here the authors sought to determine whether argumentation systems (including those that involve or support diagrammatic models of argument): help students to make arguments (*scaffolding effect*); help students to comprehend domain topics (*arguing to learn*); and help students to acquire general argumentation skills (*learning to argue*). While they found promising research supporting the first effect, the authors found no significant support for arguing to learn, and inconsistent results supporting the premise of learning to argue.

They did, however, detail interesting evidence supporting the contention that the format of an external representation can affect students' learning. They noted that studies which compared alternate visualizations such as diagrams and tables (e.g. [117] & [109]) supported the contention that the form of the representation and interaction affects student behavior and learning gains. Representations that provide more structure prompt students to use the structure which in turn leads to more elaborated arguments and argumentative discourse. This is related to a second finding that "micro scripts" which structure the students process, (e.g. instructing students in constructing and responding to arguments) encourages them to engage in better quality argumentation. Although, they focus on group, not individual work, these results are consistent with Hypothesis H_a.

For the present discussion the most relevant individual work is that of Pinkwart et al. [91, 92, 86, 4, 89], Easterday [33], and Chryssafidou [19, 20]. The work of Pinkwart et al. is focused on LARGO, an ITS for legal argumentation designed to teach students the process of making arguments with legal tests and hypothetical cases via a legal argument diagram [91]. Students use the system to diagram oral argument transcripts taken from the U.S. Supreme Court via a graphical diagramming language that reifies the *tests*, *hypothetical cases*, and *facts* in the argument as well as the relationships between them such as *distinguished-from* and *modified-to*. When using the system students are guided via a set of diagnostic patterns based upon a process model of legal argumentation described in [4, 89]. A LARGO screenshot is shown in Figure 2.1.

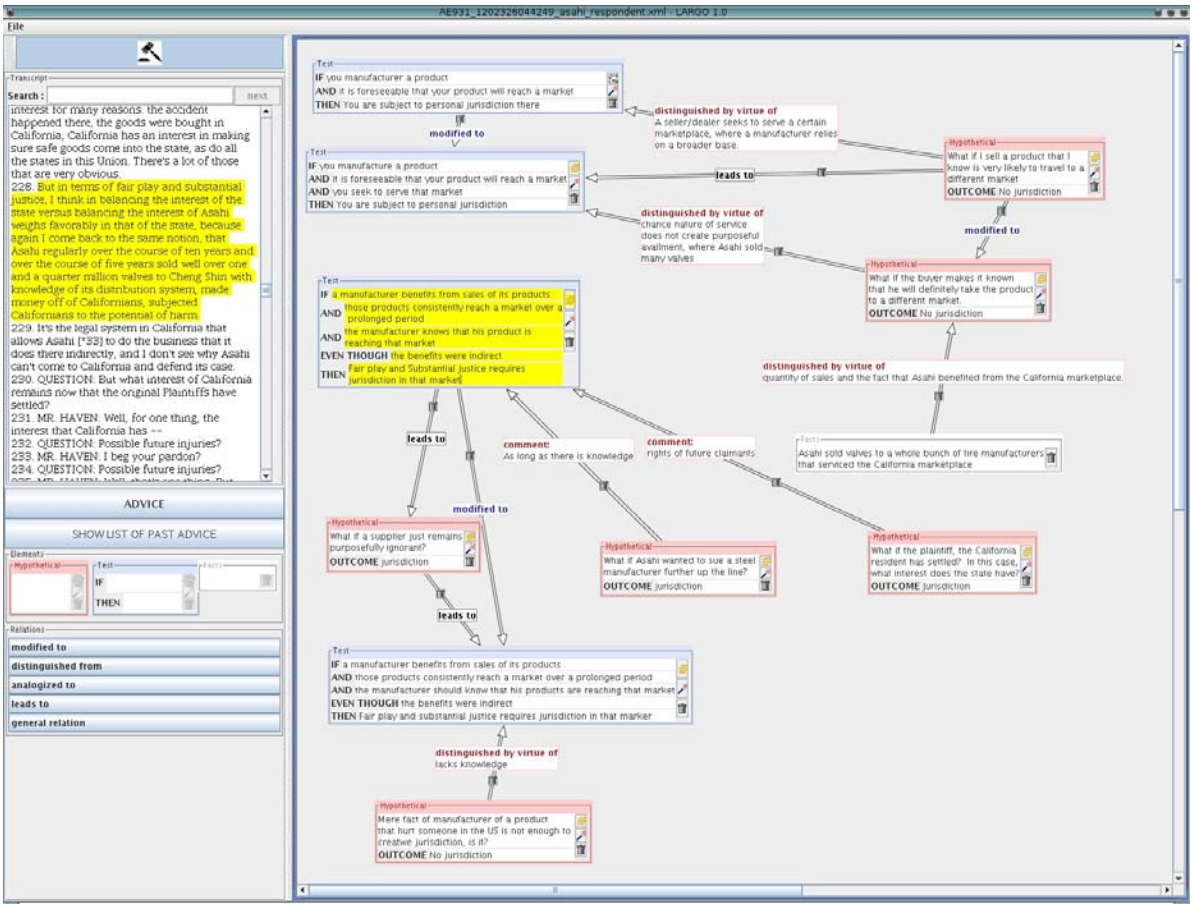


Figure 2.1: A Sample LARGO screenshot. The left-hand column contains the text transcript, advice buttons, and drawing palette. The right-hand side shows the diagramming panel. The highlighted text in the transcript is linked to the selected node.

As described in [91], LARGO has been deployed in a series of studies conducted at the University of Pittsburgh's School of Law. In these studies students, ranging from first to third-year students, made use of the system to diagram a set of oral arguments drawn from the domain of personal jurisdiction. In some, but not all of these studies, we drew a comparison between the task of annotating cases using the diagrammatic model and more standard text-based note-taking.

The primary goal of these studies was to test the effect of diagramming on students' ability to comprehend arguments (*scaffolding*) and to make novel arguments (*learning to argue*). The results of these studies were inconclusive. In some studies students using the diagramming system outperformed their text peers [86], while in others they did not [92]. As described in the next subsection, this data was used in our earlier efforts to examine the *diagnosticity* of argument diagrams. That is, can the argument diagrams be used to diagnose students' understanding of arguments and their ability to produce novel arguments as in the written essays?

Easterday et al. [33, 32] tested the impact of causal diagrams on argument comprehension. They performed a series of studies in which students were tasked with reading and responding to arguments in a policy debate. The students were split into three conditions *Text*, *Diagram* and *Tool*. The *Text* group acted as the control condition and was assigned to read the written articles in their original form. The *Diagram* group was supplied with both the original text and an expert-drafted argument diagram and was instructed to study both. And finally the *Tool* group was instructed to read the text and produce their own parallel diagram. The authors found that the diagram students outperformed the text group on basic comprehension tasks while the tool condition was not significantly different from either group. However, on a subsequent transfer task (reading a new article presented only as text) both the diagram and tool conditions outperformed the text group, suggesting that some essential lessons were transferred. Thus the work of Easterday et al. supported both the scaffolding effect defined by Scheuer et al. and the concept of *learning to argue*.

Unlike Pinkwart et al. and Easterday et al., Chryssafidou focused on the impact of diagrammatic argument models on argument making, specifically essay writing. In [19], she described Dialectic, a prototype tutoring system designed to teach paper writing through the use of argument diagrams. Like Belvedere and LARGO, Dialectic provides advice to students based upon an argumentation model. The model in question is a dialogue model that guides students to balance arguments and counterarguments when planning for paper writing.

In [20] Chryssafidou & Sharples present a comparison study in which students were either tasked with using Dialectic to plan their essay argument or with performing the same diagrammatic planning with pencil and paper. Thus the primary difference between the two groups was the absence of automatic advice. In lieu of random assignment the authors chose to assign weaker students to the computer condition in order to balance performance. The absence of truly equivalent conditions and clear results makes it difficult to assess the impact of the diagrams used in this work on students' learning.

However, in contrast to most of the work cited above, Chryssafidou & Sharples grade student essays not only at a gestalt level but structurally as well. Like Carr, they manually annotated the argument structure of each essay using an annotation model defined by Crammond [26, 27]. This rubric, like that of Toulmin, Rieke, & Janik [121, 120], renders a structured schema-based model of the argument which can then be analyzed for features such as the argument depth, complexity, and the presence of specific nodes.

The results presented by Chryssafidou and Sharples in [20] were somewhat inconclusive. Both groups improved on their essays as measured by the structural grading methodology. However each group appeared to improve differently. The pen and paper group expanded their use of counterarguments and refutations while the computer group expanded their use of supporting nodes. Additionally the authors note that the computer group changed their 'style' between the pre- and post-essays possibly masking the improvement in a strict node count. These results must be taken with a grain of salt, however, as she also reports a pre-test difference between the groups. In personal communications, Chryssafidou has stated that additional analyses are ongoing.

In more recent work Harrell and Wetzel conducted a series of studies on the use of argument diagrams to teach students argumentation and argument comprehension. In [45, 44], Harrell described a series of studies conducted in the First-Year Writing courses offered at Carnegie Mellon University. First-year writing is an introductory course in writing, and by extension critical thinking, offered to incoming freshmen. In it the students are instructed in argument processing and argument comprehension and are tasked with evaluating existing written arguments as well as producing their own.

In some years of the course the students were given instruction in argument diagramming where they used an existing diagramming toolkit or pencil and paper to read and diagram existing arguments for comprehension purposes. In other years the instruction was given "in the traditional manner." The performance of the students was assessed via pre- and post-tests that were given before and after the interpretation and argumentation section. In these studies Harrell found that all students improved in their critical thinking and argument comprehension skills and that students who were taught argument diagramming improved more than their non-diagramming peers. She also found that students who had low incoming competence gained more from the use of diagrams than higher-performing students. Harrell argued, citing Schema Theory, that the use of argument diagrams helped to reify argument schema for student comprehension.

In [46] Harrell and Wetzel discuss an extended meta-study of eighty-one students across seven sections of the interpretation and argumentation section. In these sections students were given pre- and post-tests that included two short writing assignments where they are tasked with reading a short argumentative piece and crafting a written response. The written problems were independently graded both for the quality of the argument itself, based upon the presence of crucial features, and for features of student comprehension, notably their coverage of crucial features of the target argument. In this study the authors found that students who were taught argument diagramming showed generally higher gains in both comprehension and argumentation skills over the course of the section. In particular they found strong gains in the quality of the authors' *metacommentary* which they cite as the language the authors use to make their arguments clear to users (see [50]).

These results are similar to and consistent with the benefits of argument diagramming reported by Pinkwart et al. in [93, 87, 91] and by Easterday in [33, 32]. In the LARGO work we found that diagramming aided subsequent argument comprehension, particularly for poorly-prepared students. Similarly Easterday found that diagrams, both student-constructed and instructor-provided, aided in later comprehension. Unlike the prior work, Harrell and Wetzel did not study comprehension of the diagrammed arguments only long-term comprehension ability. Moreover, unlike the work of Chryssafidou et al. and the present study the authors did not examine the use of diagrams to produce novel arguments only to annotate existing

ones, nor did they consider the structure or quality of the student-produced diagrams.

2.3 VISUAL REPRESENTATIONS OF ARGUMENTATION

As noted above prior work such as that of Suthers has compared the effect of alternate representations on student performance. In [117], Suthers describes a study of collaborative development of evidentiary reasoning in science. Students were assigned to use a diagrammatic representation from Belvedere (see [115]), a matrix layout encoded decisions in a two-dimensional form, and a word processor. This work, while informative, did not involve a direct comparison between specific diagrams only general conditions. Similarly, while the work of Chryssafidou involved a form of structured analysis including node counting and link measures similar to diagram analysis, that analysis was applied to expert-annotated student essays, not to student-produced essay diagrams.

In recent work McLaren, Scheuer, & Mikšàtko [79, 80, 81] developed tools designed to aid teachers in analyzing and guiding classroom discussion conducted via a graphical argument model and studied the automatic detection of crucial discussion components. This work, however, was not focused on drawing automatic connections between diagram structures and individual student learning, but on aiding an expert in searching, monitoring, and guiding group discussion. Indeed, of the work cited above, only Carr [14] provides any analysis of the change in student diagrams noting the increase in complexity over time for both groups.

Diagram analysis has also been the subject of prior work in psychology. Chi & Koeske [16] for example, used an expert-authored semantic network to represent a child's knowledge of dinosaurs as evidenced by their answers to interview questions. The authors performed a manual analysis of the diagrams to determine the complexity of the child's knowledge as well as to assess how well structured it was. This analysis included counting the number of links drawn from each dinosaur to other neighbors as well as the connections within and between groups of dinosaurs. While these are not argument diagrams they are relevant to this type of analysis as it shows a connection between graphical representations and knowledge.

Automated or guided diagram analysis has become increasingly important in machine learning (e.g. [23]). Joyner, et al. [53, 51], for example, present SubdueGL a grammar-induction algorithm designed to identify frequent subgraphs. An empirical comparison of graph classification algorithms is presented by Ketkar [55]. Further work in [52, 59] has extended these lines of research.

2.4 LARGO DIAGNOSTICITY ANALYSES

The LARGO evaluations reported above were focused on testing the value of LARGO as a learning intervention. Students used the system to annotate cases and the effect of their work was tested on a multiple-choice post-test covering general argumentation skills, argument comprehension, and response. In addition to these evaluations my colleagues and I have begun to investigate the diagnosticity of the oral argument diagrams drawn from the prior studies [75, 70, 67, 5, 72]. In [75] we tested the diagnosticity of individual diagram features such as the number of arcs drawn between nodes and the number of nodes produced. We also tested the predictiveness of the hand-tooled diagnostic patterns used in LARGO's help system. We found that some of these features correlated with students' incoming aptitude (as represented by their LSAT score), and with their position in law school, in this case 1^{st} year vs. 3^{rd} year, but not with their overall learning gains.

In [70] we continued this line of research focusing on the utility of the diagnostic patterns used in the LARGO help system both individually and as a basis for graph classification. For this study, rather than considering diagrams by student groups, we binned the diagrams by the students' post-test performance into *above-median* and *below-median* respectively. Our analysis showed that three of the diagnostic features were individually predictive of students' performance. Two represented failure to link nodes to the text and were significantly correlated with poor performance. The remaining feature, being prompted to revise the statement of a legal test based upon others' feedback, correlated with good performance.

We then applied both genetic programming [95] and C4.5 [8] to the construction of decision trees designed to classify student diagrams by post-test performance. The best

trees produced by both algorithms successfully classified roughly 80% of training cases and 100% of the test cases. Due to the small amount of available data, however, we were not able to perform a statistical comparison of the results. In this work we focused solely on classifying diagrams according to the students' overall performance or their incoming aptitude. We did not, for example, focus on correlating diagram features or diagnostic patterns with specific skills such as the students' ability to draw analogies and distinctions or pose novel legal tests and hypotheticals.

While this prior work has shown some successes at establishing a link between diagrammatic representations and student performance, it is not definitive. The investigation of the graph features and a-priori rules showed inconsistent predictiveness. Moreover, in neither study did we test students' ability to *make* novel complex arguments as is done here. Rather the graded performance data used in both cases covers solely students' general argument comprehension and their ability to comprehend novel arguments.

We have recently begun investigating the potential for manual diagram grading. In [71] we report on an expert grading agreement study. For this work we selected a sample of 198 diagrams drawn from our prior LARGO studies. These diagrams were individually graded by a pair of expert law school faculty according to a grading metric that included both gestalt quality grades as well as specific grades reflecting the students' comprehension of key aspects of the domain model such as tests and hypothetical cases. As we report in the paper we found substantial agreement between the faculty members leading us to conclude that reliable manual grading of student diagrams is possible. Due to the ordinal nature of the ranking we measured agreement using Spearman's ρ with the minimum correlation of 0.7. Graphical and tabular representations of the agreement are shown in Figure 2.2 and Table 2.1.

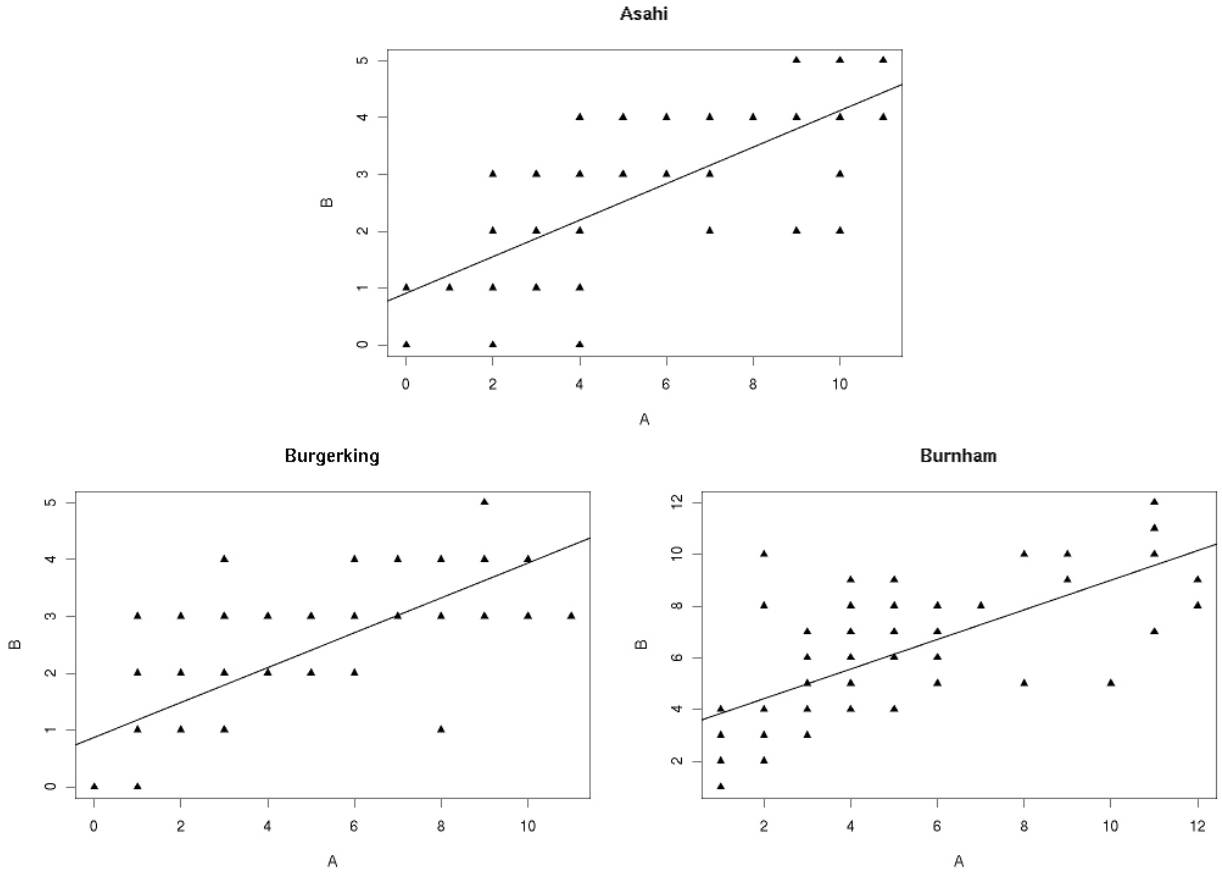


Figure 2.2: Per-case inter-grader agreement results drawn from [71]. Each plot shows the raw data for a single case with a fitted simple linear model (see [29]). The model values are specified in Table 2.1 (pp. 22).

Table 2.1: Per-case inter-grader ranking agreement drawn from [71]. The gestalt ranking is a nonparametric measure of score agreement computed using Spearman's ρ [29, 137]. The slope and intercept scores are for a simple linear model of the form $y_b = \alpha + \beta x_A + \epsilon$ (see [29, 133, 34]) computed between the scores. The models in question are plotted in Figure 2.2 (pp. 21).

	Gestalt Ranking ρ		*Overall Grade*			
			Slope		Intercept	
Case Name	ρ	*p-value*	*est.*	*p-value*	*est.*	*p-value*
Asahi	0.71	$p < 0.001$	0.32	$p < 0.001$	0.92	$p < 0.001$
Burger King	0.73	$p < 0.001$	0.30	$p < 0.001$	0.88	$p < 0.001$
Burnham	0.7	$p < 0.001$	0.57	$p < 0.001$	2.85	$p < 0.001$

2.5 LASAD

LASAD is a web-based successor to the LARGO system designed to support flexible peer-collaboration and argument diagramming [65, 13, 64]. Unlike LARGO, LASAD is designed to instantiate a variety of additional annotation and diagramming tasks, such as collaborative diagramming, and supports the introduction of flexible diagram ontologies. Introductory materials for LASAD are included in Appendix A (pp. 156).

An argument diagram ontology is a syntax describing the node and arc types usable in the diagram as well as basic features of each component. In LARGO, for example, the ontology specified the test, hypothetical, and fact nodes as well as the basic relationship arcs such as *modified-to* and *distinguished-from*. These items were specified in code within the LARGO system while in LASAD they can be specified using dynamic XML. A screenshot of LASAD in action is shown in Figure 2.3. LASAD was employed in this study for diagramming and data collection as will be described in Chapter 3.

In a recent series of studies Loll and Pinkwart tested the impact of LASAD's collaboration features and flexible ontologies on scientific argumentation [64]. Students in their studies worked either collaboratively or individually using the system to draft scientific arguments about the use of biofuels, among other topics. The authors hypothesized that manipulating the diagramming ontology from simple to complex would affect students' performance with more complex ontologies leading to more complex arguments. Ultimately they were unable to support or oppose this hypothesis but did conclude, consistent with findings by Suthers, that increased complexity of ontologies can lend itself to increased student errors (see [116]). Thus it is necessary to make representations parsimonious or risk reducing the quality of information they encode.

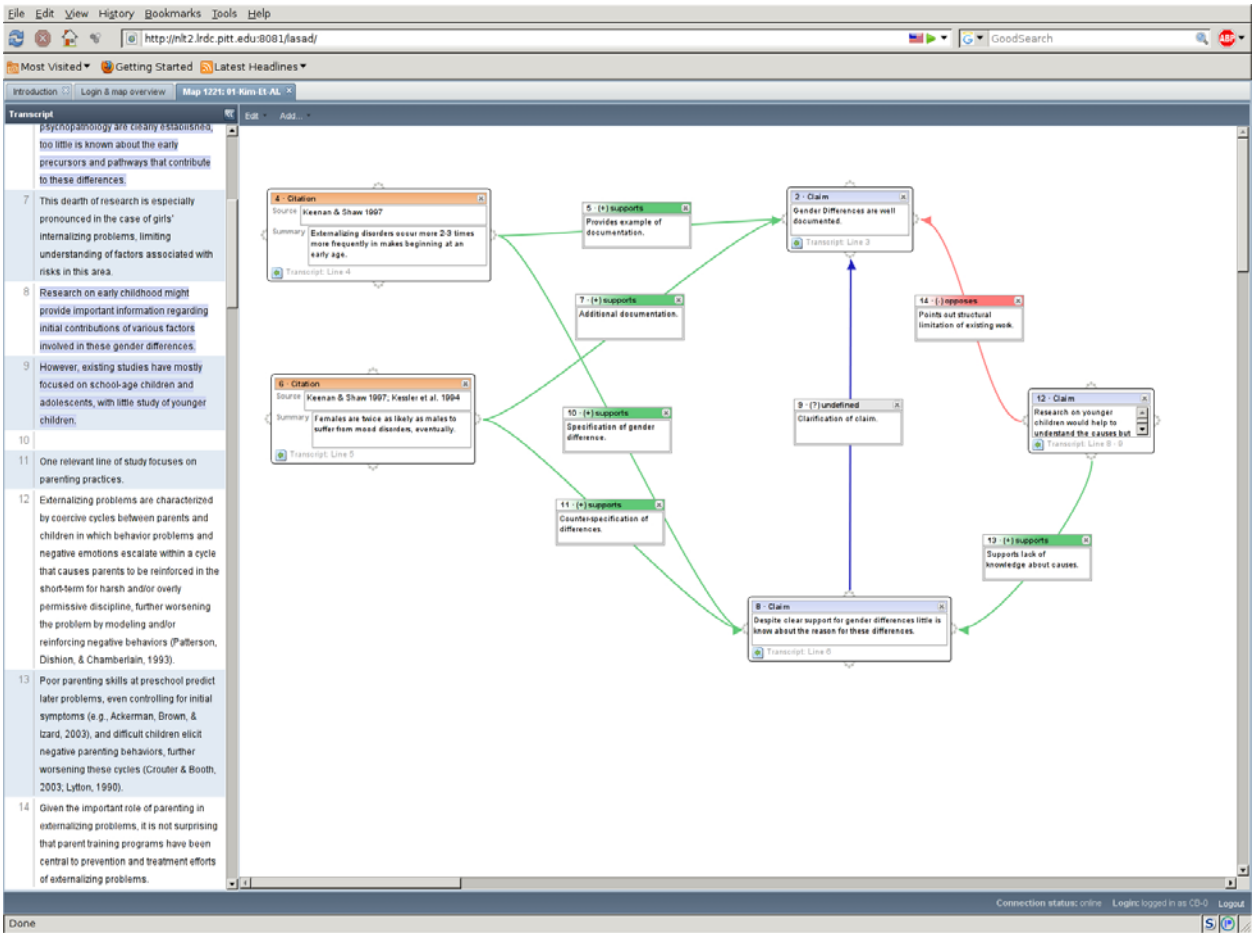


Figure 2.3: A LASAD Screenshot showing a basic argument diagram (right) and associated text transcript (left). The highlighted transcript text is linked to components of the diagram and selecting the text will flag the appropriate element.

2.6 GRAPH GRAMMARS

In LARGO, help is implemented by means of Graph Grammars [89, 4] which provide a set of a-priori rules that implement the LARGO argument model and test each graph for adherence to the model or deviations from it. A Graph Grammar, as described by Rekers and Schürr [97], is a formal grammar whose atomic components are graphs, and where the rules or *productions* transpose one graph to another.

In LARGO, augmented graph grammars were used to describe the argument model and to provide hints by means of grammar productions. These productions either matched legal modifications to the diagram, such as adding a fact node, or matched problematic graph states, such as the absence of a link between a hypothetical node and the transcript or matched pedagogically interesting patterns of argument warranted feedback and reflection. When the production rules are applied they would trigger a feedback message specified in part by the right-hand side of the production. The details of this application can be found in [89, 4]. These grammars were entirely defined by hand as guided by a domain expert. Thus all of the feedback resulting from the grammars was hand-tooled. Previous systems employing diagrammatic models of argument have used expert solutions [115] or hand-tooled graph grammars [85, 89] to provide guidance. LASAD is also equipped with an optional rule-based help system [105]. LASAD's system is constricted using the Jess rule engine for Java [94] but is formally equivalent to the augmented graph grammars defined for LARGO. Like LARGO it uses pre-compiled rules equipped with hand-authored hint messages.

In the current research these LASAD graphs will be analyzed by means of an augmented graph grammar engine developed by the author and described in detail in Appendix F (pp. 223). This engine uses a rule-based description that represents complex node and arc structures complete with type, directionality, and subfields as well as field contents. While the rules used by this system are compatible with the rules of LASAD's AFEngine the AGG graph engine is more suited to offline processing, supports non-LASAD diagrams, and can pave the way for subsequent research on rule induction. This implementation will be discussed in greater detail in Chapter 5 (pp. 61).

2.7 DIAGNOSTIC RULE INDUCTION

The induction of novel diagnostic rules and guidance is a major focus in educational datamining. Previous work in this area has included the development of detectors for gaming [7], and the induction of pedagogical tutorial guidance [17]. In recent work by Stamper et al. [113] and Nkambou et al. [83] the goal has been to extract automatic guidance for students in otherwise open domains. The goal of these projects was to induce automatic detectors for desired or prohibited student behaviors in otherwise open domains.

As described in later sections, the work described here is closer to the work of McLaren et al. [79, 81], related work by Harrer et al. for logfile analysis [47], Cluster Analysis work by Gross et al. [41, 42], and the work discussed in Lynch et al. [70, 68]. Here the goal is to support experts in identifying salient behaviors, test the validity of their assumptions using data, and to extract meaningful classifiers. Unlike McLaren et al. this research began with a set of hand-tooled rules and then focused on validating their utility and combining them to form a reliable model. However, in both cases there is interest in validating the intuitions by testing with real data.

2.8 CONCLUSIONS

As I have shown above, argument diagrams, and argumentation in general, are of interest to researchers in AI and Education. The results of this work, however, have been mixed. Researchers have shown that argument diagrams, whether instructor-provided or student-generated, can be useful scaffolds for argument comprehension and there exist promising results indicating that features of the diagram ontology do affect student performance. Argument diagrams, however, have shown limited success as planning tools and little work has been done on the use of student-produced diagrams for assessment. General techniques for diagram analysis and model production exist and have been applied to group discussions with some success. Therefore research questions Q_h and Q_a are both promising and open.

3.0 DATA COLLECTION

This chapter describes the type of scientific argumentation that is the focus of this thesis, namely empirical research reports (see Section 3.1). It also describes the diagramming ontology used (see Section 3.2), and shows sample texts (see Section 3.3). Finally, it gives an overview of the course context in which the data was gathered (Section 3.4).

3.1 SCIENTIFIC ARGUMENTATION

Science is about communication. We science and scientific discourse can be viewed as as an extended form of argument in which the interlocutors articulate general research questions and advance *defensible* answers to those questions in the form of testable hypotheses, empirical research, and sound theory. The hypotheses, in particular, must be: appropriate, relevant, and logically sound. The importance of argument in scientific research is a particular focus of Greene who notes, "The greatest advancements in science are realized through constructing convincing arguments, especially ones that resolve debatable research issues." [40]

Written research reports can thus be viewed as individual argumentative moves or self-contained arguments that both respond to and advance argumentative claims. In typical research reports, the core of this argument is presented in the introduction section where the authors state their general research questions and make broad claims, cite the relevant background literature, describe their process (at a high level), state their specific hypotheses, and draw connections between them. Their goal in doing so is to convince the reader that:

- The work is relevant to the world;
- The research question is novel, open or unanswered;
- The hypotheses are appropriate for the question;
- The hypotheses are testable; and
- The methodology is sound.

Much of this argument is made through appropriate use of citations. It is through citations that the authors highlight key support for or opposition to their hypotheses and contrast their work with prior research. This process of comparison and evaluation is discussed in detail by Beech who states that authors should take care to note and reconcile inconsistent findings in prior literature [9]. He further advises authors to note methodological problems with prior studies, and to articulate extensions or improvements that may be made. The importance of conflicts is also discussed by Greene who characterizes the identification and refutation of potential counterarguments as a central rhetorical goal [40].

All of the above requirements are meant to be addressed in the introduction. A reader should study the introduction to obtain an overview of the author's argument and to determine if the criteria have been met. The author's goal is to convince the reader to accept it as a sound and relevant piece of research, to make them willing to say: *"Yes I'll buy that."*

3.2 RESEARCH METHODS ONTOLOGY

As previously noted, argumentation is often implicit or difficult to detect and the rhetorical components of argument are often hidden from the reader or listener. One of the primary goals of scientific writing courses or methodological courses such as Research Methods (see below) is to train students to recognize and frame these components. In most courses this is carried out through a combination of example-based training with students reading and analyzing arguments and reification in lectures or other contexts. In [40], for example, the author characterizes arguments using a Toulmin-based approach where he focuses on *scientific claims*, the *warrants* backing those claims and the *data* that support them. This

is a structure similar to one of the diagramming ontologies used in [64]. In that study they also made use of the higher-level structures employed in Belvedere [116, 117].

Toulmin models are generally simpler and more domain-general than other argument ontologies. This simplicity has the advantage of reducing some types of user errors [64]. Simpler models, however, may be unsuited to educational contexts where students struggle with the basic argument structures and face difficulty translating from high-level hypotheses to more abstract components such as *warrants* and *backing*. Therefore, a higher-level, more task-specific, model of argumentation was chosen for this study. This model reifies the arguments using four specific *components*, represented as nodes, and connects them using four distinct types of argumentative *relations*, represented as arcs. I defined it in collaboration with Dr. Melissa Patchan and Dr. Chris Schunn at the University of Pittsburgh based upon an analysis of prior student work and reviewed by experienced instructors. It was also tested and refined during a pre-study conducted in the Fall of 2010, discussed below.

Examples of the four node types are shown in Figure 3.1 (pp. 32). They are:

Claims represent general research questions raised by or claims made by the author of the argument. Claims can be used to encapsulate any rhetorical point that is important for later discussion, e.g. "Argumentation is Central to problem solving, particularly in ill-defined domains.". The claim node itself has a single text field making it a simple free-text component.

Citations encapsulate literature references or other relevant materials. Citation nodes have two subfields one to represent the source or citation id, e.g. "Voss et al. 1983", and the other to include a short summary of the cited materials, e.g.: "In Voss' study of policy making they found that expert problem solvers not only justified their solutions but constructed their justifications during the problem-solving process" (see discussion in [126]). Ideally the short summative text should make clear why the citation is relevant and state what conclusions the author draws from it.

Hypotheses nodes represent a formal empirical hypothesis that can be tested empirically, preferably by the current study. The node itself represents the hypotheses as logical if-then rules with optional alternative outcomes. This logical structure is spread over two

required fields, the *conditional* or 'IF' field and the *consequent* or 'THEN' field, and one optional field the *alternative* or 'OTHERWISE' field. Thus this is the most complex of the nodes but it aligns with the structure used in the target courses and helps to scaffold the process of framing and defending it, e.g.:

> *If*: "Student diagrams adequately represent argumentation skills."
> *Then*: "The diagrams can be used to provide automatic guidance."
> *Otherwise*: "Argument diagrams may not be worthwhile as a tutoring technology."

Current Study nodes provide a mechanism to describe the structure of the current study and to highlight key features of the methodology. This is particularly important when those features are used to distinguish the study from other prior work, e.g. "In this study I will use the RM Ontology to frame student arguments as opposed to a Toulmin model.".

Examples of the four relation types can be found in Figure 3.2 (pp. 33). They are:

Supporting (+) indicates when a source node such as a citation provides support for the claims, study design, or hypothesis located in the target node. The arc contains a single text field that is be used to state the *reason* or explanation for the supporting relationship, e.g. citing a prior study that supports the current hypothesis. A supporting arc is shown connecting from Node3 to Node 2 in Figure 3.2 (pp. 33) (a).

Undefined (?) indicates that the relationship between the two nodes is indeterminate or merely factual and thus does not advance or diminish the claim being made. As before the arc has a single text field for the *reason*, e.g. "[Webster 1956] provides definition of term 'relevant' for [claim of dragnet relevance]." A sample undefined arc connects from Claim #10 to Claim #2 in Figure 3.2 (pp. 33) (a).

Opposing (-) indicates that a citation, claim, or hypothesis opposes the target item. As before a single reasons field is given, e.g. "[64] indicates that simpler ontologies are less error-prone than more complex ones.". A sample opposing arc is shown in Figure 3.2 (pp. 33) (b) connecting from Citation #6 to Hypothesis node #25.

Comparison (∼) is used to draw *analogies* and *distinctions* between nodes. This is frequently used to explain the differences between opposing citations or to highlight similarities between cited work and the current study. Comparison nodes are defined by a

flexible list of Analogy and Distinction fields, each of which should contain a specific point of comparison, e.g. "Previous studies on the use of argument diagrams in education *did not* attempt to diagnose students performance using the diagram structure.". A sample comparison arc is shown in Figure 3.2 (pp. 33) (b) connecting citation nodes #8 and #6.

The model does not include a specific node type for Research Questions. While I considered including it, I rejected it on the advice of the course instructors who advocated for a simpler ontology. In lieu of that explicit scaffold, students were instructed to represent research questions via a claim node with its contents framed as a question. This node in turn was to be connected to the hypothesis and other relevant claims making it central to the diagram structure. Similarly issues such as the novelty of the study as a whole are handled by drawing explicit comparisons between a current study node and cited materials.

This ontology was encoded in LASAD for use in the course both in annotation tasks and for original argument planning. It forms the basis of the later analytical rules. Sample written introductions and their associated diagrams can be found below.

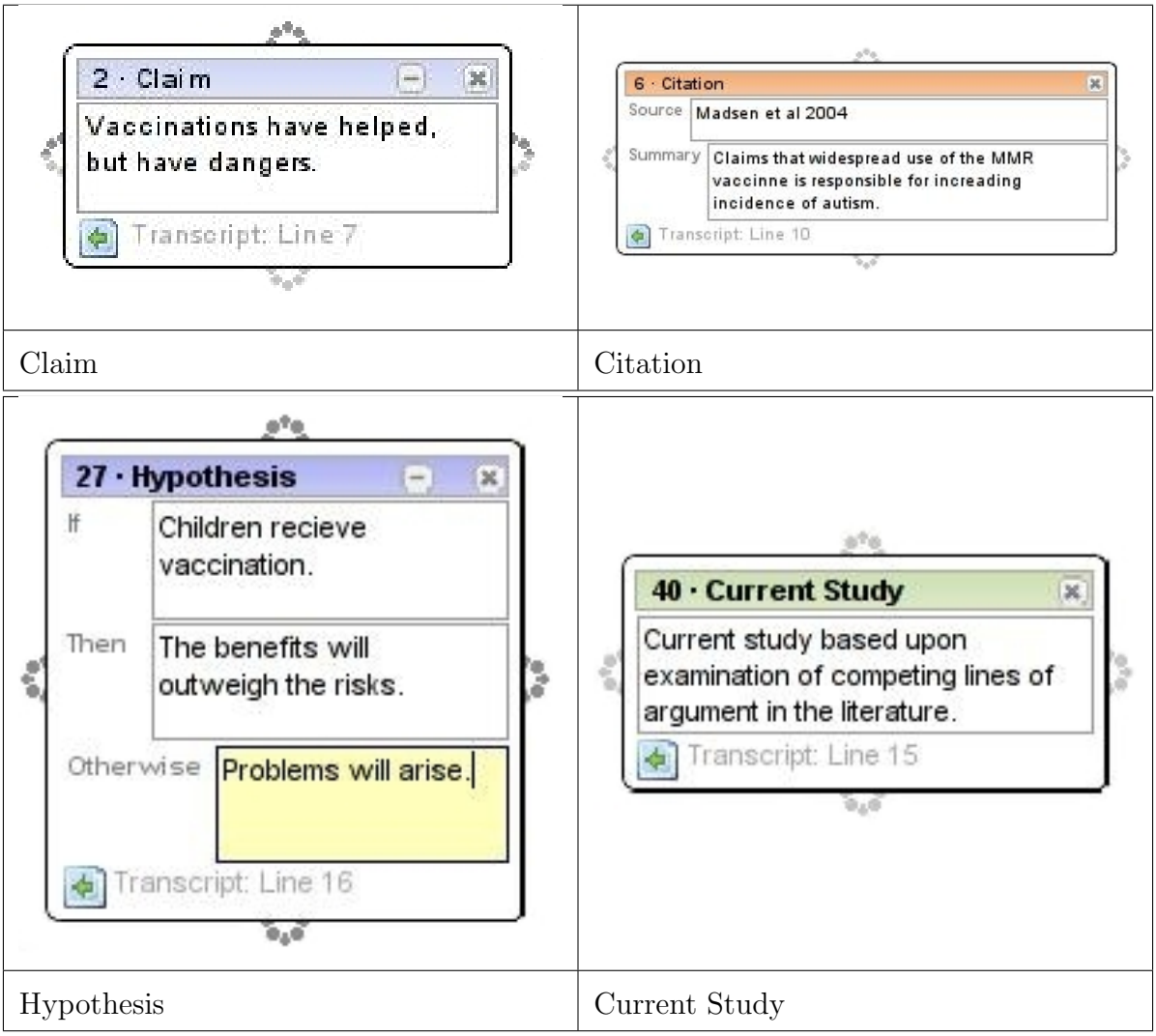


Figure 3.1: Component node types for the SciIntro ontology as they appear in the LASAD diagramming system.

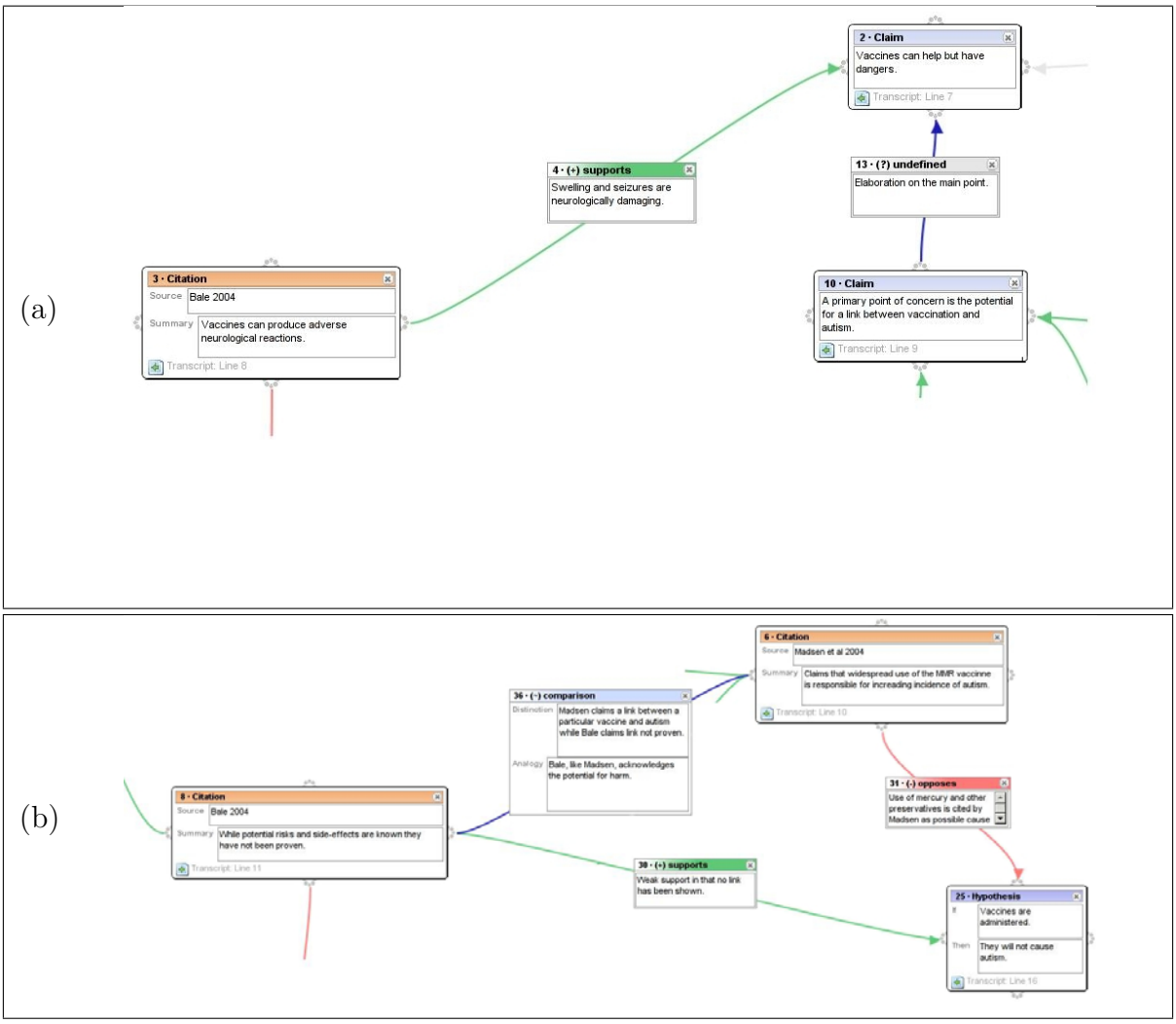


Figure 3.2: Relational arc types for the SciIntro Ontology as they appear in the LASAD diagramming system. *Supporting* and *Undefined* arcs are shown in (a) while *Comparison*, *Opposing*, and *Supporting* arcs are shown in (b).

3.3 SAMPLE INTRODUCTORY ESSAYS

This section contains two sample introductory essays along with their associated diagrams. Both were taken from the course dataset described below. Sample Essay A and its associated diagram were graded highly by both graders. Both are relatively detailed and have a clear narrative and logical flow. Sample Essay B and its diagram, however, were graded poorly. The logical flow of the argument is unclear and the citations have no connection to the primary hypothesis. Both the essays and the diagrams are typical of the dataset.

3.3.1 Sample Essay A

Effect of Symbol Use on Behavioral Compliance of an Instructional Sign

People, especially university students, see a variety of visual stimuli on a daily basis (from fliers to banners to endorsements) and learn to tune out excess information. The purpose of a sign is to inform, but if the language of the sign is not understood, pertinent information will be lost. This is especially important if the sign contains a warning of any kind, but is equally relevant in all situations; the assumption is that signs will only be present if they contain information that is necessary for the population to know. It will be beneficial to understand all aspects of attention capture and maintenance, the general effectiveness of symbols on signs, and the previous research involving the use of symbols.

To fully understand what would make a sign more efficient, it is necessary to understand the nature of attention, specifically as it relates to words and images. According to a study led by Wogalter and Leonard (1999), there are two stages of attention: the capture stage and the maintenance stage. The stimuli must stand out from its environment enough that it will be noticed by others - capturing attention. For optimal attention capture, a sign's contributing parts should complement each other in such a way as to be obvious and clear. Then, the maintenance stage will hold "attention while and until information from the warning is extracted."

But how do symbols aid the process of attention capture and maintenance? The advantage of symbols is that they transcend the language barrier, and can convey their intended meaning universally. Symbols are also useful in communication with anyone who is unable to read for a variety of reasons (Wolff and Wogalter, 1998). Research has shown that "warnings with pictorial symbols are rated more noticeable than warnings without them" (Friedmann, 1988). Symbols are more likely to capture attention because they are more familiar, and those with greater contrast have been shown to be observed and internalized faster than those that blend into the rest of the sign (Friedmann, 1988).

The symbols that stand out are especially useful in warning signs, where the information must be communicated clearly and efficiently so that the population remains safe.

Survey research has shown that "the likelihood of noticing a warning varies inversely with the amount of familiarity with the product"; therefore, a new sign or symbol on a familiar door will be more likely to be noticed than a sign that is always present (Friedmann, 1988).

In a study conducted by Kline and Beitel (1994), eleven sets of push/pull door signs (with a variety of modalities consisting of text-only, symbol-only and text-symbol combinations) were compared for their effectiveness. This "effectiveness" was determined through a ranking system that took into account the conspicuity of the sign, the reaction time of subjects to the sign, the subject's perception of the meaning of the sign, and the preference of the subject. The sign with the highest rating was one of mixed modality with a symbol of a hand pushing a door and the word "Push." (Kline and Beitel, 1994) This sign has the qualities of being easy to understand (simple text) and universally understood (simple symbol).

Another example that highlights the effectiveness of the inclusion of symbols on signs is a study by Wogalter (1997) entitled "Effectiveness of elevator service signs." In this study, the signs asked participants to refrain from using the elevator in the event that they would only have to walk up one flight of stairs in order to provide efficient elevator service to the rest of the population. Three signs that contained black and white words-only instructions were compared with a colored sign that contained words and symbols. The syntax of the text was varied on each of the signs, with some containing more or less instruction. The presentation, arrangement and font sizes were used to emphasize specific words on the sign. It was found that word order did not have a significant effect on compliance, as the three word-only signs were rated similarly. However, the presence of color and symbol (in addition to words) significantly increased the level of compliance from participants. (Wogalter 1997)

This specific study will analyze the effects of symbols on subject compliance to door signs. By comparing a sign that contained words only to a sign that contained the same words and a symbol, the direct effects of the symbol will be more observable. It is predicted that the presence of a symbol on the sign will increase the compliance rate of subjects to the instructions given.

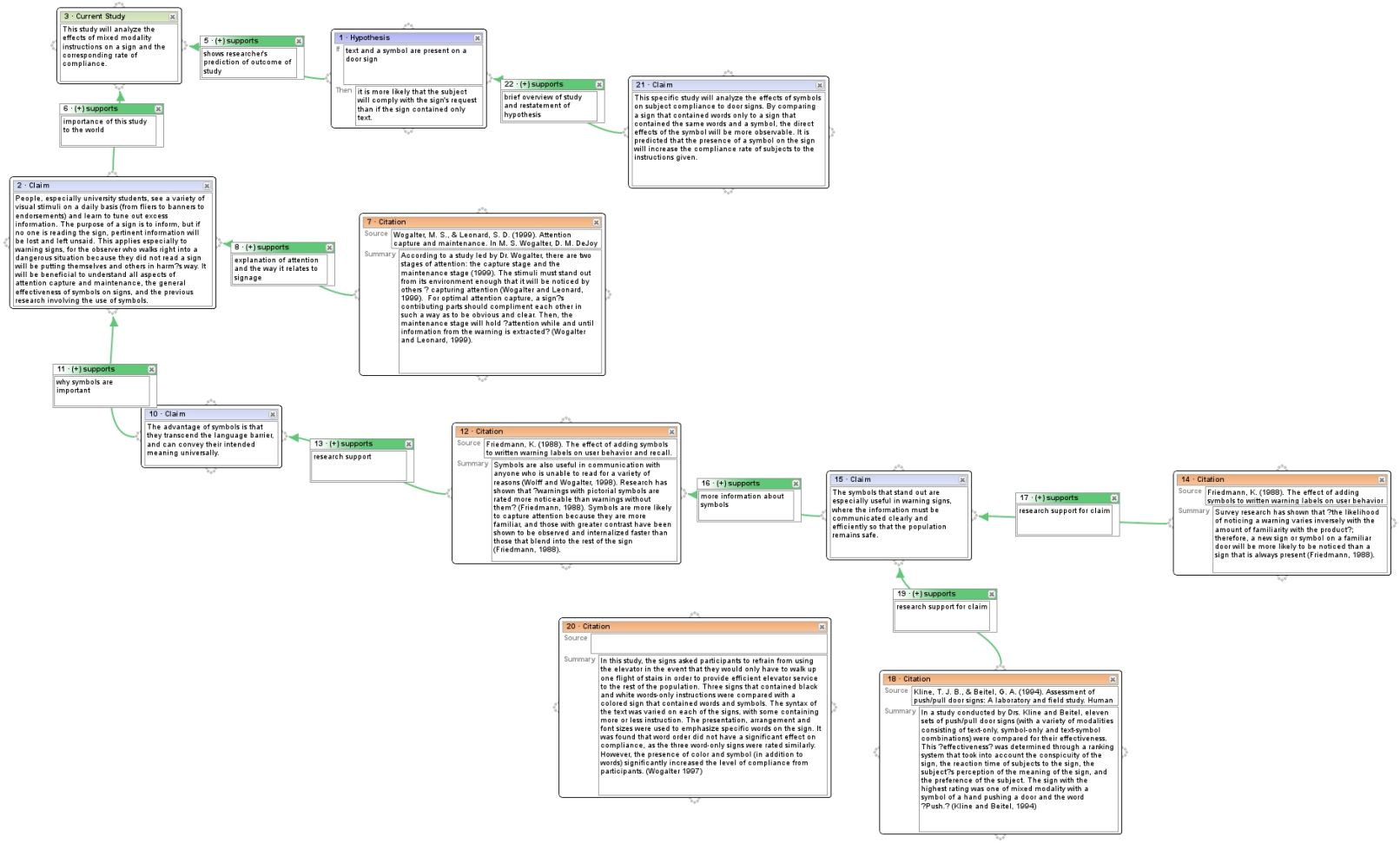


Figure 3.3: Planning diagram associated with the sample essay A.

3.3.2 Sample Essay B

The Effects of Visual Stimuli on Adults Attention Span Processes

Paradigms have been studied and replicated where the participant is more obedient or attentive when there wasn't anyone around than when others were present (Dambrun & Valentine, 2010). Depending on the participants sense of authority of the victim predicted their behavioral outcome of whether or not to obey. Other studies have shown that different types of hand gestures given to a receiver can highly affect the receiver's evaluation of the message's persuasiveness, effectiveness, and speaker's competence (Maricchiolo, Bonaiuto, & Ficca, 2009). This study infers that depending on the participant's evaluation of the sign will determine whether or not they pay attention to it.

Pinel, Long, and Crimin (2010) found, based on previous research by Asch (1951), which participants who previously shared a specific moment with another participant were less likely to conform than with participants who did not. They were also more likely to conform in public than in private. This study illustrates the level of attention of the participant regarding other participants. To be less likely to conform, the participant must have noticed whether or not others are present and how familiar their surroundings are.

Another study conducted by Tajfel (1978) noted that there is a propensity to socially differentiate than conform. This ultimately signifies that instead of people conforming to the directions of the sign, they will be more likely to ignore it, or differentiate from it. An alternative study examined the relationship between self-report and social skills (Tanimura & Watanabe, (2008). This study focused on the ability of the participant's attentiveness towards his or her social ability. This represents the degree to which the participant would have paid attention to the sign. The ability of the participant to accurately evaluate him or herself would be precursor knowledge of how the participant would react to the sign. If they are socially comfortable and not susceptible to embarrassment, then he or she may stop and turn around to look. On the other hand, if they are vulnerable to embarrassment then they may just ignore the sign and keep walking.

Many case studies have shown that certain attention processes are required for everyday tasks, such as walking through a door. The researcher's purpose of this study was to determine whether or not an individuals' attention span would be sharper during the day or at night. If they are more keen to their surroundings at night, then night time is a possible route for studying or handling a majority of complex tasks and vice-versa.

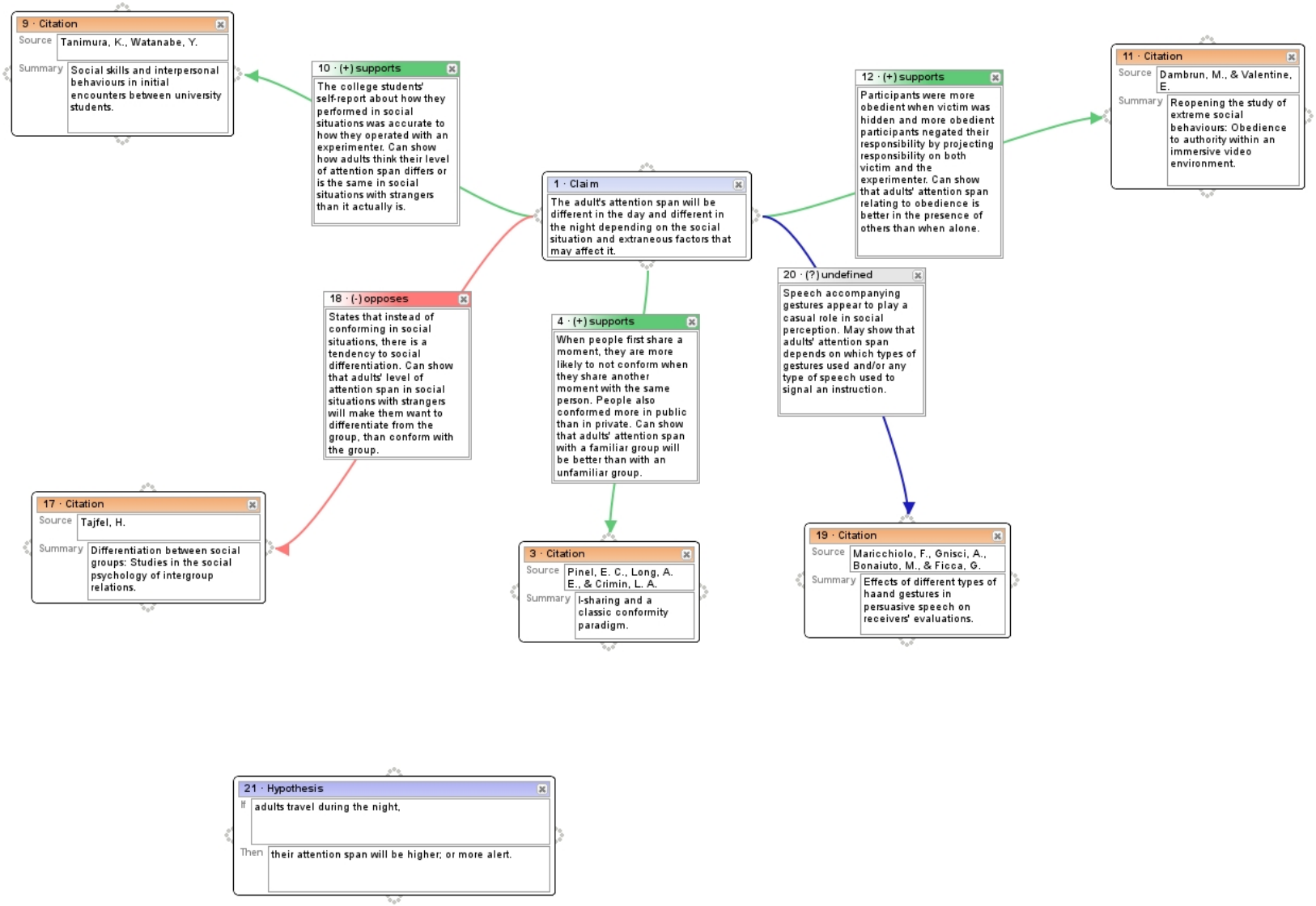


Figure 3.4: Planning diagram associated with the sample essay B.

3.4 STUDIES

For the purpose of this study I conducted two in-class experiments. The first was a pre-study experience conducted in one section of the Research Methods (RM) course at the University of Pittsburgh during the Fall of 2010 in collaboration with Dr. Melissa Patchan. Research Methods is a weeder course for all psychology majors and is required or encouraged for students majoring in Economics, Anthropology, and other social sciences. Students must complete the course in order to declare a major in Psychology. The course covers the research process from articulating hypotheses through study design, data collection, statistical evaluation, and writing. The course also includes segments on scientific ethics and reading scientific papers. The goal of the course is to train students in scientific comprehension and to make them good 'consumers' of science.

The Pre-study was a qualitative study designed to test the LASAD software and the argumentation ontology. The study took place in a single section of the Research Methods course offered at the University of Pittsburgh. Students were briefly introduced to LASAD and asked to diagram a single argument within it. Due to the small sample size the data was not used for any substantive analysis and only used for user experience testing and later grading as will be described in Subsection 4.3.1. The study led to functional improvements both in the software and the diagramming process, but no major substantive changes in the ontology.

The primary data collection phase was again in RM during the Spring of 2011. All of the discussion below and the experimental data came from this second round. The 2011 Spring course consisted of a single large-format lecture course with separate lab sections. The course enrollment at the time of data collection was 207. All of the relevant diagramming and authoring work was done in the lab section. There were a total of nine lab sections and the students were largely divided equally between them.

Within each lab section the students were required to complete two research projects including identification of a research topic, defining a hypothesis, collecting data, and writing up a research report. The first three phases were carried out collectively with the whole

lab working together to select the topic, frame the hypothesis and collect the data. The students were then required to carry out the writing either individually or as a team. This included selecting the appropriate background literature, framing of the research question, interpretation of the data and discussion of the results. A representative assignment sheet for this project can be found in Appendix B (pp. 181). Data collection for this thesis was integrated into these research and writing assignments.

For each assignment the study process consisted of three phases: introduction, diagramming, and writing. The students were *introduced* to LASAD and the diagramming ontology through a lecture and a guided reading task where they were asked to read one or more appropriate research papers and annotate the arguments made in the introduction section using LASAD. The introductory materials supplied to the students can be found in Appendix A (pp. 156). The papers being annotated were section-specific and were chosen by the Teaching Assistants or the course instructor.

During the *diagramming* phase the students were required to plan their own research paper by drafting a LASAD diagram. They carried this process out contemporaneously with data collection and background research as well as other course activities and were encouraged to update the diagram as they collected their source materials and structured their underlying claims. Once the diagram was complete they then *wrote* their research paper and submitted an initial draft for instructor review. The students were able to access their diagram during the writing process. If the students worked as a team then they were required to collaborate on both the diagram and essay components and to submit one of each per team.

Over the course of the study I collected one or more reading diagrams for each student in the course and paper planning diagrams and essays for each student and student author that completed them. All nine sections completed the paper planning diagrams and essays for the first assignment. However for the second writing assignment the sections varied with some TAs giving conflicting instructions to the students or making the assignment optional without informing the instructors. Due to intra-section variations I opted to focus my analysis on the first assignment. This one was handled consistently across the sections and all of the students were required to participate. From this process I collected a total

of 132 original diagrams and 125 essay drafts. After filtering out degenerate essays and diagrams and dealing with dropouts this resulted in a total of 119 gradable diagrams and 125 gradable essays. After purging duplicates, using the "Best Foot Forward" principle, and incomplete submissions I obtained 105 unique diagram-essay pairs, 74 of which were authored by a team with the rest being completed by individuals. All subsequent analyses described here will be based upon this set.

3.5 CONCLUSIONS

In this chapter I presented an overview of argumentation in empirical research reports. I further described the role that key components of the argument play in conveying the authors' specific goals. I then described an argument diagramming ontology tailored to the domain and included two sample arguments with associated diagrams. I then concluded by describing the data collection process and the Research Methods course at the University of Pittsburgh in which the study took place.

4.0 Q_H HUMAN-GRADING

Q_h Can student-produced argument diagrams be assessed reliably by *human* graders and are those assessments valid predictors of future performance?

4.1 INTRODUCTION

The focus of this chapter is on Q_h. In it I will discuss what it means for an educational activity, object, or test to be valid and reliable and will articulate specific, testable null hypotheses that address the substance of Q_h (see Section 4.2). This will be followed by a description of a grading study conducted to test these hypotheses (see Section 4.3) and presentation of the results. The results will be separated into discussions of reliability (see Section 4.4) and validity (see Section 4.5). The discussion will conclude with a summary of broad conclusions and applications (see Section 4.6).

4.2 RELIABILITY AND VALIDITY

In order for a graded activity to be pedagogically useful it should be both reliable and valid. A graded activity is *reliable* if multiple suitably-trained graders can assess it independently and obtain comparable results or if a single grader can re-grade it and obtain the same results [10]. The former definition is known as *inter-rater* reliability while the latter is *intra-rater* reliability. Inter-rater reliability is essential to the development of educational assessments. If an artifact cannot be graded reliably by multiple graders or a rubric cannot be applied

consistently then we have no guarantee that it measures persistent and transferable skills and no reason to believe that the assessment provided can predict future performance. High scoring students may simply have drawn a credulous jury.

As noted in Section 2.4 I have previously addressed inter-rater reliability of argument diagrams in [71]. The focus of that study, however, was conducted with LARGO diagrams. Unlike the present study students using LARGO annotate shared argument transcripts to prepare for a comprehension assignment. Therefore the reliability measure was based upon representational diagrams that covered the same underlying arguments and thus should share a structure. In the present assignment the arguments were open-ended and unique to each author while in the prior work the students were annotating a single shared argument thus the problem differs substantially.

I define a rubric or assignment as *valid* if it is predictive of or correlated with future performance on other relevant activities. If, for example, the students' graph grades are correlated with their future essay grades then the diagramming assignment is a valid predictor of their future writing performance. If the graph grades have no bearing on their essay grades however, then the exercise is invalid and there would be little point in spending time on it. Validity is a prerequisite for predicting transfer of skills to novel tasks and domains. It is particularly important when we consider the role of instructors in advising students on one activity with the goal of improving future work.

The problem of validity is coupled with the problems of translation and transfer. In the present context students are being asked to apply their skills at argumentation and their understanding of concepts such as "hypothesis" to construct argument diagrams. They are then expected to apply those same skills when drafting an essay, translating the argument from the diagram form to a written document and applying the same concepts in a novel form. Thus an exercise such as argument diagramming will only be valid if the skills it measures are transferable from the exercise domain to the target domain.

While previous researchers have sought to connect diagram usage to student performance none have conducted a systematic study of diagram assessment. Apart from [20] and [14] the authors have not reported on any rigorous annotation or grading of the essays and diagrams.

Nor have they drawn formal connections between detailed measures of diagram quality and subsequent essay performance. The only work conducted on that topic to date, [68], focused solely on automatic grading and a single gestalt assignment grade. While this supports us in our belief that the diagrams can inform subsequent grades it does not show that the diagrams are useful to human reviewers or that they can be used to provide more detailed predictions.

Consider the diagram-essay pair shown in Subsection 3.3.2 (pp. 37). The diagram shown in Figure 3.4 (pp. 38) has several clear issues. Hypothesis node #21 is disconnected from the rest of the graph and the remainder of the diagram consists only of a single star shape with arcs going from the central claim node to the citations. Nor has any attempt been made to compare the citations to one-another or to break the primary claim up into smaller claims. Therefore it seems likely that there are regularities that an expert instructor could recognize.

In order for the argument diagrams to meet the above criteria the following null hypotheses must be false:

H_{h1}: Student-produced argument diagrams *can not* be reliably graded by human graders.
H_{h2}: Student-produced argumentative essays *can not* be reliably graded using a parallel grading rubric by human graders.
H_{h3}: Human-assigned diagram grades *are not* valid predictors of parallel essay grades.
H_{h4}: Human-assigned diagram grades *are not* valid predictors of their gestalt essay grades.

These hypotheses were tested in a pair of studies focused on inter-rater reliability and validity. Both were conducted using a custom parallel grading rubric that is described in the next section. I will address the grading process and the hypotheses below.

4.3 GRADING

The diagramming ontology described in Section 3.2 (pp. 28) was designed to reify the key argumentation features present in research reports. My goal in designing this parallel structure was, in part, to ensure that the students could transfer the argumentation skills from one format to the other thus making the diagramming task a valid one. In order to test

the extent of this transfer and the validity of the task as a whole I defined a pair of parallel grading rubrics one intended for use on diagrams and the other for essays. Each rubric has 14 questions covering specific features and *gestalt* qualities of the elements. Each question or grade is paired with a comparable peer in the matching rubric.

4.3.1 Rubric

A short summary of the rubric questions is shown in Figure 4.1 (pp. 46). As that table illustrates questions *G/E.01 - G/E.11* focus on specific subsets of the argument such as the research question or the citations. Questions *G/E.12 (Arg-Coherent)*, *G/E.13 (Arg-Convincing)* and *G/E.14 (Arg-Quality)* are *gestalt* questions that cover the argument as a whole. All but *G/E.14* are graded on a scale from (-2 - 2) in increments of 0.5. Questions *G/E.02 (RQ-Link)* and *G/E.05 - G/E.11* allowed for an additional "N/A" value indicating that the question does not apply. *G/E.14 (Arg-Quality)* is graded on an 11-point scale (-5 - 5) again in 0.5 increments. A detailed summary of the questions along with keypoints for each of the grade values can be found in Appendix C (pp. 187).

The rubrics were initially developed based upon a qualitative analysis of the diagrams collected during the pre-study described in Section 3.4 (pp. 39). I then tested the draft rubric with a short pre-grading study using the same data. In this pre-grading process a pair of experienced TAs graded a sample of the diagrams independently of one-another and then met to discuss issues and differences. Both TAs had been involved in the course and were experienced with LASAD. Based upon the results of this process the rubrics were extended with additional questions and better-specified grading criteria to produce the final form used here.

4.3.2 Grading Process

The final grading itself was carried out by two independent graders. The *primary* grader was a Ph.D. student in Anthropology at the University of Pittsburgh who had previously served as a TA and grader for the RM course in 2012. During that year LASAD was again used and she led her students in comparable assignments to the ones used in the present

G/E.01 (RQ-Quality) Did the author clearly state a research question?
G/E.02 (RQ-Link) Was the research question, if stated, clearly relevant to and connected to the rest of the diagram/essay?
G/E.03 (RQ-Support) Did the author include relevant citation nodes or cite relevant background literature related to the research question?
G/E.04 (Hyp-Testable) Did the author include one or more testable hypothesis nodes in the diagram or articulate a testable hypothesis or hypotheses in the essay?
G/E.05 (Hyp-Link) Were the hypothesis nodes / hypotheses relevant and clearly connected to the research question?
G/E.06 (Cite-Conclusions) Did the author include clearly state the conclusions that he or she drew from the cited works in the citation nodes or essay?
G/E.07 (Cite-Reasons) Did the author include relevant summaries of the cited works and make clear hpow they connect to the remainder of the argument via text or arcs?
G/E.08 (Claim-Support) Did the author adequately support his or her claims via citations either textually (in the essay) or by drawing paths from the citation nodes to the claims (in the diagram)?
G/E.09 (RQ-Open) Did the author defend the opnenness of their research question by including similar research that disagrees about the hypothesis?
G/E.10 (Hyp-Open) Did the author defend the openness of their hypotheses by citing work that disagrees about it and clearly delinating the differences?
G/E.11 (Study-Novel) Did the author defend the novelty of their study by drawing explicit comparisons between their work and prior work and noting relevant distinctions?
G/E.12 (Arg-Coherent) Did the author develop a single coherent argument?
G/E.13 (Arg-Convincing) Did the author present a convincing argument?
G/E.14 (Arg-Quality) Please rate the overall quality of the diagram / essay based upon the organization, coherence, and completeness?

Figure 4.1: Summary of the Research Methods Grading Rubric. Detailed information can be found in Appendix C

study. As such she was experienced with the domain, the student body, the assignment, and LASAD. In addition to her existing experience with LASAD she was given a pre-training memo describing the goals of the grading process and instruction in the specific grading rubric. This memo can be found in Appendix D (pp. 210). She was also given instruction in the use of SNG a web-based grading toolkit that I developed for the purpose of this thesis (see Appendix E (pp. 215)).

The primary grader was given the task of grading the full set of diagrams and essays collected from the study including all planning diagrams and essays generated during assignments 1 and 2. This was a superset of the 105 unique pairs that I use for analysis below. The total number of diagrams and essays graded can be found in Table 4.1. All of the essays and diagrams were anonymized before being provided to her and the orders were randomly shuffled so that diagrams and essays could not be matched up. The essays were embedded within the web-based grading tool SNG and were graded entirely in that tool. The diagrams were posted for access in LASAD with grades being entered via SNG. In the pre-grading study the diagrams were viewed as static images however the graders found this format to be confining and expressed a preference for a dynamic format.

After conducting her initial training the primary grader was given samples of diagrams and essays from a prior year for discussion. She then proceeded to grade the diagrams and essays in batches. For the most part she did so independently, communicating only when she had technical questions. No ongoing guidance was required nor did I seek to influence her

Table 4.1: Individual grading assignment totals listed by grader.

Grader	Assignment	Diagrams	Essays
Primary	Validation	105	105
Reliability	Training 1	10	10
	Training 2	10	10
	Reliability	30	30

grade assignments. The full grading process for both the diagrams and essays required one full year once training was completed. As I will discuss below her graph and essay grades will form the basis of the validity analysis and subsequent model induction.

The *reliability* grader was an experienced Ph.D. Candidate and TA at Stanford University. She had completed an undergraduate degree in Psychology at Carnegie Mellon University where she completed a Research Methods course. She had also served as a TA for related courses at Stanford. As such she was experienced with the nature of the course and with the type of writing assignment. She was given pre-training on LASAD in addition to the grading memo and SNG manual. She was also given the opportunity to review some of the assignments from a prior year before grading. Both graders were hired separately and neither one had occasion to interact with the other.

The reliability grader was assigned the task of producing a set of independent grades to compare with the primary graders' results. The grader was given initial instruction in LASAD, SNG and the assignment. She then graded a set of 10 randomly selected diagram and essay pairs. I then compared these grades to those of the primary grader and discussed the differences with her. She then conducted a second round of training grading with 10 diagrams and essays and again compared the results. After the second round of training she was assigned a final round of 30 diagram/essay pairs and graded them independently. The complete process took roughly 3-4 months. An overview of the assignments was shown in Table 4.1 (pp. 47).

The diagram-essay pairs were chosen from the set of 105 covered by the primary grader. As with the primary grader the diagrams and essays were graded anonymously and independently of one-another. The grader was not informed that the diagrams and essays were matched sets and graded the diagrams and essays in separate batches. As with the primary grader this process was carried out online using SNG for the essays and questions and LASAD for the diagram display.

4.3.3 Grading Results

An overview of the grading results for the *primary* grader can found in Table 4.2 (pp. 51). As that table shows the grader ranked the diagrams relatively low with 10 of the 14 diagram grades having a negative average score as compared to two of the essay grades. Moreover, a visual inspection of the grade histograms for the diagrams found that 6 of the 14 had an unbalanced distribution with almost all of the students receiving the same score. The number of items receiving the majority score for each diagram and essay is indicated by the '#M' column. As that column shows in four cases (*G.01 (RQ-Quality)*, *G.02 (RQ-Link)*, *G.05 (Hyp-Link)*, & *G.09 (RQ-Open)*) more than 80 of the 105 diagrams received the same grade while for two grades (*G.10 (Hyp-Open)*, & *G.11 (Study-Novel)*) a majority of the 105 were assigned the same grade. This only occurred for two of the essay grades: *E.09 (RQ-Open)* & *E.10 (Hyp-Open)*. As the table shows the remaining scores, while well-distributed, are frequently skewed with the graph grades being positively skewed (mean and median closer to the minimum score than the maximum) and the essay grades were frequently negatively skewed (mean and median closer to the max score). This skewness does not affect the calculations described below.

Table 4.3 (pp. 52) contains the same type of overview for the *reliability* grader. In this case 9 of the diagram grades had a negative mean score and the same 6 had unbalanced distributions with four of the six (*G.01 (RQ-Quality)*, *G.02 (RQ-Link)*, *G.05 (Hyp-Link)*, *G.09 (RQ-Open)* & *G.11 (Study-Novel)*) having more than 20 diagrams assigned to the majority class and one, *G.10 (Hyp-Open)*, with more than half. For the essay grades five of the six had more than 20 essays assigned to the majority class (*G.01 (RQ-Quality)*, *G.02 (RQ-Link)*, *G.05 (Hyp-Link)*, *G.09 (RQ-Open)* & *G.10 (Hyp-Open)*) and one with half *G.13 (Arg-Convincing)*.

For the most part the distribution of minimum, median, mean, maximum, and standard deviation scores was similar between the primary and reliability graders. In both cases the students scored worse on the diagrams in general. While somewhat disappointing this is not entirely surprising as students carried out the diagramming and essay writing serially. Therefore I would expect that the essay would score marginally higher than the diagram if

only due to refinement. The primary issue of concern is the unbalanced score distributions, particularly on questions relating to the research question and the novelty of the argument. It is possible that these scores reflect the students' correct performance or the limitations of the ontology. I will address this point more fully in Subsection 4.5.1 (pp. 57).

4.4 RELIABILITY H_{H1} & H_{H2}

4.4.1 Primary Results

As indicated by [10] there are a number of competing definitions of reliability. For the present study my goal was to show that, given suitable training, the diagrams and essays can be reliably graded obtaining consistent scores. Therefore I conducted a direct comparison between the graph and essay grades assigned by the reliability grader and those of the primary grader. The raw results and comparison calculations are shown in Table 4.4 (pp. 53).

Due to the discrete ordinal structure of the grades I tested the level of agreement using Spearman's Rank Correlation Coefficient (ρ). ρ is a nonparametric test of correlation that ranges from -1 to 1 with larger values indicating stronger relationships and the sign indicating the direction [29, 130, 137]. Unlike Pearson correlation or similar parametric measurements ρ is less sensitive to the exact shape of the correlation and does not require an exact linear relationship.

As Table 4.4 (pp. 53) shows the graders obtained marginally-significant ρ scores ranging from 0.41 to 0.72 for all but two of the graph grades. For grades *G.12 (Arg-Coherent)* and *G.13 (Arg-Convincing)* the scores were marginally-significant ($\rho = 0.32, p < 0.09$) and ($\rho = 0.3, p < 0.1$) respectively. Given the strong agreement on the diagram grades it is clear that they can be graded reliably given suitable training. Therefore the null hypothesis H_{h1} does *not* hold in general.

This conclusion, however comes with some caveats. The diagram correlations are non-parametric therefore they monotonically correlate but this does not mean that they are exact. Moreover the correlations, while significant, are not perfect indicating that there are

Table 4.2: Overview of Primary Grader Results including Min, Median, Mean & Max Scores, as well as Standard Deviations (σ) and the # of items assigned to the majority score class (#M) (N=105).

Question		**Graph Questions**						**Essay Questions**					
		Min	Med	Mean	Max	σ	#M	Min	Med	Mean	Max	σ	#M
01	RQ-Quality	-2	-2	-1.47	2	1.16	**84**	-2	0.5	0.31	2	1.25	24
02	RQ-Link	-2.5	-2.5	-1.84	2	1.42	**84**	-2.5	1	0.66	2	1.44	29
03	RQ-Support	-1	0.5	0.60	2	0.68	33	-1	1	1.15	2	0.79	29
04	Hyp-Testable	-2	0.5	0.39	2	1.15	24	-2	1.5	1.07	2	0.97	28
05	Hyp-Link	-2.5	-2.5	-1.92	2	1.38	**87**	-2.5	1.25	0.69	2	1.61	36
06	Cite-Conclusions	-1.5	0.5	0.31	2	0.95	21	-1	1.5	1.27	2	0.75	37
07	Cite-Reasons	-1.5	0.5	0.28	2	1.03	20	-1	1	1.10	2	0.80	30
08	Claim-Support	-2.5	0	-0.25	2	1.09	24	-1.5	1	0.78	2	1.06	23
09	RQ-Open	-2.5	-2.5	-2.24	1	0.73	**83**	-2.5	-2	-0.94	2	1.47	*56*
10	Hyp-Open	-2.5	-2	-1.05	2	1.39	*53*	-2	-2	-0.78	2	1.44	*57*
11	Study-Novel	-2.5	-2.5	-1.79	2	1.12	*64*	-2	0.5	0.52	2	0.95	24
12	Arg-Coherent	-2	0	-0.10	1.5	1.04	23	-2	1	0.89	2	0.95	24
13	Arg-Convincing	-2	0	-0.34	1.5	1.10	20	-2	1	0.62	2	0.91	30
14	Arg-Quality	-4	-1.75	-0.76	4.5	2.48	14	-5	3.25	2.46	5	2.16	23

Table 4.3: Overview of Reliability Grader Results including Min, Median, Mean & Max Scores, as well as Standard Deviations (σ) and the # of items assigned to the majority score class (#M) (N=30).

		Graph Questions						**Essay Questions**					
Question		Min	Med	Mean	Max	σ	#M	Min	Med	Mean	Max	σ	#M
01	RQ-Quality	-2	-2	-1.58	1	0.87	**24**	-2	-2	-1.45	2	1.11	**23**
02	RQ-Link	-2.5	-2.5	-1.93	1.5	1.20	**24**	-2.5	-2.5	-1.75	1.5	1.41	**23**
03	RQ-Support	-2	0.5	0.53	2	1.14	6	-1	1.25	1.02	2	0.85	7
04	Hyp-Testable	-2	0.5	0.25	2	1.20	6	-2	0.25	0.38	2	1.05	7
05	Hyp-Link	-2.5	-2.5	-2.27	0.5	0.76	**26**	-2.5	-2.5	-1.82	1.5	1.35	**23**
06	Cite-Conclusions	-2.5	1	0.60	2	1.48	9	-1.5	1.25	1.08	2	0.82	9
07	Cite-Reasons	-2.5	0.5	0.45	2	1.31	7	-1.5	1	0.73	2	0.97	7
08	Claim-Support	-2.5	0	-0.35	2	1.54	6	-2	1	0.45	2	1.15	10
09	RQ-Open	-2.5	-2.5	-2.28	1	0.67	**24**	-2.5	-2.5	-2.38	-2	0.22	**23**
10	Hyp-Open	-2.5	-2	-1.42	1	1.15	*16*	-2.5	-2	-1.70	2	0.92	**25**
11	Study-Novel	-2.5	-2.5	-1.77	2	1.40	**21**	-2	0	0.08	2	1.11	8
12	Arg-Coherent	-2	-0.5	-0.23	2	1.20	6	-0.5	0.5	0.73	2	0.72	9
13	Arg-Convincing	-2	0	0.00	1.5	0.90	10	-1	0	0.43	1.5	0.69	*15*
14	Arg-Quality	-5	0	-0.32	4.5	2.63	6	-2.5	1	1.07	4.5	2.15	7

Table 4.4: Final results for the reliability comparison. This includes a summary of score values and ρ comparisons (n=30).

		Graph Grades G.01 - G.14						Essay Grades E.01 - E.14					
		Primary		Validation		Spearman		Primary		Validation		Spearman	
	Name (G/E)	mean	σ^2	mean	σ^2	ρ	sig	mean	σ^2	mean	σ^2	ρ	sig
01	RQ-Quality	-1.68	0.97	-1.58	0.87	**0.66**	$p < 0.01$	0.6	1.15	-1.45	1.11	**0.42**	$p < 0.02$
02	RQ-Link	-2.22	0.95	-1.93	1.2	**0.64**	$p < 0.01$	1.07	1.09	-1.75	1.41	0.15	$p < 0.44$
03	RQ-Support	0.33	0.95	0.53	1.14	**0.44**	$p < 0.01$	1.13	0.78	1.02	0.85	0.1	$p < 0.6$
04	Hyp-Testable	0.15	1.33	0.25	1.2	**0.5**	$p < 0.01$	0.9	1.0	0.38	1.05	**0.62**	$p < 0.01$
05	Hyp-Link	-2.22	1.08	-2.27	0.76	**0.68**	$p < 0.01$	1.03	1.36	-1.82	1.35	0.14	$p < 0.45$
06	Cite-Conclusions	0.1	1.26	0.6	1.48	**0.63**	$p < 0.01$	1.33	0.67	1.08	0.82	0.16	$p < 0.39$
07	Cite-Reasons	0.13	1.26	0.45	1.31	**0.59**	$p < 0.01$	1.2	0.75	0.73	0.97	**0.6**	$p < 0.01$
08	Claim-Support	-0.57	1.12	-0.35	1.54	**0.41**	$p < 0.03$	0.72	1.03	0.45	1.15	*0.35*	$p < 0.06$
09	RQ-Open	-2.42	0.23	-2.28	0.67	**0.5**	$p < 0.01$	-0.97	1.47	-2.38	0.22	-0.04	$p < 0.82$
10	Hyp-Open	-1.15	1.44	-1.42	1.15	**0.72**	$p < 0.01$	-0.75	1.47	-1.7	0.92	**0.55**	$p < 0.01$
11	Study-Novel	-1.92	0.97	-1.77	1.4	**0.72**	$p < 0.01$	0.7	0.91	0.08	1.11	0.29	$p < 0.12$
12	Arg-Coherent	-0.25	1.13	-0.23	1.2	*0.32*	$p < 0.09$	0.85	0.93	0.73	0.72	**0.5**	$p < 0.01$
13	Arg-Convincing	-0.48	1.17	0.0	0.9	*0.3*	$p < 0.1$	0.58	0.86	0.43	0.69	**0.46**	$p < 0.01$
14	Arg-Quality	-1.05	2.51	-0.32	2.63	**0.51**	$p < 0.01$	2.72	1.65	1.07	2.15	**0.56**	$p < 0.01$

some real deviations between the graders. As I will discuss in Chapter 8 (pp. 131) these differences may stem from real disagreements about the meaning of the structural concepts or the argument model. Nevertheless the correlations are cause for satisfaction but may not generalize to all cases.

For the Essay grades the answers were more problematic with statistically-significant scores ranging from 0.42 to 0.6 on 7 of the 14 grades and a marginally significant ($\rho = 0.35, p < 0.06$) on *E.08 (Claim-Support)*. The remainder of the grades did not find significant agreement. The unreliable grades were: *E.02 (RQ-Link)*, *E.03 (RQ-Support)*, *E.05 (Hyp-Link)*, *E.06 (Cite-Conclusions)*, *E.09 (RQ-Open)*, and *E.11 (Study-Novel)*. Of these all of them are focused on the relationships *between* the argument components.

These structural relationships are difficult for novice arguers and can easily be lost in the essay format while the diagrams were explicitly designed to reify them. Moreover the relevant textbooks such as [9] and [40] and the course lectures discuss individual argumentation components in detail. Thus there is relatively widespread agreement on how individual components such as hypotheses and citations should be framed in research reports and the students received explicit instruction on how to do so, more than they received on the logical relationships.

Therefore it seems likely that the low reliability on these scores can be attributed to the lack of clear guidance on how such relationships should be expressed in an essay, and the lack of clear conventions for the graders. In the absence of such guidance the students either failed to include the relationships, already encoded in their diagrams, in the essays or did so inconsistently. In the absence of clear conventions the graders could not obtain agreement. Thus it is clear that agreement is possible but not for all of the criteria. Therefore hypothesis H_{h2} is rejected in general but clearly some of the essay grades were not reliable. This reliability has important consequences that I will discuss below.

Moreover the important caveats discussed above continue to apply. These correlations are nonparametric and are consequently less sensitive to direct disagreements. Thus while these results are sufficient to establish general reliability they should not be read as guaranteeing complete agreement.

4.4.2 Reliability Filtering

In order for any statistical analysis to be useful it is necessary to ensure that the manually-assigned grades are sufficiently reliable. In the remainder of this thesis I will be using the manually-assigned graph grades solely to predict the essay grades as discussed in Chapter 7 (pp. 97). As noted above the graders found statistically-significant or marginally-significant agreement for all of the diagram grades. Therefore they are sufficiently reliable for use as predictive variables.

The reliability of the essay grades are a more complex issue. In general 7 of the 14 essay grades were reliable: *E.01 (RQ-Quality)*, *E.04 (Hyp-Testable)*, *E.07 (Cite-Reasons)*, *E.10 (Hyp-Open)*, *E.12 (Arg-Coherent)*, *E.13 (Arg-Convincing)*, and *E.14 (Arg-Quality)*. There was also marginally-significant agreement on 1: *E.08 (Claim-Support)*. The relevant ρ scores are shown in Table 4.5 (pp. 56), third column.

However, given the fact that this work focuses on paired diagrams and essays individual essay reliability is a minimum standard. For the present analysis I will set a higher threshold for reliability of the essay grades that incorporates the reliability of the associated graph grades. This *paired reliability* will be estimated as a multiple of the statistically-significant ρ scores for the corresponding grades of the form ($\tau = (\rho_{g_i}\ \&\ \rho_{e_i})$). Thus the threshold value for *E.14 (Arg-Quality)* is $\tau_{E.14} = (0.51 * 0.56) = 0.29$. Correlations that are not statistically- or marginally-significant will not be used.

Therefore an essay grade will be considered reliable if and only if the graders found statistically-significant or marginally-significant agreement on both the relevant graph and essay grades *G.x* and *E.x* and if the combined reliability score $\tau_{E.x}$ meets or exceeds 0.2. This calculation is shown in Table 4.5 (pp. 56). As shown there only 5 of the 14 grades meet these standards: *E.01 (RQ-Quality)*, *E.04 (Hyp-Testable)*, *E.07 (Cite-Reasons)*, *E.10 (Hyp-Open)*, and *E.14 (Arg-Quality)*. Therefore I will focus on these grades as dependent variables for the remainder of this thesis.

Table 4.5: Reliability Filter calculations including ρ scores and $\tau_{E.x}$ scores for the manually-assigned graph and essay grades. **Bold** scores represent statistically-significant ρ values while *italics* represent marginally-significant results.

Name	Reliability ρ		$\tau_{E.x}$	Decision
	Graph	Essay		
01 *RQ-Quality*	**0.66**	**0.42**	**0.28**	**Keep**
02 *RQ-Link*	**0.64**			
03 *RQ-Support*	**0.44**			
04 *Hyp-Testable*	**0.5**	**0.62**	**0.31**	**Keep**
05 *Hyp-Link*	**0.68**			
06 *Cite-Conclusions*	**0.63**			
07 *Cite-Reasons*	**0.59**	**0.6**	**0.35**	**Keep**
08 *Claim-Support*	**0.41**	*0.35*	0.14	
09 *RQ-Open*	**0.5**			
10 *Hyp-Open*	**0.72**	**0.55**	**0.4**	**Keep**
11 *Study-Novel*	**0.72**			
12 *Arg-Coherent*	*0.32*	**0.5**	0.16	
13 *Arg-Convincing*	*0.3*	**0.46**	0.14	
14 *Arg-Quality*	**0.51**	**0.56**	**0.29**	**Keep**

4.5 VALIDITY: H_{H3} & H_{H4}

For the present purposes there are two relevant measures of validity: direct correlation of paired grades of diagrams and essays, and correlation with gestalt scores. For the former we test whether the corresponding question grades in the paired rubrics are correlated that is, whether *G.01 (RQ-Quality)* is significantly correlated with *E.01 (RQ-Quality)*. For the latter the question is whether the individual graph grades are significantly correlated with the reliable *gestalt* essay grade *E.14 (Arg-Quality)*. That is, do the graph features correlate in any way with the overall essay quality. I will discuss both definitions below focusing solely on the five reliable essay grades. As before I will rely on Spearman's ρ as a measure of correlation.

4.5.1 Direct Validity

As shown in Table 4.6 (pp. 57) statistically significant positive correlations were found between the graph and essay grades for 4 of the 5 reliable questions *G/E.04 (Hyp-Testable)*, *G/E.07 (Cite-Reasons)*, *G/E.10 (Hyp-Open)*, and *G/E.14 (Arg-Quality)*. In all cases the correlations were positive and ranged from $\rho = 0.254$ for question *G/E.12 (Arg-Coherent)* to $\rho = 0.443$ for question *G/E.07 (Cite-Reasons)*. Interestingly *G/E.01 (RQ-Quality)* had a strongly unbalanced distribution with the primary grader scores having median scores of *-2* and 0.5 for *G.01* and *E.01* respectively (see Table 4.2 (pp. 51)). This was not the case for three of the other four scores where the graph grade median was closer to that of

Table 4.6: A listing of the direct validity comparison results including ρ and p-values (n=105).

Questions	Speaman's ρ	sig
G.01/E.01 (RQ-Quality)	r=-0.102	p<0.299
G.04/E.04 (Hyp-Testable)	**r=0.254**	**p<0.009**
G.07/E.07 (Cite-Reasons)	**r=0.443**	**p<0.001**
G.10/E.10 (Hyp-Open)	**r=0.254**	**p<0.009**
G.14/E.14 (Arg-Quality)	**r=0.331**	**p<0.001**

the essay grades. Only *G/E.14* was similarly unbalanced with relative median scores of *-1.75* and *3.25* respectively While the direct correlations are not maximal ($\rho = 1$) they are relatively strong particularly for *G/E.07 (Cite-Reasons)*, and *G/E.14 (Arg-Quality)*. Therefore while it appears that the unbalanced results may have adversely impacted the direct validity calculations it is also apparent that the parallel rubrics do correlate with one-another making the some of the paired grades directly valid. Thus H_{h3} is rejected. These nonparametric correlations give us reason to believe that performance is, in fact, predictive.

It is important to note, however, that they should not be read as showing that diagram performance improves writing only that performance is correlated across the board.

4.5.2 Gestalt Validity

As shown in Table 4.7 (pp. 58), statistically significant positive correlations were found between grades *G.04 (Hyp-Testable)*, *G.07 (Cite-Reasons)*, *G.10 (Hyp-Open)*, and *E.14 (Arg-Quality)*. *G.01 (RQ-Quality)* was, again, uncorrelated. As above these results are not exact but show general performance agreement and are conditioned by the fact that the gestalt grades were assigned by the same grader as the individual grades. Indeed the grader was encouraged by the structure of the assignment to consider the specific components and then the gestalt grades thus this level of agreement was expected and is satisfying. Therefore most of the reliable graph grades also satisfy a test for gestalt validity and H_{h4} is rejected.

Table 4.7: Gestalt Validity Comparisons using Spearman's Rho. This lists direct nonparametric correlations between the reliable graph grades and *E.14 (Arg-Quality)*.

Questions	Names	ρ	Sig
G.01-E.14	(RQ-Quality – Arg-Quality)	r=-0.075	p<0.449
G.04-E.14	(Hyp-Testable – Arg-Quality)	**r=0.237**	**p<0.015**
G.07-E.14	(Cite-Reasons – Arg-Quality)	**r=0.419**	**p<0001**
G.10-E.14	(Hyp-Open – Arg-Quality)	**r=0.237**	**p<0.015**

4.5.3 Summary & Analysis

As noted above the grade pair *G/E.01 (RQ-Quality)* was neither directly valid nor did it have gestalt validity. The remaining four reliable pairs, *G/E.04 (Hyp-Testable)*, *G/E.07 (Cite-Reasons)*, *G/E.10 (Hyp-Open)*, & *G/E.14 (Arg-Quality)* however were both reliable and valid.

Interestingly, the four valid reliable grades can be readily classified by the most relevant argument component. Grade *G/E.07*, focuses on the students' use of citations and their connection with other features. Grades *G/E.04* & *G/E.10* deal primarily with hypotheses. And *G/E.14*, of course, deals with the gestalt quality. By contrast grade *G/E.01* focuses on the central research question and had an unbalanced distribution. As with the essay reliability it is possible that this lack of validity can be attributed to the structure of the diagrams and to the lack of instructional support.

Research questions are an essential part of a well-written research report. They serve to generalize the research connecting it to larger problems and provide a basis to discuss the openness of the research more broadly. As I noted in Section 3.2 research questions are not represented directly in the diagramming ontology but via a precisely-formatted claim node. Expressing the novelty of the study requires a similarly complex process. Students were instructed to defend the novelty of their study by generating a current study node that summarized their study methodology and other key features of the work and then to draw comparison arcs between this node and the citations. They were then instructed to use these arcs to state distinctions between their work and the cited works. Thus there was no single node used to defend the full novelty of the study and, as I will discuss in later chapters, the students made little use of the current study nodes and comparison arcs.

While this was discussed with the students during the LASAD introduction subsequent discussions with the TAs indicated that they did not reinforce the concepts of novelty and comparison of related citations nor did they support students in the inclusion of research questions or current-study nodes in their diagrams. Most chose instead to focus students attention on the use of hypothesis statements, citations, and the connections between them.

In the absence of explicit framing by the diagramming ontology and persistent guidance

from the primary point of contact, the TAs, it is unsurprising that most of the students omitted research questions from their diagrams. It is also unsurprising that they did not make efforts to distinguish their study from prior work at that time. The resulting unbalanced distribution is thus unable to reflect the real variation in the associated essay grades. The one valid yet unbalanced grade *G.10 (Hyp-Open)* was only marginally so and thus variable enough to apply.

4.6 CONCLUSIONS

Having individually rejected all four of the null hypotheses introduced in Section 4.1 it is clear that the general question Q_h holds true. The diagrams can be reliably graded by human graders and the resulting grades are significantly correlated with the subsequent essay grades. Indeed it appears that the reification offered by the diagrams makes them easier to grade reliably than the comparable essays. Therefore the argument diagrams are valid structures for argument planning and a viable target for intervention by expert instructors so long as an appropriate set of rubrics is used. Clearly the parallel rubrics designed here were appropriate in general. These conclusions, however, are conditioned by the limits both on both the reliability and validity results that I discussed above. Based upon these limitations it is not clear that all of the grades are suitable targets for automated analysis. I will address this point again in Section 5.5.

5.0 Q_A AUTOMATIC GRADING

Q_a Can argument diagrams be analyzed *automatically* to *diagnose* students' argumentation skills and to *predict* future performance on "real-world" tasks?

5.1 INTRODUCTION

In this part of the thesis I will focus on question Q_a. Having shown that argument diagrams can be evaluated reliably by human graders and that the resulting grades are valid predictors of subsequent performance, it remains to be seen whether or not the diagrams can be subjected to automatic analysis. Such analysis is useful if it is possible to define a graph model, such as graph rules or neural networks, that can be used to *diagnose* individual student problems by identifying specific conceptual errors or problematic structural features that correlate with future performance. For the present study I will focus on specific a-priori graph rules. Such rules can be used as the basis for direct feedback or other student guidance. An individual rule or graph structure that has been shown to correlate with future behavior is *empirically-valid*. Automatic analysis is also useful if we can use it to reliably *predict* student performance. Reliable predictive models, if available, can then be used to rank student diagrams and to flag students who require additional support. Thus Q_a can be broken down into two specific null hypotheses:

H_{a1}**:** It is not possible to define *empirically-valid* diagram rules that correlate with students' novel written argumentation ability.

H_{a2}**:** Automatic features of student diagrams *can not* be used to predict students' novel written argumentation ability.

The remainder of this chapter will focus on background material relevant to H_{a1} & H_{a2}. This will include: a brief discussion of relevant prior work in Section 5.2, an introduction to Augmented Graph Grammars (Section 5.3), a summary of the rules used in the present work (Section 5.4), and a detailed discussion of the graph and essay grades used for automatic analysis (Section 5.5). The two hypotheses will then be addressed in the following chapters with H_{a1} being tested in Chapter 6 and H_{a2} in Chapter 7.

5.2 PRIOR WORK

As I discussed in Section 2.2 several existing argumentation systems such as LARGO, [91, 88, 90], LASAD [103], and Belvedere [117] employ graph grammars or other rules to provide automated graph analysis. The rules are used to trigger automatic guidance such as hint messages and to enforce higher-level semantic and syntactic constraints and to address the more complex (mis)uses of the graph structures such as those shown in Figure 3.4 (pp. 38).

In LARGO, for example, students were instructed to draw arcs representing logical rule modification from one legal test node to another, not from a legal test to a current fact situation. Rules were encoded for the help system to flag when a student did so and to provide on-demand advice against it. Similar techniques have been used in other systems. Scheuer et al. survey such established advice techniques in [105]. In systems such as LARGO the rules are only used to provide at-will advice, not to impose hard syntactic constraints.

In general, the rules used are, like the ontologies, defined *a-priori* by domain experts based upon normative rules of argumentation and current pedagogical goals. The help rules used in LARGO were defined by Kevin Ashley, an experienced law professor, based upon a process model of argument and then encoded directly into the system [4, 6]. One key advantage of pedagogically-driven, at-will guidance of this type is that it can be tuned to the level of complexity desired by the instructors and students but can also be safely ignored by more experienced users. This latter feature was particularly desirable in LARGO as the students were tasked with diagramming existing dialogues that did not always conform to the ideal model. This makes such rules ideally-suited to open-ended tasks such as argumentation

and to ill-defined domains [74].

While such rules have strong pedagogical support, little work has been done to empirically validate them using student data or to assess their individual role in subsequent student performance. For the present discussion a rule is *empirically-valid* if it correlates with subsequent measures of student performance. Thus, if a rule is designed to detect problems with an argument diagram the rule would be empirically valid if violations of the rule correlate with poorer scores on subsequent arguments. Thus empirical validity is defined relative to a given task and grading metric. In general it is assumed that rules with strong pedagogical support are empirically valid but this is not always tested directly. Crucially, empirical validity is not necessarily the same as predictiveness. As with the inter-grader reliability and validity discussed in Chapter 4 the goal is to demonstrate that the rules correlate with desired behavior but not necessarily that an individual rule can be used to accurately predict subsequent performance.

In a prior study my colleagues and I used decision trees to predict students' post-test performance based upon features of their LARGO diagrams [70]. The features used in that study were the LARGO help rules themselves. Comparisons were made based upon the rule counts with the goal of classifying students into good and poor performers, that is, above and below the mean score on a subsequent test of argument *comprehension*. In the course of that study we found that some of the graph features were individually and collectively predictive and that both C4.5 and Genetic Programming yielded successful decision trees. A decision or comparison tree is a rule-based classifier that operates by recursively partitioning cases based upon variable values [56]. Each internal node in the tree represents a feature variable with the outgoing paths representing possible values of the target variable. Leaf nodes encode possible classifications. The induced trees in this study succeeded in accurately classifying 100% of the test cases. Interestingly the study also found that one of the error rules was, in fact, positively correlated with student performance. The rule was a more complex error and as such was only triggered by advanced students.

In a more recent study reported in [68] we compared open-ended student-produced argument diagrams with diagrams produced by a domain expert. The comparison was made based upon the order, size, density, and other basic graph features. No attempt was made

to combine features as in [70]. The individual features were also used to predict students' subsequent performance on a writing assignment for which the diagram served as an outline. The study found that some features, notably the length of the summative text in the citations, was predictive of subsequent performance and served to distinguish student from expert diagrams. These features, however, were simple graph components and did not include any complex a-priori semantic rules as we use here. Moreover the final grade being predicted was a single letter grade that covered the entire assignment, including the writing quality, argument, scope of cited materials, and communication. As such it was driven by many factors beyond the written argument quality.

Thus while these studies have been successful and appear to support general question Q_a, neither one was directed to the specific question at hand. The study reported in [70] focused on representational note-taking diagrams drawn from a set of shared transcripts and argument comprehension scores as an outcome measure. In [68], by contrast, no complex rules of the type described below were tested and the outcome measure was an overall grade, not just an argumentative one. The present study extends the prior work by focusing on open-ended argument diagrams and on argument-specific grades.

5.3 AUGMENTED GRAPH GRAMMARS

As noted previously, Augmented Graph Grammars are a rule formalism for classification and mining. For the present work I elected to implement all of the graph analysis rules using this formalism. These rules, described below, were designed *a-priori* with the guidance of domain experts and examination of prior data. The rules were not used to provide help to the students during the course of the study but were solely used for the analyses described below. In Section 2.6 (pp. 25) I briefly introduced Graph Grammars, a formal production grammar defined over graphs and subgraphs. Basic graph grammars represent formal graph productions that are analogous to context-free string grammars (see [111]) and were defined by Rekers and Schürr [97] in terms of individual productions.

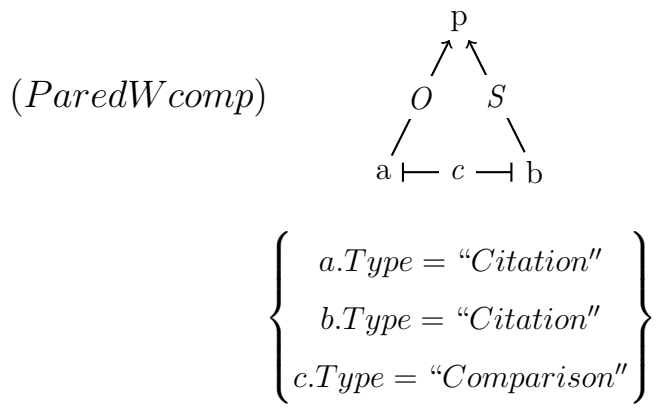


Figure 5.1: A simple augmented graph grammar rule that detects paired counterarguments. The rule shows a single parent node related to two citation nodes that both disagree and are compared via a comparison arc. The rule depends in turn upon two recursive rule productions shown in Figure 5.2 (pp. 79) and Figure 5.3 (pp. 80)

Graph grammars represent rules as graph structures of the type shown in Figure 5.1 (pp. 65) making them a natural representation for graph matching and classification. As with standard string grammars (see [111]) they can be used to match corresponding graphs by recursive matching (e.g. [146]) where the nodes and arcs within the rules are matched to components of the target diagram subject to the constraints of arc direction and type or content constraints. Graph grammars allow for recursive rule productions of the type shown in Figures 5.2 - 5.3 which map variable nodes and arcs, identified by capitalized names, to appropriate subgraphs.

Graph Grammars can be evaluated using standard graph-matching algorithms such as tree-search and, while arbitrary graph matching is NP-Hard, some heuristic restrictions such as layered graph grammars make the search process tractable. In the present case rules such as the one shown in Figure 5.1 are matched to diagrams like the one shown in Figure 5.4 (pp. 81). The matching is completed in a top-down manner with the code first matching rule node p to hypothesis node *#2* in the diagram. It would then map rule node a to citation

node *#15* and rule node *b* to citation node *#6*. Having mapped the two nodes the mapper would automatically map rule arc *c* to the comparison arc *#21* and then recursively map the two production arcs *O* and *S*. Arc *O* would be mapped using case O_{p3} to the single opposing arc *#16* while rule arc *S* would require a recursive application of subrules S_{p1} and S_{p2} to cover arcs *#19* and *#21*.

Basic graph grammars such as those defined by Rekers and Schürr and in [51] are focused on fixed diagrams with a limited alphabet of node and arc types. Augmented graph grammars, by contrast, allow for the use of an argument ontology with complex element types that include subsidiary features such as textual fields. The LARGO process model was initially defined in terms of graph grammars before being implemented in Java [89]. The augmented graph grammars used here allow for named *textual fields*, and *access functions* as well as the basic types and directional information. Formally, these additional features impose additional complex constraints on the nodes and arcs to be matched in the rules as well as shared constraints that cover subgraphs. An named production of the type used in this study is shown in Figures 5.5 (pp. 82) - 5.6 (pp. 83). This is a typical example of an Augmented Graph Grammar structure. As with the grammars defined by Rekers and Schürr, the formalism used here makes some restrictive assumptions such as a requirement that the productions be "expansive" or layered. A more detailed discussion of the formal structure can be found in Appendix F (pp. 223).

Graphs that can be collapsed according to a given set of productions are thus matches for the language of graphs described by it. I define a set of one or more named graph productions as a graph *rule*, and any subgraph that can be collapsed by a given rule is a *feature*. In the discussion below, I will refer to simple and complex graph features. These are, in turn, defined by appropriate rules. The production shown in Figures 5.5 (pp. 82) - 5.6 (pp. 83) therefore represents a single self-contained graph rule.

I have developed an augmented graph grammar engine called AGG. This engine implements the augmented graph grammar formalism described in Appendix F (pp. 223). With the exception of the chained features, all of the complex feature rules listed below were implemented using this engine. In general, Augmented Graph Grammars have proven to be a

and robust formalism for the definition of graph rules and rule analysis, but some tuning for efficiency is required.

5.4 GRAPH FEATURES AND GRAPH GRAMMARS

The present analysis is concerned with computable features of the argument diagrams. For analytical purposes the features were divided into two classes: *simple* features such as the number and type of components present, and *complex* features such as disjoint subgraphs and chained counterarguments. I will describe these two types below.

5.4.1 Simple Features

The 34 simple graph features are basic graph-theoretic structures (see [11]) that are relevant to all graphs. The simple features fall into four types: Size and Density, Ontology, Textual, and Visual. I will describe each class in turn and illustrate them with reference to the example shown in Figure 5.7 (pp. 84).

Size and Density such as the *order* of the graph (number of nodes) and *size* (number of arcs); the average number of child nodes or neighbors of a given node, and so on.

- **Order**: The number of nodes in the graph. Figure 5.7 has 5.
- **Size**: The number of arcs in the graph. Figure 5.7 has 4.
- **Average/Min/Max Children**: The average, minimum, or maximum number of child nodes in graph. A child node is a node that has an arc *to* a parent. For undirected nodes the child status is based upon the order in which the arc was drawn. In Figure 5.7 claim node #4 has 4 children, 3 of which are shown.
- **Average/Min/Max Children Ignore Empty**: Average number of nonempty child nodes. Nodes are empty if they have no text in the field(s). Figure 5.7 contains no empty nodes.
- **Average/Min/Max Parents**: average number of parent nodes in graph. Node 16 in Figure 5.7 has 1 parent.

- **Average/Min/Max Parents Ignore Empty**: Average number of nonempty parent nodes over the whole diagram.
- **Average/Min/Max Degree (δ - Δ)**: Average number of neighbor nodes for a graph over the whole diagram (both parent and child). Node 4 has a degree of 5 while node 16 has $\delta = 2$.

Ontology Element the presence and number of specific ontology features such as the number of *citation* and *current-study* nodes as well as the number of *supporting* and *opposing* arcs.

- **Elt_Claim**: Number of claim nodes in the graph.
- **Elt_Citation**: Number of citation nodes in the graph.
- **Elt_Hypothesis**: Number of hypothesis nodes in the graph.
- **Elt_CurrStudy**: Number of currstudy nodes in the graph.
- **Elt_Supporting**: Number of supporting nodes in the graph.
- **Elt_Opposing**: Number of opposing nodes in the graph.
- **Elt_Undefined**: Number of undefined nodes in the graph.
- **Elt_Comparison**: Number of comparison nodes in the graph.

Textual Features focus on the content of the nodes and arcs, specifically the length of student summaries and other info on a per-field basis.

- **Avg/Min/Max Field Sent Len**: The length of the fields in terms of sentences. In Figure 5.7 every visible node and arc has a field sentence length of 1.

- **Minus citation Avg/Min/Max FieldSentLen**: The length of text fields in terms of sentences minus all citation nodes. This measurement was developed based upon the work reported in [68] and is intended to test the distinct influence of the, often more detailed, citation nodes.

Visual Features focus on layout issues such as the number of overlapping nodes and crossing arcs or general *messiness*. The importance of these features was suggested by both student reviewers and graders who consistently cited messiness as a problem when examining diagrams.

- **Crossed Arcs**: Number of arcs that cross one-another in the diagram.
- **Overlapping Nodes**: Number of node boxes that overlap in the diagram.
- **Visual Messiness**: CrossedArcs + Overlapping Nodes.

5.4.2 Complex Features

In systems like LARGO and Belvedere help is driven by complex pattern-matching rules designed to detect diagram features that violate the argument process model (see [4]) or other higher-level semantic constraints (e.g. a citation supporting another citation). These rules are typically drafted by a domain expert and used without any independent empirical validation. A similar development process was taken here. The 43 complex features represent violations of the argumentative norms (e.g. unfounded claims), or pedagogically important features (e.g. paired counterarguments), and were identified with the help of psychology domain experts including the course instructors for Research Methods. Rules to match the features were then implemented as augmented graph grammars. The rules *were not* used to provide help to the students during the data collection phase of the study but are being used for that purpose in subsequent courses.

A number of the rules overlap or cover similar features. Some, for example, are designed to accept arcs of any direction while other rules make the more restrictive requirement of directed versions. The overlapping rules were designed to test the impact of arc direction on the results.

The complex features can be classified into the following types: Chained, Single Component, Node/Arc Pair, Text Field, Triplet, Grounding, & Disjoint Subgraph. I will summarize each type below along with selected grammar examples and pseudocode for some of the relevant rules.

Chained Features are specifically focused on the use of counterarguments to present complex support and opposition structures. Unlike the other complex feature rules these were not implemented as graph grammars.

- **Paired Counterarguments**: Paired counterarguments are a graphical structure representing disagreement over a node. The structure consists of a single *subject* parent node and two *disputants* or child nodes one of which is connected to the parent via a supporting arc and the other via an opposing arc. A corresponding graph-grammar rule can be seen in Figure 5.8 (pp. 85) (Paired). A structure of this type can be seen in Figure 5.7 consisting of node #4 (subject), node #28 (supporting disputant via arc #29), and node #16 (opposing disputant via arc #17).

  During the instruction phase of the 2011-PittRM study, students were shown examples of paired counterarguments and were instructed to use them to represent disagreement. They were further instructed to use the comparison arc to draw analogies and distinctions between the children if they were citation or current study nodes.

- **Chained Argument Nodes**: Chained counterarguments are an alternative argumentation structure consisting of a chain of three nodes. A root subject node with an opposing child node which is itself opposed by a subsequent child node. A graph-grammar rule for this feature can be seen in Figure 5.8 (Chained). No examples of this feature are shown. This structure was *not* described to students during the 2011 RM study. However, the domain experts identified it as an important argumentative structure and it was included in the instructional materials during a subsequent study.

Single Component Features focus on the existence or absence of a single feature. As such they are similar to the simple ontology features described above. They also include negated rules that record the nonexistence of a given component and complex components such as the Research Question. It also includes some simple rules that duplicate the simple ontology features. These were developed for validation purposes. They can be grouped into three groups shown below.

- The *R01p*_Has_ (CurrStudy/Hypothesis/Claim/Cite)* rules replicate the simple ontology node features exactly, recording the presence of basic ontology items and were reimplemented for testing purposes.

- *R01pd_Has_RQ* tests for the existence of a "Research Question," that is, a claim node with the text inside of it framed as a question. For the present this detector simply checks for a nonempty node that ends with a '?'. Pseudocode and a graph grammar structure are shown in Figure 5.9 (pp. 85). Clearly this framing has some weaknesses that could be rectified with further conditions, but is consistent with, the instructions supplied to the students.

- *R01n*_No_(CurrStudy/Hypothesis/Claim/Cite/RQ)* rules check for the nonexistence of a basic ontology item and are binary with 1 indicating the item is not present or 0 if it is. This is simply a negation of the existing *R01p* rules. By definition these rules are binary values as shown in the pseudocode below:

  R01n_NoCurrstudy: *If* there exists no Current Study node in the graph *then* return 1 *else* return 0.

Node/Arc Pair Features focus on an individually desirable or undesirable node and arc relationships such as non-citation node with no inbound arc. These represent argumentative structures that the students were instructed to avoid but were not forced to do so.

- *R02: NonCurrStudy w/o outlink*: A non-current study node without any outgoing arc. This is problematic as the designated role of the current study node is to state features of the study (e.g. a focus on college students) that serve to frame the hypothesis, support claims, or distinguish it from other prior work. As such it should not be isolated but should be, at least, connected to one other part of the argument.

(Pseudocode) Count the number of nodes N in the graph such that: (1) N is not a current study node; *and* (2) there is no directed arc $\overrightarrow{e(N, X)}$ from N to another node X in the graph.

- *R02a: NonHypo w/o outlink*: A non-hypothesis node without an outgoing arc. Citations, claims, and current study nodes are introduced to frame the discussion and to support or oppose the working hypothesis. As such all of them should have an outgoing arc.

- *R02b: NonHypo/Claim w/o outlink*: A non-hypothesis *or* Claim node without an outgoing arc. This is a subset of the feature rule above, targeting only hypothesis and claim nodes.

- *R02c: NonHypo/RQ w/o outlink*: A non Hypothesis or research question node without an outgoing arc. A subset of the feature above targeting only hypothesis and research question nodes.

- *R03: Noncite w/o inlink*: A non-citation node without any incoming arc. Citations are used to provide evidence that supports or opposes other components. As such it is appropriate for them to have only outgoing arcs. It is not necessarily appropriate for any other type of node.

- *R03b: Noncite/CurrStudy w/o inlink*: A node that is neither a citation nor a current-study node without an incoming arc. This is a refinement of the rule above that also ignores current study nodes.

- *R08: Unsupported Hypo*: A hypothesis node without any incoming supporting arcs. Students are expected to present some claims and citations that support their hypotheses rather than simply stating them in isolation. This confirms that they have provided some form of support for it.

(Pseudocode) Count the number of nodes H in the graph such that: (1) H is a Hypothesis node; *and* (2) there are no incoming Supporting arcs $\overrightarrow{s(*, H)}$ coming to H from another node.

- *R08: Unopposed Hypo*: A hypothesis node without any opposition arcs inbound. If a hypothesis has no opposition then it is not open or in doubt. Students were instructed to find at least one opposing citation and this tests for the presence of an opposition arc to the hypothesis.

- *R10a: Hyp/Claim comp*: A hypothesis node compared to a claim node. Comparison arcs should be used to differentiate opposing citations or to draw distinctions between citations and the current study nodes. This rule tests for violations of that. In the preliminary study we observed students using the comparison arcs incorrectly despite instructions to the contrary.

  (Pseudocode) Count all pairs of nodes H & C such that: (1) H is a Hypothesis node; *and* (2) C is a Claim node; *and* (3) there exists a Comparison arc $m(H, C)$ between them.

- *R10b: CurrStudy/CurrStudy comp*: A pair of current study nodes with a comparison arc between them. As with *R10a* this is designed to detect improper use of the comparison arcs.

- *R10c: Claim comp*: A Claim node with a comparison arc to it. This is a less sensitive rule that tests for one half of the feature covered in *R10a*.

- *R10d: Hypothesis comp*: A hypothesis node with a comparison arc to it. Again, this is a less sensitive rule that tests for a subfeature of the one covered in *R10a*.

- *R12: Undefined Cite Claim*: Use of an undefined node to connect a citation or claim node to any node type. During the pre-study we found that students who used undefined nodes frequently appeared to be uncertain in their later conclusions. This rule was introduced to detect cases where a student appears to be unclear as to whether the source claim or citation backs its target. In the 2011 study the students were encouraged to use the undefined relationship for factual sources and definitions. In subsequent studies use of the undefined arc was generally discouraged.

  (Pseudocode) Count all pairs of nodes C & O such that: (1) C is a Citation or Claim node; *and* (2) there exists an Undefined arc $u(C, O)$ between them.

Text Field Features match empty required text fields within the nodes and arcs. The nodes and arcs are structural components and, absent containing text, have no semantic content.

- *R04a Empty Node Fields*: Any node with an empty text field.

  (Pseudocode) Count all nodes N such that: one or more of the required text fields in N is empty.

- *R04b Empty Arc Fields*: Any arc with an empty text field.

Triplet Features reflect the existence or nonexistence of relationships between two nodes.

- *R05: Hypo Supports Cite*: hypothesis node with a supporting arc to a citation. In general the hypothesis nodes should not have any outgoing arcs, particularly to citations.

  (Pseudocode) Count all pairs of nodes H & C such that: (1) H is a Hypothesis node; *and* (2) C is a Citation node; *and* (3) there exists a Supporting arc $\overrightarrow{s(H, C)}$ from H to C.

- *R05a: Hypo Opposes Cite*: hypothesis node with an opposing arc to a citation.

- *R05b: Hypo To Cite*: hypothesis node with any arc to a citation. This rule will cover both *R05* and *R05a*.

- *R06: Uncompared Cite & CurrStudy*: citation node that has not been compared to any current study node. As noted in Section 3.2 the comparison arc is designed to state analogies and distinctions between two components, primarily citations and current study nodes. During the course of the study the students were encouraged to draw analogies and comparisons for all citations and required to do so for all opposing citations. This rule tests for the existence of a pair of nodes, one cite one current study, that are not connected by a comparison arc.

- *R06a: Uncompared CurrStudy & Cite*: current study node that has not been compared to any citation. This is the logical inverse of *R06*, designed to detect isolated current study nodes.

- *R07: Uncompared Opposition*: two citation nodes that disagree about a shared hypothesis or claim node and are not connected via a comparison arc. As per the discussion above, this tests for the required case of two opposing citations that have no comparison arc between them.

  (Pseudocode) Count the set of all nodes C_i, C_j, & O such that: (1) C_i is a Citation node; *and* (2) C_j is a Citation node; *and* (3) O is a Hypothesis or Claim node; *and* (4) there exists a supporting *path* from $\overrightarrow{Supp(C_i, O)}$ from C_i to O; *and* (4) there exists an opposing path $\overrightarrow{Opp(C_j, O)}$ from C_j to O.

  Supporting Paths $\overrightarrow{Supp(N_0, N_j)}$ are directed paths $\overrightarrow{s(N_0, N_1)}, \ldots, \overrightarrow{s(N_{j-1}, N_j)}$ such that every arc $\overrightarrow{s(N_x, N_{x+1})}$ in Supp is a supporting arc from N_x to N_{x+1}.

An *Opposing Path* $\overrightarrow{Opp(N_0, N_j)}$ is a directed path $\overrightarrow{e(N_0, N_1)}, \ldots, \overrightarrow{e(N_{j-1}, N_j)}$ such that an *odd number* of arcs on the path $\overrightarrow{e(N_x, N_{x+1})}$ are Opposing arcs, and all the remaining arcs are Supporting arcs.

- *R07b: Undistinguished Opposition*: A comparison arc exists between disputant pairs but no distinction is specified to explain the disagreement. That is given a pair of opposing citations *with* a comparison arc does that arc have a distinction stated in the distinction field or not?

- *R07u: Undir Uncompared Opposition*: This is a generalization of R07 which permits undirected paths. In R07 the supporting and opposing paths must be directed from the citation nodes to the shared claim or hypothesis. Here the direction of the interim arcs on each path is immaterial, only their type. This was added to test for cases where students were confused about the direction of the arcs but not their role. That is, the rule no longer requires that the arcs in the supporting and opposing paths proceed from N_x to N_{x+1} but permit reversed arcs along the path as well. This rule was added to account for cases where the students had drawn citations that disagree but had done so with some directional error in the paths. These were added based upon the observation of student directionality errors in the pre-study.

- *R07ub: Undef Undistinguished Opposition*: An undirected generalization of *R07b*.

Grounding Features deal with the extent to which claims are 'grounded' or 'founded' in a citation. *Unfounded claims* have no connection to a citation and thus no basis in the literature. Similarly *Ungrounded Hypotheses* are those that have no path to a citation and thus are not based upon the relevant literature.

- *R11: Ungrounded Hypo Claim*: A hypothesis node which is supported by a claim node that is itself unfounded. Thus it tests for claims that are used to support hypotheses but themselves have no indicated basis in the literature.

(Pseudocode) Count the set of all nodes N such that: (1) N is a Hypothesis or Claim node; *and* (2) there does not exist a node C such that: (2a) C is a Citation Node; *and* (2b) there exists a Supporting Path $\overrightarrow{S(C,N)}$ from C to N

- *R11a: Ungrounded Hypo*: A hypothesis that is not grounded in the literature via directed paths (of any type) from a citation. This includes neighboring claims nodes that are themselves unfounded.

- *R11b: Unfounded Claim*: A claim node that is not connected in some way to a citation node, via supporting, opposing, or undefined paths from the citation to the claim. This uses the path construct so the claims can be supported by citations via intermediary claims.

- *R11u: Undef Ungrounded Hypo Claim*: This is a generalization of R11 that allows for undirected paths, thus incorporating cases where the direction of the arcs is wrong.

- *R11ua: Undef Ungrounded Hypo*: This is a generalization of R11a for undirected paths.

- *R11ub: Undef Unfounded Claim (neg)*: This is a generalization of R11b for undirected paths.

Disjoint Subgraph *R13: Disjoint Subgraphs*: In order for an argument to be coherent it must be the case that each piece of it relates to the others. Absent such a connection there is little reason to believe that the arguments form a coherent whole. This rule counts the number of disjoint nodes in the graph. This rule can range as high as $\frac{|G_n|^2}{2}$.

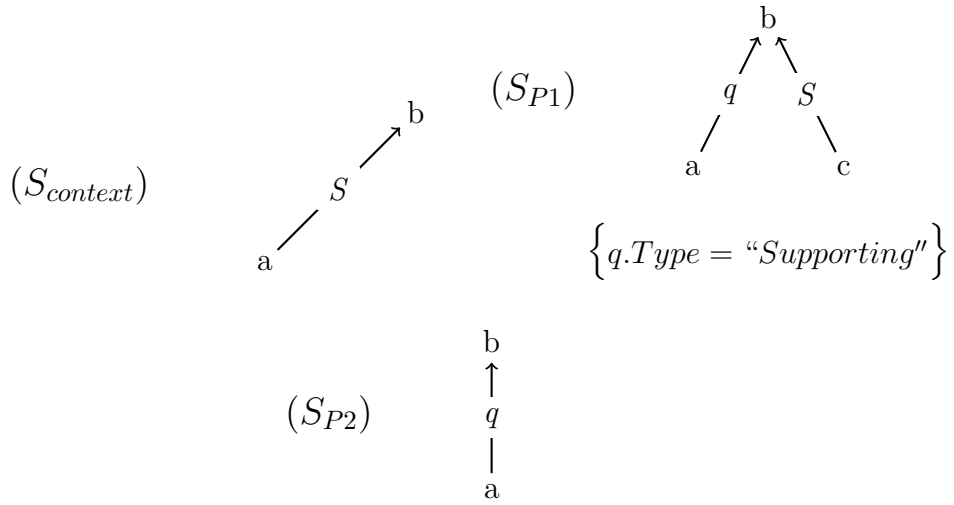
(Pseudocode) Count the set of all node pairs N & M such that: there does not exist an undirected path $P(N,M)$ between N and M.

5.5 RELIABILITY

In order for the empirical validation to be useful it is necessary to ensure that the manually-assigned grades are sufficiently reliable. In the present analysis I will be using the manually-assigned graph grades as independent variables along with the features to predict the essay grades. As before it is necessary to show that all of these features are sufficiently reliable. The graph features are calculated automatically from the individual graphs. As such they have 100% test-retest reliability. The reliability of the diagram and essay grades was previously discussed in Section 4.4 (pp. 50). For the analyses reported here I will continue to use the filtering discussed there. More specifically, all of the diagram grades will be used in the predictive models while I will focus on five of the essay grades: *E.01 (RQ-Quality)*, *E.04 (Hyp-Testable)*, *E.07 (Cite-Reasons)*, *E.10 (Hyp-Open)*, and *E.14 (Arg-Quality)*.

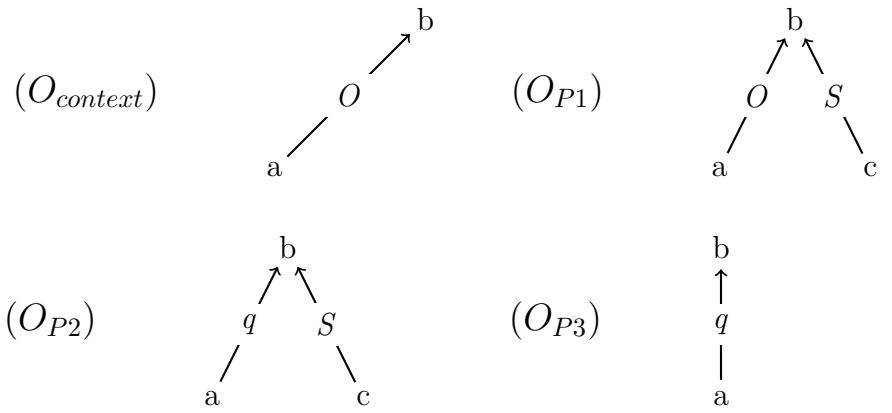
5.6 CONCLUSION

This chapter framed the null hypotheses H_{a1} and H_{a2} and described the augmented graph grammar framework that I will use to address them. The chapter further described the specific graph features of interest to this analysis and sets the thresholds used for reliability of the graph and essay grades. In the following two chapters I will build upon this work and analyze the two hypotheses by means of a direct comparison study and a greedy modeling process.



$\overrightarrow{S(i,c)}$ define a supporting path as a right-recursive production that maps variable length paths of parallel directed supporting arcs.

Figure 5.2: A simple recursive rule production that defines a *supporting path* in an argument diagram.



$\overrightarrow{S(i,c)}$ define an opposing path as a right-recursive production that maps to an opposing path preceded by a supporting path (O_{p1}), an opposing arc followed by a supporting path (O_{p2}) or a single opposing arc (O_{p3}).

Figure 5.3: A simple recursive rule production that defines an opposing path in the diagram.

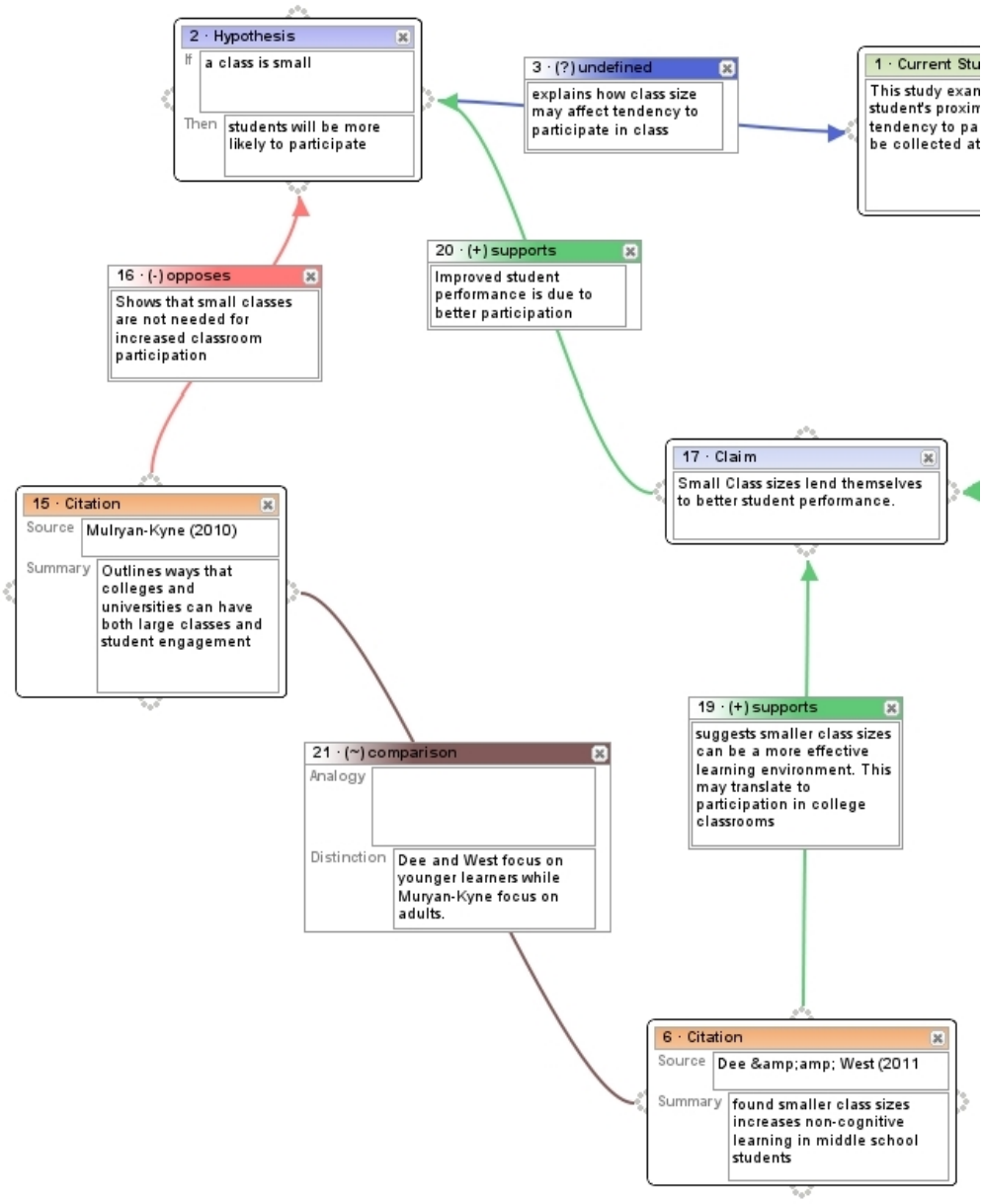
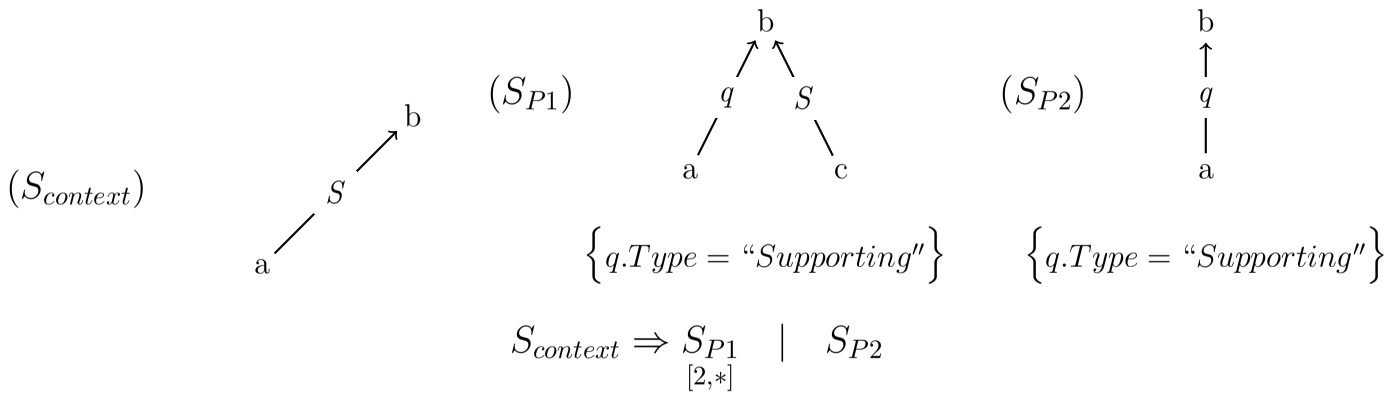


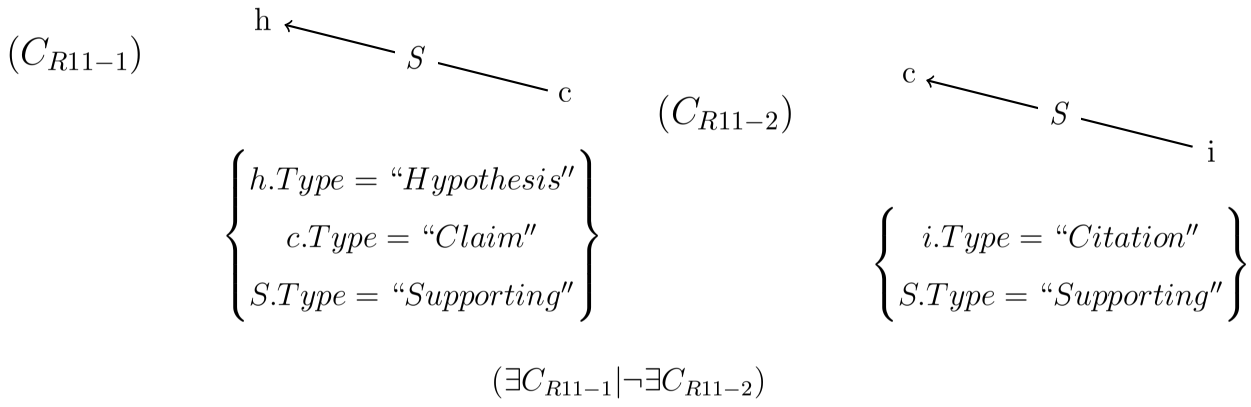
Figure 5.4: A sample argument subgraph that matches the recursive rule shown in Figure 5.1

.



$\overrightarrow{S(i,c)}$ define a supporting path as a right-recursive production that maps variable length paths of parallel directed supporting arcs.

Figure 5.5: Augmented Graph Grammar Example Supporting Path



Count the set of all nodes N such that: (1) N is a Hypothesis or Claim node; *and* (2) there does not exist a node C such that: (2a) C is a Citation Node; *and* (2b) there exists a Supporting Path $\overrightarrow{S(C, N)}$ from C to N

Figure 5.6: Augmented Graph Grammar rule example for the complex rule: *R11: Ungrounded Hypo-Claim.*

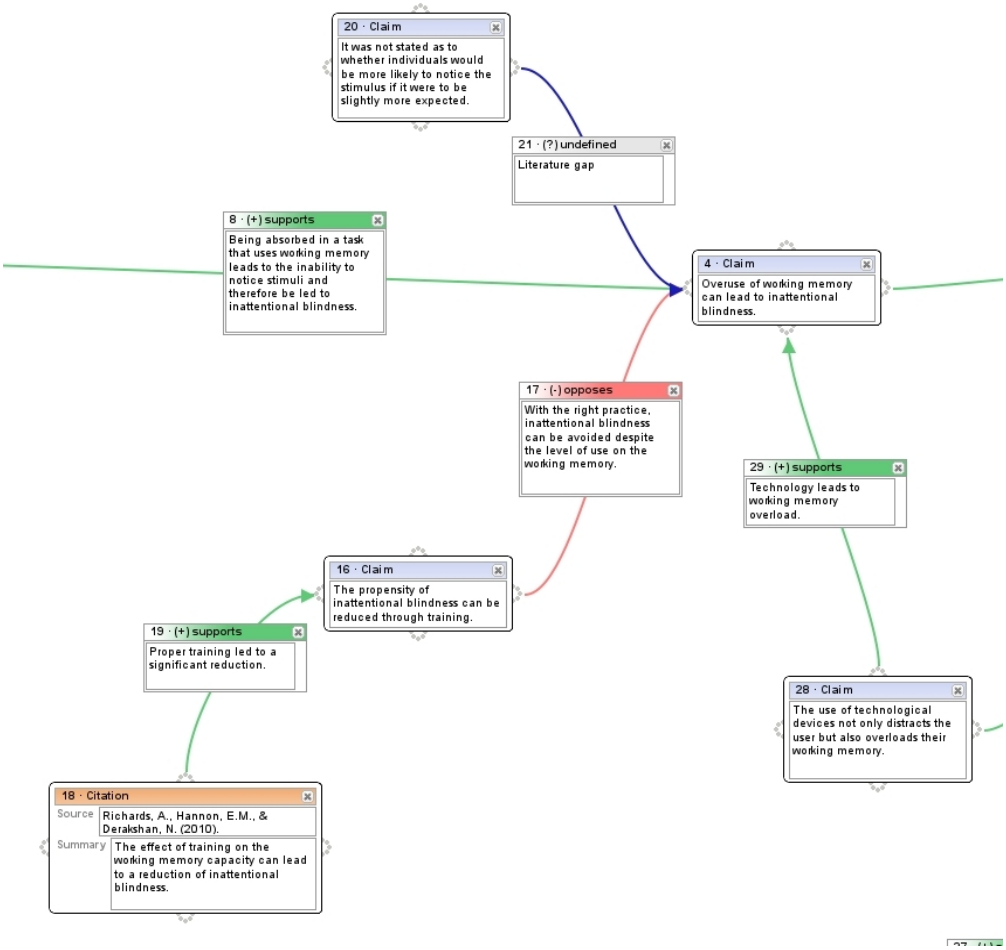


Figure 5.7: Reference argument diagram segment drawn from a student-produced argument used in the study.

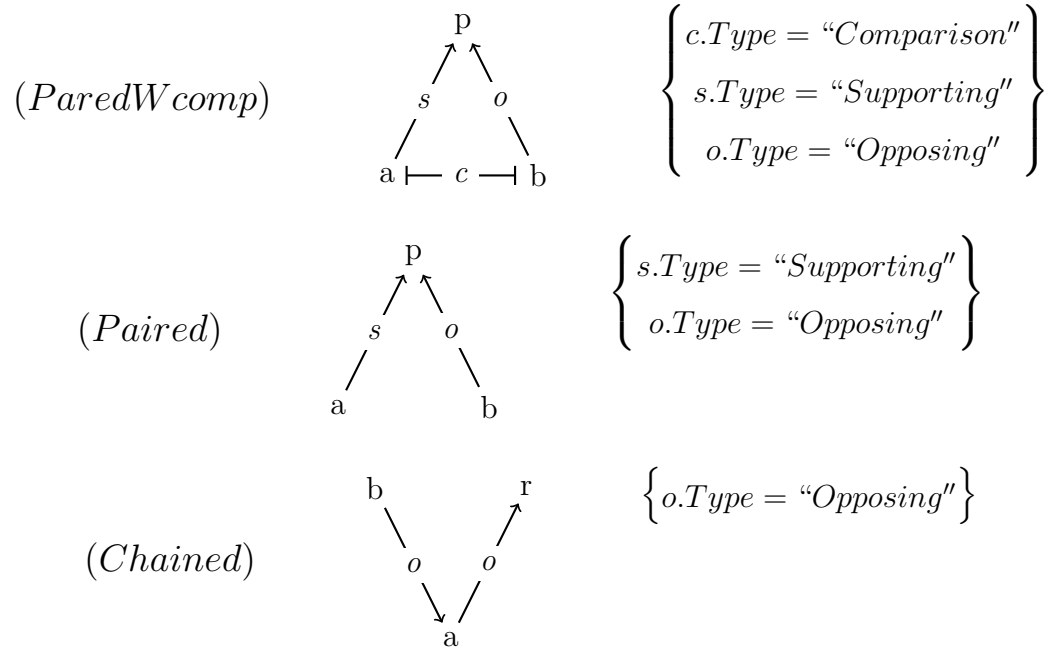


Figure 5.8: Augmented Graph Grammar examples for the Chained rules including *Paired Disagreement* both with and without a comparison arc, and *Chained Disagreement*.

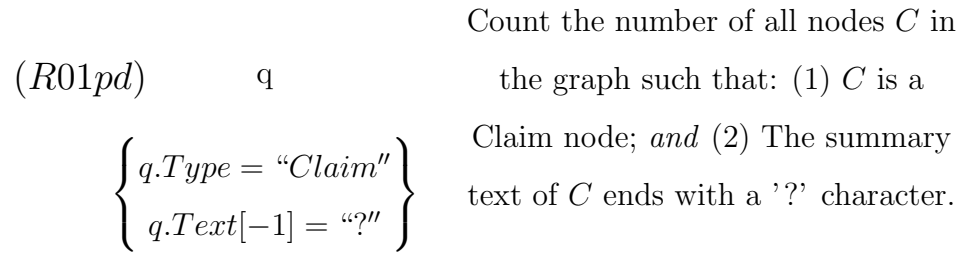


Figure 5.9: An example Augmented Graph Grammar rule for evaluation in this case: *R01pd Has RQ*.

6.0 H_{A1} EMPIRICAL VALIDITY

H_{a1}**:** It is not possible to define *empirically-valid* diagram rules that correlate with students' novel written argumentation ability.

6.1 INTRODUCTION

The focus of this chapter is on the empirical validation study conducted to test hypothesis H_{a1}. As noted in Section 5.2 previous studies have tested the relationship between graph features and subsequent comprehension measures as well as gestalt scores. The focus here is on the use of individual graph features to assess students' specific argumentation abilities.

The study conducted here is a data analysis study that draws on the simple and complex graph features described in Section 5.4 and the reliable essay grades described in Section 5.5. The study will be described briefly in the next section along with a listing of the results. Analysis of the results by feature type will be presented in Sections 6.3 (pp. 87) & 6.4 (pp. 94). I will present my overall conclusions in Section 6.5.

6.2 RESULTS

For the individual empirical validation I performed a series of pairwise comparisons linking each of the independent features to the essay grades. Prior to performing these comparisons I normalized each of the essay grades to a range of $(0-1)$ based upon the observed values. I then calculated the total scores for each of the features per diagram and generated three

distributions: raw scores; log of the raw scores; and a binary distribution where the score was 0 if the raw count was 0 or 1 otherwise. For each distribution I then calculated a direct comparison with the target essay variables using Spearman's ρ [29, 137].

I report the statistically- and marginally-significant predictive results for the simple features in Table 6.1 (pp. 88) and Table 6.2 (pp. 89). None of the complex textual or visual features were individually significant. The comparison results for the remaining complex features can be seen in Table 6.3 (pp. 90), Table 6.4 (pp. 91), and Table 6.5 (pp. 92). I omitted results for the positive individual counts *R01p: ** as those mirror the *Elt ** simple features. For each feature I report only the strongest statistically-significant or marginally-significant relationship if any with the highest $|\rho|$.

6.3 ANALYSIS: SIMPLE FEATURES

As Tables 6.1 - 6.5 illustrate a number of the features were predictive. The strongest rank correlation was $\rho = 0.383$ between the existence of a hypothesis node (*Elt hypothesis (binary)*) and the manually assigned grade for the testable hypothesis (*E.04 (Hyp-Testable)*). This was also the only simple feature to be significantly correlated with the overall grade *E.14 (Arg-Quality)* at $\rho = 0.228$. This result is positive and confirms that the graph measures are capturing the central role that hypotheses play in this type of argument. The correlation between the amount of opposition (*Elt opposes (log)*) and *E.10 (Hyp-Open)*, which grades the extent to which the author has defended the openness of his/her hypothesis ($\rho = 0.264$) is similarly positive, as is the correlation between the number of citations (*Elt citation (log)*) and *E.10 (Hyp-Open)* ($\rho = 0.213$). In both cases this is consistent with the instructions given to the students and with the expectations of the course instructors.

Several of the other correlations or absent correlations, were more surprising. I am surprised by the absence of any significant correlation between the visual features and the grades. It is not understatement to say that "messiness" was frequently cited by students, TAs, and the graders as a problem that adversely impacted their work and peace of mind. Several advocated for the addition of normalization features to the system that would restrict

Table 6.1: Empirical Validation: Size and Density Simple Features Includes item names, distribution, grade and score values. $p \leq 0.05$ are boldfaced $p \leq 0.1$ is italics.

Names			Feature Values		Grade Values		Spearman	
Feature	Dist	Grade	Mean	σ^2	Mean	σ^2	ρ	p-value
Order	log	E.07 (Cite-Reasons)	0.44	0.17	0.69	0.27	0.162	*0.098*
Order	log	E.10 (Hyp-Open)	0.44	0.17	0.3	0.36	0.172	*0.079*
Size	log	E.07 (Cite-Reasons)	0.6	0.14	0.69	0.27	0.193	**0.049**
MinDegree	raw	E.04 (Hyp-Testable)	0.74	0.44	0.77	0.24	-0.213	**0.029**
MaxDegree	raw	E.04 (Hyp-Testable)	0.47	0.18	0.77	0.24	-0.194	**0.047**
MaxDegree	raw	E.10 (Hyp-Open)	0.47	0.18	0.3	0.36	0.178	*0.07*
MaxChildren	log	E.04 (Hyp-Testable)	0.59	0.18	0.77	0.24	-0.187	*0.056*
MaxChildren	log	E.10 (Hyp-Open)	0.59	0.18	0.3	0.36	0.294	**0.002**
AvgChildren IgnoreEmpty	log	E.10 (Hyp-Open)	0.54	0.15	0.3	0.36	0.226	**0.02**
MaxChildren IgnoreEmpty	log	E.04 (Hyp-Testable)	0.59	0.18	0.77	0.24	-0.187	*0.056*
MaxChildren IgnoreEmpty	log	E.10 (Hyp-Open)	0.59	0.18	0.3	0.36	0.294	**0.002**
AvgParents IgnoreEmpty	log	E.10 (Hyp-Open)	0.47	0.11	0.3	0.36	-0.179	*0.067*

Table 6.2: Empirical Validation: Ontology Simple Features Includes item names, distribution, grade and score values. $p \leq 0.05$ are boldfaced $p \leq 0.1$ is italics.

Names			Feature Values		Grade Values		Spearman	
Feature	Dist	Grade	Mean	σ^2	Mean	σ^2	ρ	p-value
Elt citation	log	E.10 (Hyp-Open)	0.62	0.15	0.3	0.36	0.213	**0.029**
Elt comparison	log	E.01 (RQ-Quality)	0.11	0.22	0.59	0.31	0.234	**0.016**
Elt comparison	log	E.07 (Cite-Reasons)	0.11	0.22	0.69	0.27	0.177	*0.071*
Elt hypothesis	bin	E.04 (Hyp-Testable)	0.91	0.28	0.77	0.24	0.383	**0.001**
Elt hypothesis	log	E.07 (Cite-Reasons)	0.36	0.16	0.69	0.27	0.249	**0.01**
Elt hypothesis	bin	E.14 (Arg-Quality)	0.91	0.28	0.74	0.22	0.228	**0.019**
Elt opposes	log	E.10 (Hyp-Open)	0.25	0.29	0.3	0.36	0.264	**0.006**
Elt supports	log	E.07 (Cite-Reasons)	0.55	0.14	0.69	0.27	0.18	*0.067*
Elt unspecified	bin	E.01 (RQ-Quality)	0.5	0.5	0.59	0.31	-0.184	*0.06*

Table 6.3: Empirical Validation: Chained & Negated Individual Ontology Complex Features Includes item names, distribution, grade and score values. $p \leq 0.05$ are boldfaced $p \leq 0.1$ is italics.

Names			Feature Values		Grade Values		Spearman	
Feature	Dist	Grade	Mean	σ^2	Mean	σ^2	ρ	p-value
PairedCounterarg	raw	E.10 (Hyp-Open)	0.13	0.23	0.3	0.36	0.323	**0.001**
R01na: NoHypothesis	bin	E.04 (Hyp-Testable)	0.09	0.28	0.77	0.24	-0.383	**0.001**
R01na: NoHypothesis	bin	E.07 (Cite-Reasons)	0.09	0.28	0.69	0.27	-0.166	*0.09*
R01na: NoHypothesis	bin	E.14 (Arg-Quality)	0.09	0.28	0.74	0.22	-0.228	**0.019**
R02a: NonHypowoOut	log	E.07 (Cite-Reasons)	0.42	0.26	0.69	0.27	-0.207	**0.034**
R02c: NonHypoRQwoOut	log	E.07 (Cite-Reasons)	0.42	0.26	0.69	0.27	-0.204	**0.037**
R08: Unopp Hypo	log	E.04 (Hyp-Testable)	0.4	0.22	0.77	0.24	0.196	**0.045**
R10a: Hypo or Claim Comp	bin	E.01 (RQ-Quality)	0.16	0.37	0.59	0.31	0.17	*0.083*
R12: UndefinedCiteClaim	bin	E.01 (RQ-Quality)	0.23	0.42	0.59	0.31	0.225	**0.021**
R12: UndefinedCiteClaim	bin	E.07 (Cite-Reasons)	0.23	0.42	0.69	0.27	0.172	*0.079*

Table 6.4: Empirical Validation: Textual and Triplet Ontology Complex Features Includes item names, distribution, grade and score values. $p \leq 0.05$ are boldfaced $p \leq 0.1$ is italics.

Names			Feature Values		Grade Values		Spearman	
Feature	Dist	Grade	Mean	σ^2	Mean	σ^2	ρ	p-value
R04a: EmptyNodeFields	log	E.07 (Cite-Reasons)	0.42	0.18	0.69	0.27	0.173	*0.078*
R05a: HypoOpposesCite	bin	E.04 (Hyp-Testable)	0.03	0.17	0.77	0.24	0.173	*0.077*
R06: Cite Uncompared w Curr	log	E.10 (Hyp-Open)	0.61	0.15	0.3	0.36	0.211	**0.031**
R07: UncomparedOpp	log	E.10 (Hyp-Open)	0.14	0.26	0.3	0.36	0.396	**0.001**
R07b: UndistinguishedOpp	log	E.10 (Hyp-Open)	0.14	0.26	0.3	0.36	0.396	**0.001**
R07u: Undef UncomparedOpp	log	E.10 (Hyp-Open)	0.17	0.27	0.3	0.36	0.321	**0.001**
R07ub: Undef UndistinguishedOpp	log	E.10 (Hyp-Open)	0.17	0.27	0.3	0.36	0.321	**0.001**

Table 6.5: Empirical Validation: Grounding & Disjoint Complex Features Includes item names, distribution, grade and score values. $p \leq 0.05$ are boldfaced $p \leq 0.1$ is italics.

Names			Feature Values		Grade Values		Spearman	
Feature	Dist	Grade	Mean	σ^2	Mean	σ^2	ρ	p-value
R11: Ungrounded Hypo Claim	bin	E.07 (Cite-Reasons)	0.3	0.46	0.69	0.27	0.196	**0.045**
R11: Ungrounded Hypo Claim	bin	E.14 (Arg-Quality)	0.3	0.46	0.74	0.22	0.213	**0.029**
R11u: Undef Ungrounded Hypo Claim	bin	E.04 (Hyp-Testable)	0.4	0.49	0.77	0.24	0.171	*0.081*
R11u: Undef Ungrounded Hypo Claim	bin	E.07 (Cite-Reasons)	0.4	0.49	0.69	0.27	0.219	**0.025**
R11ua: Undef Ungrounded Hypo	log	E.07 (Cite-Reasons)	0.12	0.27	0.69	0.27	-0.226	**0.02**
R11ua: Undef Ungrounded Hypo	log	E.14 (Arg-Quality)	0.12	0.27	0.74	0.22	-0.219	**0.025**
R11ub: Undef Unfounded Claim	log	E.04 (Hyp-Testable)	0.14	0.24	0.77	0.24	0.225	**0.021**
R13: DisjointSubgraphs	bin	E.04 (Hyp-Testable)	0.47	0.5	0.77	0.24	0.183	*0.062*

diagram design or provide automatic cleanup. Still others questioned the utility of the project given the difficulty of producing and analyzing argument diagrams. Moreover there is some support in the literature for the belief that visual or structural features of the diagramming tool will affect performance [117]. Therefore either the problem was overstated or the simple structural methods did not adequately capture the concept.

It is also notable that no statistically-significant correlation was found between the amount of text contained in the node and arc fields and the grades. In the study reported in [68] my colleagues and I found that the amount of text included in the diagram fields, particularly the citation nodes, was both a significant predictor of the overall grade and could be used to distinguish student diagrams from those of an expert. I therefore expected it to be meaningful here. The difference however, may be explained by the assignment context. The study reported in [68] took place in a law school course where students were tasked with researching an area of law and drafting an argument for a real case. Law is a scholarly pursuit with a strong emphasis on citing existing works and students are encouraged not only to cite all relevant law but to quote from it as required. The quality and comprehensiveness of the cited works was a factor in the students' final grades. Legal citation is described as "a fine art"[1] and is central to legal education. This lends itself to larger more complex citation nodes. In Research methods, by contrast, students are encouraged to summarize the relevant works briefly and were given a fixed number of required citations, which encouraged them to focus on other things.

The remaining simple feature correlations are less specific. For example, Graph *Size* is significantly correlated with *E.07 (Cite-Reasons)*. This grade reflects whether or not the students explain why the cited works are relevant to their argument and the quality of that explanation. While this is a clear correlation it does not lend itself to precise advice. We can, for example, encourage students to flesh out anemic diagrams but it does not necessarily allow us to tell them what specific part of the diagram should be expanded. Therefore it may be difficult for novice students to act on the advice given in a useful way.

[1] Professor Kevin D. Ashley, University of Pittsburgh, private communication

6.4 ANALYSIS: COMPLEX FEATURES

The number of paired counterarguments (*PairedCounterarg*) was positively correlated with the openness of the hypothesis (*E.10 (Hyp-Open)*) at $\rho = 0.323$. This was consistent with the directions given to the students who were told to represent disagreements in this way. Similarly the absence of a hypothesis (*R01na: NoHypothesis (bin)*) was negatively correlated with: the testability of the hypothesis (*E.04 (Hyp-Testable)*, $\rho = -.383$), the presence and quality of reasons for the cited works (*E.07 (Cite-Reasons)*, $\rho = -0.166$), and the overall quality of the argument (*E.14 (Arg-Quality)*, -0.228).

Negative correlations were found between the presence of a non-hypothesis node without an outgoing arc (*R02a: NonHypoWoOut (log)*) and *E.07 (Cite-Reasons)*; and between the presence of a non-hypothesis or research question node without an outgoing arc (*R02c: NonHypoRQwoOut (log)* and *E.07 (Cite-Reasons)*. They were also found between the number of ungrounded hypothesis nodes using undirected paths (*R11ua: Undef Ungrounded Hypo (log)* and both the presence and quality of reasons for the citations (*E.07 (Cite-Reasons)*) and argument quality (*E.14 (Arg-Quality)*).

The remaining rules are, if anything, more surprising as they include wholly counterintuitive results such as the positive correlations between: the number of empty text fields in the nodes (*R04a: EmptyNodeFields (log) & E.07 (Cite-Reasons)* ($\rho = 0.173$); the presence of disjoint subgraphs (*R13: Disjoint Subgraphs (bin)*) & *E.04 (Hyp-Testable)* ($\rho = 0.183$); and the amount of uncompared opposing citations (*R07* (log)*) and *E.10 (Hyp-Open)*.

These results indicate the limits of direct nonparametric measures of validity. In the case of *R02a: NonHypoWoOut* for example it seems likely that the correlations with *E.07 (Cite-Reasons)* are an artifact of the data. Students who performed better generally were also more apt to state reasons than students who did not, hence the correlation. In the case of *R07** by contrast, the results run counter to our expectations. In this case, however, I think that it reflects persistent student limitations. Students were instructed to use opposing arcs to explain their arguments and the results indicate that some of the higher-scoring students did so. As shown in Table 6.2, *Elt opposes* was positively correlated with *E.10 (Hyp-Open)* as was *PairedCounterarg*. Both such features are subsumed by the *R07** rules and it appears

that few if any of the students included the comparison arcs necessary to distinguish the *PairedCounterarg* case from the *R07** cases. Thus the uncompared rules are, in effect, matching the same subgraphs as the *PairedCounterarg* rule and covering the same cases.

In future analyses of this type, it may be necessary to condition the more complex rules on the smaller ones. That is, one could look at the empirical validity of the *R07** rules on only those cases where opposing arcs have already been included. Collecting sufficient data for such a case, however, may be prohibitively expensive.

6.5 CONCLUSIONS

The goal of this chapter was to frame the discussion of Question Q_a and to address Hypothesis H_{a1} in detail. Necessary background for the general hypothesis was presented in Sections 5.1 (pp. 61) - 5.5 (pp. 78). The individual empirical validation was discussed in Section 6.1. As discussed above several of the graph features were correlated with one or more of the essay grades. These results contradict null hypothesis H_{a1} and provide a basis to develop empirically-validated guidance for future students.

Those correlations, however, were not always consistent with the *a-priori* assumptions that motivated their construction. Likewise some of the anticipated features (e.g. the messiness and text criteria) were *not* significantly correlated with the final grades despite user reports and prior research. Some of these results are due to assignment-specific differences that warrant a comparison with novel datasets. The others are driven both by the complexity of the rules and some limitations of the dataset which is, after all, not an exploratory coverage of the design space but a representation of real students who are receiving specific guidance. Therefore H_{a1} does *not* hold.

In this case, as in the Validity assessment reported in Chapter 4 the correlations come with some caveats. First and foremost, the results are nonparametric correlations and are frequently of relatively low strength (e.g. $\rho = 0.171$). Therefore this empirical validity should not be read as predictive validity, only as a conclusion that the rules themselves do correlate with the behaviors whether desired or undesired. Secondly, as noted above, some of the

correlations violated our *a-priori* pedagogical assumptions. It is possible that these cases indicate artifacts of the data or advanced rules that require lower-level performance which results in the improvement. This requires further analysis. And finally the performance is based upon hand-tooled rules and states nothing about whether better rules could be chosen or whether these correlations generalize to other assignments or domains.

7.0 H_{A2} MODEL PREDICTION

H_{a2}**:** Automatic features of student diagrams *cannot* be used to predict students' novel written argumentation ability.

7.1 INTRODUCTION

While previous chapters have shown that some of the diagram features are correlated with students' essay grades, these individual correlations we observed were nonparametric relationships and, as such, could not be used for robust grade prediction. Such robust prediction, however, is important for use of diagrams in the real world. In an educational setting automated advice and automated grading are useful particularly for immediate homework-helper interventions. Absent a direct connection between the advice given and subsequent performance, however, it can be difficult to integrate such a system into the classroom setting. If the diagramming system can also be used to predict future performance and to rank students generally, then it is possible for instructors to deploy it both for immediate guidance and to flag students who require individual attention or additional practice.

This chapter will focus on hypothesis H_{a2}. My goal will be to develop a reliable model that predicts students' essay grades based upon their diagram features. The analysis conducted here will use the same feature and score data described in Chapter 5 with the same normalization and filtering. The next section will describe the basic model induction process and present some relevant background information on linear models. Section 7.3 will then describe the specific study methods. This will be followed by a statement of the results (Section 7.4) to be followed by analysis and conclusions as well as implications for future

research (Section 7.5).

7.2 LINEAR REGRESSION & MODEL INDUCTION

The present goal is to develop a predictive model that estimates students' essay performance based upon features of their planning diagrams. In [70] we opted to discretize the resulting grade splitting students by median grade and then applied a decision-tree classification. While that approach can readily split up the students, it is relatively insensitive and is unsuitable for the current dataset. As noted in Subsection 4.3.3 (pp. 49) the raw essay grades are negatively skewed. While this poses no problem for the nonparametric statistics used previously, it would complicate any classifier approach making the classes unbalanced and inconsistent.

Therefore the present work will treat this as a regression problem. The graph features will be treated as numeric counts and will be combined using standard linear regression models (see Subsection 7.2.1) and later Generalized additive models (see Subsection 7.4.4). As will be discussed below, both models are robust well-known methods that can be readily inspected by domain experts for subsequent analysis. They also have well-known training methods that could in theory be used to train them directly using the full space of graph features. Such models, however, would be prone to over-fitting and would be less informative than more parsimonious approaches.

Therefore for the purposes of this thesis, I will take a two-pass approach to prediction based upon linear regression models. Candidate models will be generated using a greedy generation approach and then trained using standard methods for later evaluation. The remainder of this section will summarize the theory and training of standard linear models. It will then describe comparative evaluation metrics for models (Subsection 7.2.2) and conclude with a discussion of the greedy generation process used to define the model structure (Subsection 7.2.3).

7.2.1 Standard Linear Regression Models

Linear regression models are theoretically-grounded robust models, widely used in psychology, the social sciences, and other empirical domains. They are additive models that represent the relationship between one or more *independent variables* $x_0, \ldots x_n$ (sometimes called the input or explanatory variables) and one or more *dependent variables* (sometimes called the output or conditional variable) y_j [29, 34, 133]. When a model contains only one independent variable it is referred to as *Simple Regression*, while models with more than one independent variable are called *Multiple Regression* models. In the present work multiple regression models with a single dependent variable will be used. They can be defined formally as:

$$y_i = \alpha + \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \epsilon_i \tag{7.1}$$

where α defines the *Intercept Point* or base value of the model; each β_k is a *coefficient* that defines the strength and the sign of the relationship between the independent variable x_k and the dependent variable y; and the *error term* ϵ_i defines the error term specific to the data term y_i.

Linear models are advantageous as they provide a clear and comprehensible model for the empirical relationship between the variables with the magnitude and sign of each β_k serving to indicate the magnitude and polarity of the relationship. Given suitable data the unknown values α, β_k, and ϵ_i, can be efficiently estimated using the *Least-Squares Regression* (also known as the *Method of Least-Squares* or *Ordinary Least Squares*). Least-squares regression operates by solving for the values above such that the sum of squared residual errors is minimized.

Given a predetermined set of variables, Least-Squares estimation is a robust and potentially optimal method for linear models. The appropriateness and efficiency of the algorithm, however, rests on seven assumptions shown below [34, 133]:

Linearity: The independent variables are linearly related to the dependent variable with constant individual effects represented by the β_k values. For generalized additive models (see [141, 54]) this assumption is relaxed to an assumption of *Additivity* (see: Section G.1 (pp. 247)).

Independence: The samples were taken independently and thus any error terms ϵ_i and $\epsilon_{j \neq i}$ are independent of one another as are the errors across the dependent variables y_j and y_k (see: Section G.2 (pp. 248))

Variability: The individual independent variables are non-constant (see: Section G.2 (pp. 248)).

Non-Multicollinearity: The independent variables are independent and not collinear. That is, no variable x_i is dependent on other variables $x_0, \ldots, x_{m \neq i}, \ldots, x_n$ (see (see: Section G.3 (pp. 248)).

Homoscedasticity: The variance of the error terms ϵ_i is constant for all observations. That is, the error variance is not affected by the independent variables (see Section G.4).

Normally-Distributed Errors: (aka *Normality*) The error terms of the model are normally-distributed: $\epsilon \sim N(0, \sigma_\epsilon^2)$ (see Section G.5).

Weak Exogeneity: The independent variables are error-free either because they are set by experimental condition (dividing students by age) or because they can be measured without error. As such they do not introduce a significant source of error into the model (see: Section G.6).

If the assumptions of *Linearity*, *Independence*, and *Homoscedasticity* are met then according to the *Gauss-Markov* theorem the least-squares estimator will be the most efficient unbiased linear estimator available [34]. If the additional assumption of *Normally-Distributed Errors* is met then the theorem states that least squares regression will be the most efficient estimator over *all unbiased estimators* even nonlinear ones. I describe these assumptions in more detail in Appendix G (pp. 246) and I will address them in the analysis and results sections below.

It is important to note that this list does *not* include a requirement that the data itself be normally distributed. While it is commonly assumed that data must be normally-distributed to use linear models this is not, in fact, a requirement. In order to conduct tests of statistical significance it is necessary to make general distributional assumptions about the data. When calculating f-scores, p-values or confidence intervals it is necessary to assume that the independent variables are drawn from a normal distribution or another fixed distribution if *Generalized Linear Models* are used (see [34, 141]). When using empirical model evaluation such as RMSE, as I do here, such assumptions are not required. For more discussion (see Section G.7 (pp. 251)).

7.2.2 Model Evaluation

There exist several alternative methods to rank and evaluate linear models. Typical evaluation metrics focus on the f-score and p-values which evaluate the statistical significance of the fit. As noted above that is unsuitable in this case because the essay grades are not normally-distributed. Alternative empirical methods include the Adjusted-R^2 ($\bar{R}^2$) score [130, 49], Akaike Information Criterion (AIC) [48], Bayesian Information Criterion (BIC) [48], and the *Root Mean Squared Error (RMSE)* [135, 48]. $\bar{R}^2$ is routinely used in empirical research and is robust but is not always optimal for model comparison. AIC and BIC, by contrast, explicitly penalize larger models and permit us to compare disparate models easily, the reported values, however, are not reflective of the actual grades. Moreover, both depend upon log-likelihood calculations which make assumptions about the underlying error distribution [48]. As I noted above in the discussion of *homoscedasticity* this is not always a safe assumption to make. Therefore I will make use of RMSE here. RMSE strictly relies on the observed errors and reports expected error scores in the range of the original variables. Penalization of larger models is handled by reality and the cross-validation process itself. Therefore this measure is robust in the face of unreliable data.

RMSE is a global estimate of model stability or reliability (see [135]) and is defined as:

$$RMSE = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n}} \tag{7.2}$$

where SSE is the sum of squared error of the model taken over the population. Thus RMSE is an empirical measurement of the distance between the observed value y_i and the model prediction $\hat{y}_i$. RMSE has the distinct advantage of summarizing the model error in the same units and range as the output variable. Therefore it can be used for both absolute model evaluation and relative comparisons. Unlike other measures, however, RMSE does not penalize models for increased complexity and thus is not robust against over-fitting. For this reason RMSE is often calculated in *Cross-Validation (CV)* studies of the type described here.

Cross-validation is an empirical method used to test the generality of a given model when no separate testing or validation data is available. In a cross-validation study the training

data is partitioned into subsets and the model is iteratively trained on all but one of the bins with the remainder being used as a test set [48]. In studies of this type the RMSE scores are collected over the test iterations and thus represent the performance of the model on the unseen data. Therefore these estimates are robust against over-fitting for the α and β_k values.

For the present work two types of cross-validation will be used. During the generation process discussed in Subsection 7.2.3 *10-fold* cross-validation with *balanced random assignment* will be used. Balanced random assignment partitions the dataset by rank, in this case the essay scores. The goal of this process is to ensure that each partition contains the same distribution of scores thus avoiding biased or unrepresentative samples. Balanced random assignment is frequently performed in study design where the goal is to generate equivalent participant groups.

Once the final models are generated they will be compared using *Leave-One-Out* cross-validation. In this form the model is tested once for each record in the dataset. On each iteration it is trained on all of the remaining records. Leave-one-out cross-validation is often more accurate than 10-fold as each training iteration takes advantage of almost all of the available data. It also guarantees, in this case, that models which are trained independently have covered the same training data. This additional training, however, is costly. With the present dataset that would increase the time per iteration by an order of magnitude. Therefore the leave-one-out cross-validation will only be used to evaluate the final models for comparison in Sections 7.4 & 7.5.

One disadvantage of the raw RMSE score is that it penalizes all errors equally. While that is desirable in most contexts, it is not necessarily appropriate here. In an educational context the cost of ignoring a student who needs help can be high relative to the cost of burdening an already successful student with excess advice. Thus when training a predictor model there may be a practical advantage to biased error detection. With that in mind we defined the *Conservative Mean Squared Error (CMSE)* as follows:

$$CMSE = \sqrt{\frac{\sum e^{(\hat{y}_i - y_i)}(\hat{y}_i - y_i)^2}{n}} \tag{7.3}$$

This score will penalize over-estimations of students' performance more than any underestimations, thus catching the lost at the risk of annoying the found. In the analyses below, both RMSE and CMSE will be reported for the resulting models.

7.2.3 Model Generation

While Least-squares regression is a robust mechanism for model fitting it does not support model construction. In psychology and other research domains where linear models are used, the sets of independent and dependent variables are typically specified *a-priori* in order to test an existing hypothesis. In the present context, however, the goal is to search for useful predictive models, if any, within an existing space of variables. Search within the space of variables is handled by the *greedyLM* algorithm shown in Algorithm 7.1 (pp. 118).

This is a simple *greedy-search algorithm* (see [25, 131]) over the space of independent variables that assumes a fixed dependent variable. It begins by considering all single-variable models ($\forall x_j : y_i = \beta_j x_j + \alpha$) and greedily expands upon the best one by adding additional independent variables until no improvement is made or the space of variables is exhausted. The algorithm is restricted to each iterative choice and does not reconsider any decisions.

As shown in the pseudocode, the models are ranked using RMSE under cross-validation. The algorithm also uses a p-value test to filter unacceptable models. While the estimated p-values cannot be trusted for hypothesis testing they will also be used as a heuristic to narrow the search space. On the first round, models are rejected unless the reported p-value is less than or equal to 0.5. It then drops to 0.1 for the second round and 0.05 for all rounds thereafter. This decay parameter was incorporated into the model after initial testing showed that early round models could not meet the higher standard. The *greedyLM* algorithm is guaranteed to complete with a worst case running time of $O(|Predictors|)$.

This algorithm is similar to the *Forward Stepwise Selection* approach described in [48]. That too is a greedy constructive algorithm. However, in lieu of the p-threshold and RMSE scoring it relies entirely on comparing the individual F-scores. That approach, while useful, relies on the assumption of normally-distributed data. I have also tested variants of the *greedyLM* algorithm that use $\bar{R}^2$, AIC, and BIC scores. For the present I will focus solely

on the RMSE version.

As constructed the algorithm has three inductive biases (see [82]). First it retains the assumption of *linearity* discussed above. Second it selects solely for *constructive* models. That is, it assumes that the best model with n independent variables can be made by extending the best model containing $n-1$ such variables. And finally it assumes that the p-values, while unreliable for hypothesis tests will not rule out any optimal models. That is there exist no optimal models of size $n > 2$ with p-values greater than 0.05.

7.3 METHODS

In the present analysis I will test the validity of hypothesis H_{a2} by inducing predictive linear models for the five reliable essay grades based upon the predictive graph features. The *greedyLM* algorithm was also applied to induce models from the manually-assigned graph grades for comparison purposes. The models will be contrasted with the baseline grades calculated by computing the most frequent score for the associated essay and for models induced from the graph grades. For each of the final induced models the RMSE and CMSE scores calculated under leave-one-out cross-validation will be reported, and the models will be compared based upon those results.

7.3.1 Graph Feature Sets

The actual model-induction process was based upon four sets of features: *Total*, *Intervention*, *Intuitive*, and *Intuitive-NoP*. The sets are described in Table 7.1 (pp. 119). I will use these names when referencing the induced models below. These separate feature sets were formed to test the predictive utility of relevant subsets of the full feature space. As noted in Chapter 5, the complex graph features were developed with *a-priori* assumptions about their impact on students' work. Most of the features were designed with *a-priori* intuitive notions about the relationships. Most were assumed to be negative and some of them (e.g. *R08: Unsupported Hypo*) were designed to support immediate feedback.

As noted in the previous chapter, the complex features were defined based upon an *a-priori* model of argumentation and assumptions about the relationship between the rules and subsequent performance. For some (e.g. the presence of a non-hypothesis node without an outgoing arc *R02a: Non-Hypo w/o Out*) the empirical results validated those assumptions. For others (e.g. the use of a comparison arc between a hypothesis and a claim node *R10a: Hypo Or Claim Comp*) they did not, showing instead that the features were positively correlated with performance.

In the present analysis four sets of features will be considered. The sets are summarized in Table 7.1 (pp. 119). Of these sets two are restricted to features where the empirical results validated the *a-priori* assumptions. One set includes all features for which some form of descriptive advice can be defined and the remainder covers all of the simple and complex features. When forming these sets the individual feature results were used to select the best distribution for later comparison (raw, bin, or log). For statistically-significant features the best match was chosen. For the others a qualitative evaluation was made. Thus each feature used in this analysis is unique.

7.3.2 Tolerance Reduction

While the raw feature sets were logically sound, they were also multicollinear. As noted above, multicollinearity can reduce the performance of a model by making it prone to over-fitting. With that in mind I filtered each of the datasets using a simple greedy approach based upon the approach described in [108] and the heuristic threshold defined in [1]. Pseudocode for this greedy approach is shown in Algorithm Algorithm 7.2 (pp. 120). The algorithm relies on the concept of variable *Tolerance* discussed in detail in Section G.3 (pp. 248) and has a guaranteed worst-case bound of $O(|IndependentVars|)$. This process was handled manually for the present work but can be easily implemented for later use. The number of variables removed from each set is shown in Table 7.1 (pp. 119).

As noted in Subsection 7.2.1 multicollinear data may contain many overlapping collinear subsets. Therefore while the greedy approach has been effective in the present study it is not guaranteed to be optimal.

7.4 RESULTS

In this section I report on the results of the induced models and the basic comparisons. When referring to the models I will name them by their source and type. The baseline grade predictions (see Subsection 7.4.1 will be labeled as *Baseline-** (e.g. *Baseline-E.01*) while the direct and manual grade models (see Subsection 7.4.2) will be labeled *Direct-** and *Manual-** respectively (e.g. *Manual-E.14*). Similarly, the induced linear models (see Subsection 7.4.3) will be designated by the feature set and type thus "*Intervention-Full-E.14*" denotes the induced linear model for *E.14 (Arg-Quality)* drawn from full *Intervention* dataset while "*Intuitive-Trimmed-E.01*" denotes the model for *E.01 (RQ-Quality)* drawn from the trimmed *Intuitive* dataset. The Generalized Additive Models (see Subsection 7.4.4) will be denoted similarly, e.g. "*Intuitive-NoP-Trimmed-GAM-E.01*".

7.4.1 Baseline

The RMSE and CMSE scores for the baselines are shown in Table 7.2 (pp. 121). These scores were not calculated under cross-validation as they are a constant prediction. As the table shows the results are relatively poor for the baseline scores with RMSE scores ranging from 0.23 almost 1/4 of the score range for *Baseline-E.04 (Hyp-Testable)* to 0.46 (nearly 1/2 of the range) for *Baseline-E.10 (Hyp-Open)*. With the surprising exception of *Baseline-E.10* the CMSE scores were far higher than the corresponding RMSE.

When examining the RMSE and CMSE scores, it is important to bear in mind that the scores have been normalized based upon the minimum and maximum *observed* values as shown in Table 7.3. Thus an RMSE score of 0.3 on questions *E.01 (RQ-Quality)*, *E.04 (Hyp-Testable)*, and *E.10 (Hyp-Open)* reflects an error rate of 1.2 points on the original range of -2 to 2.

7.4.2 Graph to Essay Grades

As discussed above, the direct correlation between the graph and essay grades in Section 4.5 (pp. 57). Table 4.6 (pp. 57) showed that 9 of the 14 graph grades were correlated with the

associated essay grade at a statistically significant level. That table, however was focused solely on nonparametric correlations. Table 7.4 (pp. 122) shows an updated parametric analysis with the mean and standard deviation for the paired graph and essay grades as well as RMSE and CMSE scores for a simple linear model of the form:

$$e_i = \alpha + \beta_i g_i + \epsilon_i \tag{7.4}$$

This model was evaluated using leave-one-out cross-validation. As the table illustrates the direct graph grades were better predictors than the most-frequent-essay-grade baseline for all but *Direct-E.04 (Hyp-Testable)*. For that model both the RMSE and CMSE scores were close. This result further strengthens the validity arguments made previously.

Similar performance results for the induced models based upon the diagram grades are shown in Table 7.5 (pp. 122). The individual features present in the models can be found in Table H1 (pp. 254). As the tables show the resulting models were mixed. For *Direct-E.01 (RQ-Quality)* and *Direct-E.04 (Hyp-Testable)* the single grade models outperformed more complex alternatives and other grades. While for *Direct-E.07 (Cite-Reasons)*, *Direct-E.10 (Hyp-Open)* and *Direct-E.14 (Arg-Quality)* the induced models outperformed the single-grade results indicating that the addition of other graph grades added useful information. In each case, however, the improvement was small suggesting that the added information was marginal at best.

7.4.3 Induced Feature Models

As described above predictive models were induced for the reliable essay grades based upon the separate feature sets. Two induction passes were used, one on the raw feature-sets before tolerance trimming, and the other on the trimmed sets. The overall RMSE and CMSE scores for the raw and trimmed results are shown in Table 7.6 (pp. 123). As shown here eight models were induced for each of the essay grades. Four of these models were drawn from the raw feature sets before trimming took place and four were drawn from the later trimmed sets. More detailed summaries of the raw and trimmed models can be found in Tables H2 (pp. 255) - H11 (pp. 264).

The induced models can be compared via their RMSE and CMSE scores. As expected, the larger feature sets improved the performance of the models with the *Total* models outperforming their smaller peers. Interestingly the tables also show that the effect of trimming the datasets was inconsistent across questions. On *E.04 (Hyp-Testable)* and *E.07 (Cite-Reasons)* the models drawn from the trimmed feature-sets generally performed worse than their raw counterparts both in terms of the RMSE and CMSE scores. On the remaining questions the trimmed models generally outperformed their raw peers or were competitive with them. Thus it is apparent that the trimming process removed some information of value to questions *E.04* and *E.07* while reducing noise for the remaining models. The implications of these results will be discussed in more detail below.

While these models are useful and sufficient to address the primary hypothesis of this chapter it has not yet been shown that they are optimal and meet all of the assumptions of the Gauss-Markov theorem. For the trimmed models it is possible to guarantee that all of the assumptions hold save for homoscedasticity and normality. If it can be shown that these hold, then we may reasonably interpret the performance of the trimmed models as a lower bound on the RMSE scores of the linear models containing these features. Absent those assumptions, however, it is possible that alternative unbiased estimators could yield better performance.

There are a variety of tests that could be applied to check the assumptions of homoscedasticity and normality. As these tests are model-specific they must be calculated after the model is trained. In the present context the first can be tested directly via the Breusch & Pagan (BPCW) test, also known as the Cook-Weisenberg test [34, 129]. This tests the null hypothesis that the model results are homeostatic. Therefore p-values above 0.05 give one no reason to reject the assumption. P-values at or below the threshold will cause one to conclude that it does not. Similarly, the assumption of normality can be tested via the Shapiro-Wilk (SW) test of normality. This too, tests the null assumption that the hypothesis holds (i.e. that the residuals are normally-distributed) and p-values below 0.05 cause one to reject that [136, 100].

While these tests are useful, they are no substitute for plots. A p-value of 0.02 will not, after all, demonstrate why the residuals are not normal. It will only tell us that they are not. With that in mind, statisticians often recommend testing the results visually by plotting the studentized residuals against the dependent variables and by plotting histograms of the observed errors. Plots of this type paired with results of the BPCW and SW tests are shown in Figures 7.1 (pp. 124) and 7.2 (pp. 125). These figures focus solely on the models for question *E.14 (Arg-Quality)* induced from the tolerance trimmed *Intervention* and *Total* feature sets: *Intervention-Trimmed-E.14* and *Total-Trimmed-E.14*. These two results were chosen as they were the best performing of the full set.

As the figures show, the models fail on both assumptions. The tests themselves return p-values well below the threshold and a visual glance of the plots highlights the problems. In both cases the studentized residuals are close to the horizontal where there are positive errors and far below it for negative errors. Similarly the error histograms both show negative skew. All of these problems can likely be attributed to the distribution of the raw score variable. The negative skew of the score variable is clearly mirrored by the error distribution and may explain the split on the studentized residuals between a large number of smaller positive residuals and a few extreme negative errors. Consequently the assumptions of homoscedasticity and normally-distributed errors do not hold and thus the criteria for the Gauss-Markov theorem is not met. Therefore while these models can be used for prediction, they are not guaranteed to be optimal.

7.4.4 Generalized Additive Models (GAMs)

As I showed in Subsection 7.5.1 (pp. 112), it is possible to induce predictive feature models that are competitive with manually-assigned graph grades. However these models rest on the strong assumption of *linearity*. This assumption has been treated as an inductive bias of the generation procedure but it has not been shown independently and visual inspection of the raw feature scores suggests that some may not have a linear relationship with the output variables. One option is to use *generalized additive models (GAM)* which relax this assumption in favor of a general assumption of *additivity*. GAMs are a semi-parametric

analogue to linear regression [34, 141, 54] of the form:

$$y_i = \alpha + \beta_0 x_0 + \ldots + \beta_j x_j + f_{j+1}(x_{(j+1)}) + \ldots + f_n(x_n) + \epsilon_i \tag{7.5}$$

This form is similar to the one shown in Equation 7.1 where as before: y_i is the dependent variable; $\{x_0, \ldots, x_n\}$ is the set of independent variables; α is the offset or *intercept*; and ϵ_i is the error term. In this model, however, the independent variables fall into two classes: $\{x_0, \ldots, x_j\}$ are the *parametric* terms and are connected with the dependent variable via a static coefficient as in general linear regression; $\{x_{(j+1)}, \ldots, x_n\}$ are the *nonparametric* terms and are connected via nonlinear *smoothing functions*. These are local regression or mapping functions f_n such as splines that are used to modify the impact of the independent variable over the range of x_n or y_i.

One of the simplest smoothing functions is to take a weighted average of local y_i values around a given datapoint for x_m and then to fit a static β_v value to it. Thus in lieu of a single static β_m coefficient one would have a set of static coefficients for each value of x_m. Then when applying the model for prediction we would choose the appropriate local coefficient based upon the independent variable's value. More complex smoothing functions are possible such as local polynomial regression and splines which I will employ here.

As with least-squares regression, when GAMs are fitted using *Generalized Cross-Validation (GCV)* (see [141]) they remain sensitive to the assumptions of independence, weak exogeneity, multicollinearity, and homoscedasticity. Distributional assumptions as discussed in Subsection 7.2.1 are only required if we wish to carry out statistical hypothesis tests or to calculate significance.[1] Unlike linear models, however, GAMs relax the assumption of linearity in favor of a more general assumption of *additivity*. Thus the functional influence of each independent variable $x_{*,m}$ is no longer a constant value, but it is still assumed that the relative influence of distinct variables $x_{*,p}$, and $x_{*,q}$ can be expressed by an additive relationship $f_p(x_{i,p}) + f_q(x_{i,q})$ independent of their specific values. GAMs thus have the advantage of being both robust in the face of locally nonlinear relationships and of being interpretable. GAMs can be extended to include more complex multi-argument terms of the form $f_{pq}(x_{i,p}, x_{i,q})$, however, I do not do so here.

[1] Professor Simon N. Wood, University of Bath, United Kingdom, personal communication.

While GAMs have a number of advantages over standard linear models they are substantially more complex. Fitting the models requires iterative estimation via algorithms such as GCV and backfitting (see [141, 48]) which are more computationally intensive than ordinary least-squares regression. As such they are impractical for use in the greedy model-construction procedure that I discuss below. Therefore, for the purposes of the present work, I will use the GAMs as a post-hoc extension to the induced models with the goal of testing whether relaxing the additivity assumption for a given linear model will improve the overall performance. I will make use of the *mgcv* package for GAMs authored by Simon Wood [143, 142] in the R statistical language [29]. This implementation fits GAMs using penalized regression splines via the GCV algorithm.

For the present analysis I tested the reliability of the GAM model and the impact of relaxing the linearity assumption by training analogous models for the questions. These models used the same feature sets as the best models drawn from the Intervention and Trimmed feature-sets. They were then trained using generalized cross-validation and evaluated for RMSE and CMSE. The scores are shown in Table 7.7 (pp. 126).

As the table shows, the additional complexity of the GAM model did not result in improved performance. While the models were not substantially worse than the linear models they were not an improvement either. Part of this problem may lie in the fact that the GAMs were based upon the induced linear models rather than being induced as GAMs. It is possible that a GAM-based induction process would select other, more advantageous features. It is more likely, however, that the additional sensitivity of the GAM model makes them prone to over-fitting and thus less robust in the face of cross-validation. Therefore the linear models are preferred.

7.5 ANALYSIS AND CONCLUSIONS

7.5.1 H_{a2} Primary Hypothesis

The primary goal of the present discussion was to address H_{a2} and determine whether or not automatic features can be used to predict subsequent argumentation scores. That hypothesis was tested via a machine-learning study in which predictive linear models were induced from the diagram features and compared with baseline grades and predictive models based upon the manually-assigned graph grades. Linear models were chosen as they provide a robust and well-supported formalism for regression problems and can be inspected by domain experts. Said models, however, rely on the restrictive inductive bias of linearity. Therefore alternative generalized additive models were also tested. These models were compared via their RMSE and CMSE scores as calculated under leave-one-out cross-validation.

As the prior tables have shown, the feature models were competitive with the models based upon student grades and both outperformed the baseline scores. In addition to collecting the RMSE and CMSE scores, the cross-validation analysis also collected the raw squared error scores $(\hat{y}_i - y_i)$. Therefore it is possible to perform a pairwise t-test to assess whether the differences are significant. Table 7.8 (pp. 127) shows the RMSE and CMSE scores for the relevant models.

As the table shows the best possible predictor was the linear models induced from the diagram features. This in turn appeared to outperform the comparable diagram grades and both beat the baseline. The induced models were also competitive with the best linear model, performing comparably to the grade-based model and beating both the more sensitive GAM and the direct grade (on RMSE, not CMSE). This analysis, while useful, is qualitative. Appropriate statistical comparisons of the models are shown in Table 7.9.

This table lists the results for two-sample Wilcoxon signed rank tests calculated over the three classes of models: *Baseline*, *Manual* Grade (either the direct grade or best model), and the *Total-Trimmed* models. The Wilcoxon test is a nonparametric test for two population means akin to a t-test [29, 138]. Unlike the standard t-test, however, it does not assume that the samples are normally distributed. The comparisons shown in Table 7.9 were based

upon the squared errors $(\hat{y}_i - y_i)^2$ collected during leave-one-out cross-validation. The non-parametric test is insensitive to the distribution of the errors and the absolute value of the scores. This test was chosen after evaluation of the error distributions showed that they were non-normal. The test was calculated as an unpaired comparison as the focus is not on absolute improvement on every score but improvement of the overall performance. Because multiple comparisons were run, the p-values of the models were corrected on a per-question basis via *Holm Correction* (see [29, 132]).

As the table shows, the differences between the models were statistically- or marginally-significant for questions *E.10 (Hyp-Open)* and *E.14 (Arg-Quality)*. On *E.10* both the *Manual* grade and the *Total-Trimmed* model differed from the baseline to a marginally significant degree. Moreover, they differed from each other to a statistically-significant degree as well. All of this is consistent with the observed RMSE scores which dropped significantly from 0.4627 for the manual grade to 0.3159 for the linear model. On *E.14* both the manual grade and the feature model differed from the baseline but did not significantly differ from one-another. It is important to note, however, that these are overall changes. As noted above the tests were calculated in an unpaired form due to the focus on overall improvement not improvement on each item. If paired tests are used and we ask about improvement on every item then the results are not significant, suggesting that the improvement is not uniform but trades some additional errors for other substantive improvements. Therefore, while expert grades are clearly more detailed, the *Total-Trimmed* induced feature models are competitive and the null assumption of H_{a2} is consequently rejected. This rejection is qualified by the fact that the predictive models had an RMSE of ≥ 0.2 for all of the relevant grades. Thus the predictions are not exact.

The fact that the *Total-Trimmed* models are competitive with the manual grades is both positive and surprising. The primary threshold of comparison was the most frequent grade. I had anticipated that the manual grades would obtain a substantial advantage because of the text. While the graph grammars are complex they make no substantive analysis of the text embedded within the diagrams. Even though the ontology was designed to reify the all important argument structure, the overall quality of the essay also depends upon

the students' ability to frame the text of a hypothesis or claim. Poor framing can and should doom an essay even if the diagram structure holds, and the graders reported that they considered the textual content of the nodes during grading as they were asked. It is therefore surprising that the human graders did not perform noticeably better than the automatic approach.

This lack of advantage may be explained by several factors. Firstly, it is possible that there exists a ceiling effect for the information contained in the graphs. That is the 0.2 RMSE is simply the best that could be achieved given the amount of effort students spent on the later writing. Secondly, it is possible that the collaborative aspects of the assignment, where the class worked together to define the research study and even the hypothesis before diagramming took place, served to standardize the framing used in the key nodes and arcs to set a threshold for quality. As a consequence the textual content conveyed no meaningful unique information. Finally, it is possible that the textual information does convey additional pedagogical information but that the graders too ignored it, focused on different features of the argument, or were merely unable to synthesize the information effectively.

In future work, it may make sense to test these hypotheses by conducting more detailed data analyses and additional grading. If, for example, the textual information is relevant to the quality estimation then it would be expected that feature-based models of the graph grades will perform no better than the essay models or be similar to them. If, however the textual features are not as relevant and the explanation is information loss or other changes, then feature-based models of the graph grades would perform better than the essay models or differ substantially. This would be a surprising result but one that merits further study.

It is also surprising that the CMSE scores were not vastly improved under the induced models or grades. Their performance, while better than the baseline, was consistent with the change in RMSE suggesting that the induced models performed better overall but did not systematically underestimate scores to do it. In future work it may make sense to focus on inducing models via CMSE to test their overall behavior.

7.5.2 Model Inspection

Total-Trimmed-E.14 (Arg-Quality), the best induced model, is shown in Table 7.10 (pp. 129) with its coefficient values. It is relatively parsimonious relying on only 9 of the 61 available features. As noted above, the model was generated in a greedy manner focusing on the most predictive item first and thus, while it is difficult to interpret the coefficients in isolation, it is interesting to note which features made the cut and how they are weighted.

The first point of note is the α value which is 0.82. This puts the base score for the trained model just below the most common score value of 0.85. The next feature is a binary value indicating the presence or absence of a comparison arc, *Order_Elt_Comparison_bin*. If a single comparison arc is present in the diagram then 0.1776 will be added to the base score. This is consistent with the positive nonparametric correlation shown in Table 6.2 (pp. 89). The model, then, has strong negative coefficients for rules *R01na* and *R11ua* in binary form. This is also consistent with prior analyses and with their *a-priori* interpretations.

The strong negative coefficient for *Order_MaxParents_IgnoreEmpty_log*, the maximum number of nonempty parent nodes in the diagram, is harder to interpret. At face value this coefficient (-0.3369 1/3 a full score) is selecting for graphs with a small number of parents, possibly graphs with parent-free nodes. However it is not entirely surprising. In any non-circular graph this number will be 0. Therefore this is selecting, in a roundabout way, against cyclical graphs where every node has a parent and thus at least one node is the child of its own child. Deeper analysis would be required to determine how often this is the case.

Deeper analysis may also be required for the otherwise sensible negative coefficient for *R10c*. The remaining coefficients, however, are wholly counterintuitive. According to our domain experts disjoint graphs are an error *a-priori*. However, in this case the coefficient for the model gives *R13* a positive coefficient of 0.1646. The absent cite and the remaining coefficients function similarly.

Ultimately the individual coefficients are, like the individual analyses discussed in Chapter 5 a suggestive but not definitive result. Some but not all of the coefficients are consistent with our expectations. While the models suggest future hypotheses for analysis, the impact of the individual coefficients is not clear. For a better assessment of the individual rules'

impact we should focus on the nonparametric correlations in Chapter 6 (pp. 86).

7.5.3 Multicollinearity

The relative assumptions of linear regression are often ignored in practice. This is true both for the efficiency assumptions such as homoscedasticity and the hypothesis-testing assumption of normal data. For the purposes of the present analysis, however, it was important to address these assumptions and to note their effect on the generality and reliability of the resulting models. In this respect violations of the non-multicollinearity assumption are a primary issue.

As noted in Subsection 7.2.1 multicollinear datasets can have unstable β coefficients and thus be prone to inflated variance and over-fitting (see also Section G.3 (pp. 248)). I therefore expected that the multicollinear feature sets would have higher RMSE and CMSE scores than trimmed datasets and should trim whenever possible in machine learning. I also expected that trimming the feature-set would result in smaller, more parsimonious models. These assumptions were tested by performing a filtering process on the datasets and inducing models for both the raw and trimmed data. The resulting models were compared in Table 7.6 (pp. 123). A more detailed comparison of the top tier models for *E.14 (Arg-Quality)* is shown in Table 7.11 (pp. 130).

Interestingly, while the *Total-Trimmed-E.14* model is more parsimonious, the predicted relationship between the addition of multicollinear terms (in the Raw dataset) and the RMSE and CMSE scores did not entirely hold. If we consider the Total dataset which had the highest number of such terms, 16 of 77 having been removed, then it is apparent that the added information of the extra terms outweighed the cost of over-fitting. In the overall comparison, shown in Table 7.6 (pp. 123), the raw models had comparable RMSE and CMSE scores or even outperformed the trimmed feature set. This improved performance may be explained in part by the use of RMSE as a target for the greedy induction in *greedyLM*. That alone, however does not provide a full explanation. Given the strong theoretical support for the problems of multicollinearity, this empirical evidence should not be read as positive proof that multicollinear terms can be ignored. However, it is apparent that in an empirical

context they may not pose an existential crisis and that case-by-case testing is warranted. This should be explored in more detail in future work.

7.5.4 Greedy Induction

In the course of this analysis two greedy algorithms were introduced, *GreedyLM* and *Greedy-Tol* shown in Algorithms 7.1 (pp. 118) and 7.2 (pp. 120) respectively. The former was used to search for optimal (as measured by RMSE score) linear models while the latter performed a greedy tolerance reduction of the feature space. Both algorithms make strong assumptions about the structure of the space being searched and it was an open question whether or not they would function adequately here.

The empirical results described above validate the use of these algorithms. While we did not induce optimal models or identify the minimum possible tolerance reductions, it does appear that the algorithms worked in the present case. Subsequent analysis will be required, however, to determine whether their success was grounded in the particulars of this feature space or whether other better models can be found through exhaustive methods. Given the structure of the graph grammars with deliberately isomorphic features, it does seem likely that the space was an easy one for tolerance trimming. It is not clear whether this made the process of identifying linear models easier. In either case later comparisons with brute-force methods may be desirable but depend upon substantial computing power and patience.

Algorithm 7.1 Greedy Induction Model:

greedyLM(OutputVar, Predictors, Partitions, DataFrame)

1: $UpdateFlag \leftarrow TRUE$;
2: $BestP \leftarrow \emptyset$;
3: $BestLM \leftarrow \emptyset$;
4: $PThresholds \leftarrow [0.5, 0.1, 0.05]$;
5: **while** $(UpdateFlag = TRUE)$ **do**
6: $OldP \leftarrow BestP$;
7: $UpdateFlag \leftarrow FALSE$;
8: $NewPredictors \leftarrow (Predictors - BestP)$;
9: $ThresholdIdx \leftarrow min((len(PThresholds) - 1), len(OldP))$;
10: $PThreshold \leftarrow PThresholds[ThresholdIdx]$;
11: **for** $i = 0$ **to** $len(NewPredictors)$ **do**
12: $CurrP \leftarrow NewPredictors[i]$
13: $NewModel \leftarrow lm(OutputVar \sim OldP + CurrP, DataFrame)$;
14: **if** $(pVal(NewModel) \leq PThreshold)$ **then**
15: **if** $(BestLM = \emptyset) \vee (rmse(BestLM, Partitions) < rmse(NewModel, Partitions))$ **then**
16: $BestP \leftarrow (OldP + CurrP)$;
17: $BestLM \leftarrow NewModel$;
18: $UpdateFlag \leftarrow TRUE$;
19: **end if**
20: **end if**
21: **end for**
22: **end while**
23: **return** BestLM;

Table 7.1: Individual Feature Sets.

Total: (77 predictors; 16 trimmed due to multicollinearity) The *Total* set contains every one of the basic graph features and graph rules discussed in Chapter 5 (pp. 61).

Intervention: (43 complex features; 5 simple features; 6 trimmed due to multicollinearity) The *Intervention* set contains only the hand-tooled graph rules and the graph features (e.g. Paired Counterarguments) that may be appropriate for student intervention. These are chiefly specific rather than aggregate features. It makes sense, for example, to direct students to generate a research question node or to add grounding to an ungrounded hypothesis. It does not make sense, however, to advise them to raise and lower the average degree of the graph.

Intuitive: (15 complex features; 3 simple features; 1 trimmed due to multicollinearity) The *Intuitive* set is a subset of the Intervention set containing only the rules and features that matched our intuitive assumptions about their performance. Thus rules such as *R07_UncomparedOpp* were omitted from the set.

Intuitive-NoP: (11 complex features; 1 trimmed due to multicollinearity) The *Intuitive-NoP* set is a subset of the Intuitive set that omits all positive node rules. Positive node rules such as *R01_HasHypothesis* track the occurrence of the beneficial graph nodes such as hypothesis nodes, citation nodes, and so on. The remaining features in this set are negative features such as *R11us_Undef_Ungrounded_Hypo*, complex features such as *Paired Counterarguments*, or arc counts such as the number of supporting or opposing arcs.

Algorithm 7.2 Greedy Tolerance Testing and Reduction:

greedyTol(DependentVar, IndependentVars, Data)

$CurrVars \leftarrow IndependentVars$;

while $TRUE$ **do**
  $MinVar \leftarrow CurrVars[0]$;
  $MinTol \leftarrow (1 - R^2_{MinVar})$;
  for $x_i \in CurrVars[1:]$ **do**
    if $(1 - R^2_{x_{*,i}}) < MinTol$ **then**
      $MinVar \leftarrow x_i$;
      $MinTol \leftarrow (1 - R^2_{x_{*,i}})$;
    end if
  end for

  if $MinTol > 0.01$ **then**
    return $CurrVars$;
  else
    $CurrVars \leftarrow CurrVars - MinVar$;
  end if
end while

Table 7.2: Baseline scores using the most-frequent essay grade as score. The table reports the most frequent baseline value as well as the RMSE and CMSE scores for the baseline values on a per-question basis. When referenced in the text these models will be called *Baseline-**.

Question	Baseline Value	RMSE	CMSE
E.01 (RQ-Quality)	0.75	0.3437	0.4626
E.04 (Hyp-Testable)	0.75	0.2372	0.285
E.07 (Cite-Reasons)	0.666	0.2697	0.2909
E.10 (Hyp-Open)	0	0.4627	0.318
E.14 (Arg-Quality)	0.85	0.2445	0.3254

Table 7.3: Normalization table showing the minimum and maximum observed values for the essay grades coupled with the normalization formula used and the equivalent raw score value of the benchmark error rate of 0.3.

Predictors	Min	Max	$f(x_{raw}) \rightarrow x_{norm}$	Raw (0.3) equivalent
E.01 (RQ-Quality)	-2	2	$(x_{raw} + 2)/4$	1.2
E.04 (Hyp-Testable)	-2	2	$(x_{raw} + 2)/4$	1.2
E.07 (Cite-Reasons)	-1	2	$(x_{raw} + 1)/3$	0.9
E.10 (Hyp-Open)	-2	2	$(x_{raw} + 2)/4$	1.2
E.14 (Arg-Quality)	-5	5	$(x_{raw} + 5)/10$	3

Table 7.4: Direct Graph/Essay Baseline Scores with RMSE and CMSE values. Calculations via leave-one-out cross-validation. Mean and σ^2 values are included to compare the scores while the RMSE and CMSE values are shown on an individual basis. When referenced in the text these will be named *Direct-G/E**.

Question	Graph Grade		Essay Grade		Scores	
	Mean	σ^2	Mean	σ^2	RMSE	CMSE
G/E.01 (RQ-Quality)	0.1405	0.2945	0.5929	0.3072	0.3106	0.3592
G/E.04 (Hyp-Testable)	0.5798	0.2947	0.7667	0.2377	0.2322	0.2828
G/E.07 (Cite-Reasons)	0.5321	0.2763	0.6936	0.2697	0.2475	0.2754
G/E.10 (Hyp-Open)	0.3026	0.3017	0.2988	0.355	0.3389	0.3316
G/E.14 (Arg-Quality)	0.4105	0.2759	0.7386	0.2186	0.2135	0.2565

Table 7.5: Model results for the prediction of reliable essay grades from the graph grades. *Italicized* models are cases where the individual graph grade was the best performing model. The scores were calculated using leave-one-out cross-validation.

Question	RMSE	CMSE
E.01 (RQ-Quality)	0.3106	0.3592
E.04 (Hyp-Testable)	0.2322	0.2828
E.07 (Cite-Reasons)	0.2447	0.2701
E.10 (Hyp-Open)	0.334	0.333
E.14 (Arg-Quality)	0.2062	0.2410

Table 7.6: RMSE and CMSE scores for the induced feature models linking graph features to essay grades. RMSE and CMSE calculated via leave-one-out cross-validation. The scores are grouped on a per-question basis with the models and model values listed by featureset.

		Raw Featureset		Trimmed Featureset	
Grade	Predictors	RMSE	CMSE	RMSE	CMSE
E.01 (RQ-Quality)	Intuitive-NoP	0.3108	0.3604	0.3104	0.3535
	Intuitive	0.3014	0.3396	0.3014	0.3396
	Intervention	0.2924	0.3248	0.2907	0.3208
	Total	0.2902	0.3199	0.2896	0.3257
E.04 (Hyp-Testable)	Intuitive-NoP	0.219	0.2726	0.2202	0.2747
	Intuitive	0.2194	0.2747	0.2194	0.2747
	Intervention	0.2099	0.2411	0.211	0.2557
	Total	0.207	0.2492	0.2119	0.254
E.07 (Cite-Reasons)	Intuitive-NoP	0.2642	0.2979	0.2667	0.3042
	Intuitive	0.2489	0.2752	0.2477	0.2736
	Intervention	0.241	0.2627	0.2371	0.262
	Total	0.227	0.2401	0.2434	0.26
E.10 (Hyp-Open)	Intuitive-NoP	0.358	0.3401	0.3582	0.3401
	Intuitive	0.3228	0.3168	0.3288	0.3284
	Intervention	0.3235	0.323	0.3229	0.3121
	Total	0.3217	0.3279	0.3159	0.3282
E.14 (Arg-Quality)	Intuitive-NoP	0.2148	0.2609	0.2148	0.2609
	Intuitive	0.2126	0.2547	0.2143	0.2569
	Intervention	0.2122	0.2473	0.2126	0.251
	Total	0.2079	0.2415	0.2065	0.2369

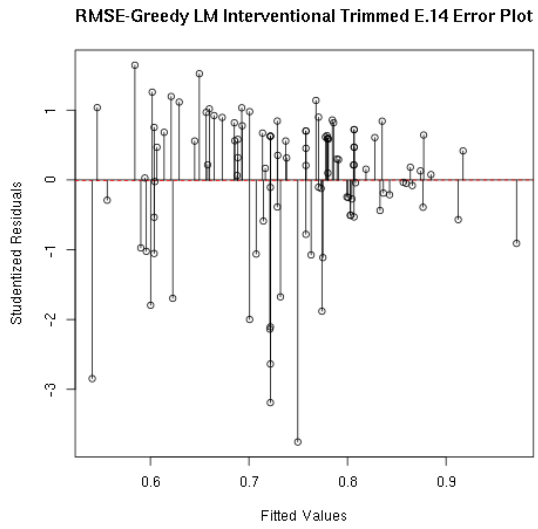
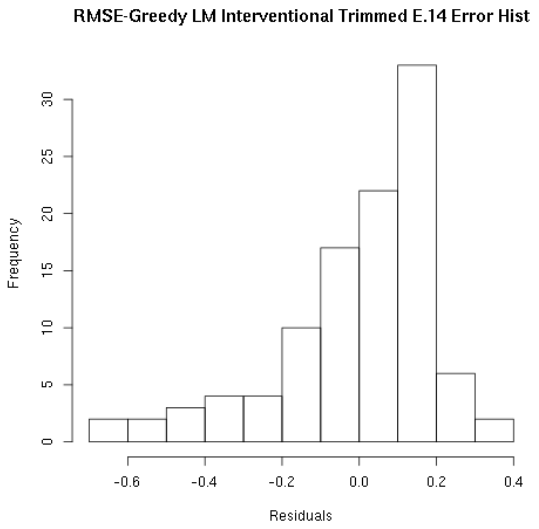
Breusch-Pagan: $p < 0.0017$	Shapiro-Wilk: $p < 2.434e-07$
	

Figure 7.1: Error plots, histogram distribution and statistical tests for the assumptions of Homoscedasticity and Normality for the *Intervention-Trimmed-E.14* model induced for *E.14 (Arg-Quality)* from the Tolerance-trimmed *Intervention* featureset.

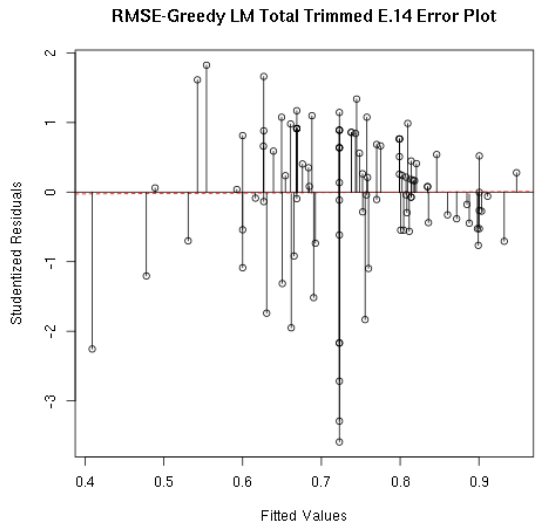
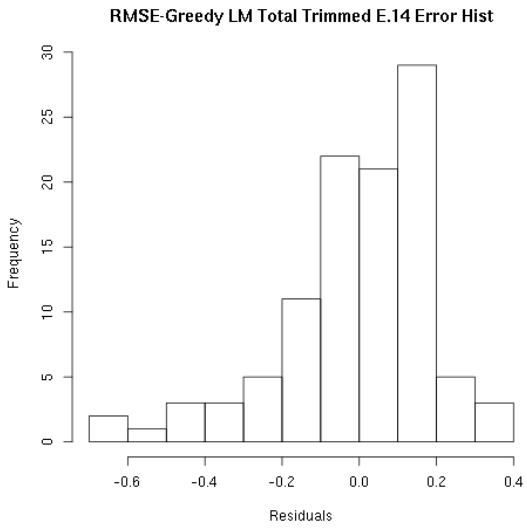
Breuch-Pagan: $p < 0.0005$	Shapiro-Wilk: $p < 1.003e-05$
	

Figure 7.2: Error plots, histogram distribution and statistical tests for the assumptions of Homoscedasticity and Normality for the *Total-Trimmed-E.14* model induced for *E.14 (Arg-Quality)* from the Tolerance-trimmed *Total* featureset.

Table 7.7: RMSE, and CMSE scores for the top-tier Generalized Additive Models defined from the Trimmed *Intervention* and *Total* featuresets. Scores calculated by leave-one-out cross-validation.

	Predictors	RMSE	CMSE
E.01 (RQ-Quality)	Intervention	0.2907	0.3208
	Total	0.2896	0.3257
E.04 (Hyp-Testable)	Intervention	0.211	0.2557
	Total	0.2123	0.2545
E.07 (Cite-Reasons	Intervention	0.24	0.2643
	Total	0.2425	0.2564
E.10 (Hyp-Open)	Intervention	0.3192	0.315
	Total	0.3203	0.3361
E.14 (Arg-Quality)	Intervention	0.2108	0.247
	Total	0.2089	0.2401

Table 7.8: RMSE/CMSE score comparisons for *Baseline* essay grades, *Manual* grader-assigned grades (direct or best model), and the *Total-Trimmed* diagram feature models. Scores calculated using Leave-one-out cross-validation. For statistical comparisons see Table 7.9.

Grade	Class	RMSE	CMSE
E.01 (RQ-Quality)	Baseline	0.3437	0.4626
	Manual	0.3106	0.3592
	Total-Trimmed LM	0.2896	0.3257
	Total-Trimmed GAM	0.2896	0.3257
E.04 (Hyp-Testable)	Baseline	0.2372	0.285
	Manual	0.2322	0.2828
	Total-Trimmed LM	0.2119	0.254
	Total-Trimmed GAM	0.2123	0.2545
E.07 (Cite-Reasons)	Baseline	0.2697	0.2909
	Manual	0.2475	0.2754
	Total-Trimmed LM	0.2434	0.26
	Total-Trimmed GAM	0.2425	0.2564
E.10 (Hyp-Open)	Baseline	0.4627	0.318
	Manual	0.3389	0.3316
	Total-Trimmed LM	0.3159	0.3282
	Total-Trimmed GAM	0.3203	0.3361
E.14 (Arg-Quality)	Baseline	0.2445	0.3254
	Manual	0.2135	0.2565
	Total-Trimmed LM	0.2065	0.2369
	Total-Trimmed GAM	0.2089	0.2401

Table 7.9: Wilcoxon Signed Rank Test results for comparison of the *Baseline*, *Manual* grade (direct or grade model) and *Total-Trimmed* models summarized in Table 7.8. Test statistics and p-values calculated using unpaired Wilcoxon tests. P-values were corrected using Holm correction for multiple tests. (see [29, 138, 132])

Grade	Pair		W	p-value
E.01	Baseline	Manual	5113	1
	Baseline	Total-Trimmed	5311	1
(RQ-Quality)	Total-Trimmed	Manual	5774	1
E.04	Baseline	Manual	5435	1
	Baseline	Total-Trimmed	5602	1
(Hyp-Testable)	Total-Trimmed	Manual	5499	1
E.07	Baseline	Manual	5859	0.861
	Baseline	Total-Trimmed	6035	0.703
(Cite-Reasons)	Total-Trimmed	Manual	5738	0.861
E.10	Baseline	Manual	4599	*0.071*
	Baseline	Total-Trimmed	4595	*0.071*
(Hyp-Open)	Total-Trimmed	Manual	6769	**0.013**
E.14	Baseline	Manual	4411	**0.037**
	Baseline	Total-Trimmed	4400	**0.037**
(Arg-Quality)	Total-Trimmed	Manual	5501	0.98

Table 7.10: Coefficient values for *Total-Trimmed-E.14 (Arg-Quality)* the best induced model from the trimmed Total featureset.

Dataset	# Predictors	RMSE	CMSE
Trimmed	61	0.2065	0.2369

x_i Name	β value
α (Intercept)	0.8204
+ Order_Elt_comparison_bin	0.1776
+ Rule_R01na_NoHypothesis_bin	-0.1379
+ Rule_R11ua_Undef_Ungrounded_Hypo_bin	-0.1941
+ Order_MaxParents_IgnoreEmpty_log	-0.3369
+ Rule_R13_DisjointSubgraphs_log	0.1646
+ Rule_R01nc_NoCite_bin	0.2479
+ Order_MinChildren_IgnoreEmpty_log	0.1222
+ Order_MinDegree	0.0479
+ Rule_R10c_Claim_Comp_bin	-0.0866

Table 7.11: Comparison of the best linear models *Total-Raw-E.14* and *Total-Trimmed-E.14* induced from the Raw and Trimmed Total featuresets for *E.14 (Arg-Quality)*.

Dataset	# Predictors	RMSE	CMSE
Trimmed	61	0.2126	0.251
$E.14 \sim Order\_Elt\_comparison\_bin$			$+Rule\_R01na\_NoHypothesis\_bin$
$+Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$			$+Order\_MaxParents\_IgnoreEmpty\_log$
$+Rule\_R13\_DisjointSubgraphs\_log$			$+Rule\_R01nc\_NoCite\_bin$
$+Order\_MinChildren\_IgnoreEmpty\_log$			$+Order\_MinDegree$
$+Rule\_R10c\_Claim\_Comp\_bin$			
Raw	77	0.2065	0.2369
$E.14 \sim Order\_Elt\_comparison\_bin$			$+Order\_MaxParents\_IgnoreEmpty\_log$
$+Order\_Elt\_hypothesis\_log$			$+Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$
$+Order\_MaxChildren\_IgnoreEmpty\_log$			$+Order\_PairedCounterarg$
$+Rule\_R01na\_NoHypothesis\_bin$			$+Order\_MaxChildren\_log$
$+Order\_MaxParents\_log$			$+Rule\_R01pa\_HasHypothesis\_log$
$+Rule\_R12\_UndefinedCiteClaim\_bin$			$+Rule\_R10c\_Claim\_Comp\_bin$
$+Rule\_R04a\_EmptyNodeFields$			$+Order\_AvgChildren\_IgnoreEmpty\_log$
$+Order\_MinChildren\_IgnoreEmpty\_log$			$+Rule\_R08\_Unopp\_Hypo\_log$
$+Rule\_R10a\_Hypo\_or\_Claim\_Comp\_bin$			

8.0 ANALYSIS & CONCLUSIONS

8.1 CONCLUSIONS

Can argument diagrams be used to diagnose and predict argument performance?

Yes

As I noted in Chapter 1 argument diagrams have a long history in philosophy and AI (e.g. [120, 124, 37, 128]). Yet despite serious research over the years their performance as educational tools has been mixed. Prior research has shown that diagrams can be used to communicate information to students but the effect of diagrams either as static communication tools or exercises has been inconsistent. Some researchers (e.g. [33]) have found that diagrams can help students to recognize key features of a domain or to transfer argument recognition skills, particularly for poor-performing students [92]. Yet other researchers have shown little to no effect from argument use (e.g. [14, 12]). Moreover, no researchers to-date have shown a clear structural connection between student-produced argument diagrams and subsequent performance nor have they convinced educators generally that such diagrams can encode pedagogically useful information.

All of this has been problematic for advocates of argument diagrams. Proponents of such diagrams, myself included, have long argued that argument diagrams have distinct advantages. Argument diagrams can reify important structural concepts, thus helping students to focus on crucial features (see [69]). They can be flexible and open-ended, thus allowing for realistic argumentation even in ill-defined domains (see [76, 74]). And they can be readily evaluated using AI techniques (see [68, 90]). Yet none of these has been shown to hold consistently.

In order for an educational intervention to be useful, it should meet standards of reliability and validity. That is, the argument diagram structure should encode pedagogically useful information that can be graded *reliably* by domain experts and the grading should be a *valid* predictor of student's subsequent performance. If automatic grading of the diagrams can be equally reliable and valid, the structure will be more applicable in real-world tasks. My goal in this thesis was to address these criteria by answering the following questions:

Q_h Can student-produced argument diagrams be assessed reliably by *human* graders and are those assessments valid predictors of future performance?

Q_a Can argument diagrams be analyzed *automatically* to diagnose students' argumentation skills and to predict future performance on "real-world" tasks?

To that end I broke question Q_h into the four hypotheses $H_{h1} - H_{h4}$ shown below which addressed the reliability and validity of human grading for both the diagrams and argumentative essays.

H_{h1}: Student-produced argument diagrams *cannot* be reliably graded by human graders.

H_{h2}: Student-produced argumentative essays *cannot* be reliably graded using a parallel grading rubric by human graders.

H_{h3}: Human-assigned diagram grades *are not* valid predictors of parallel essay grades.

H_{h4}: Human-assigned diagram grades *are not* valid predictors of their gestalt essay grades.

Question Q_a was similarly broken down into two hypotheses H_{a1} and H_{a2}, which addressed the utility of automatic grading rules both individually and via trained models.

H_{a1}: It is not possible to define *empirically-valid* diagram rules that correlate with students' novel written argumentation ability.

H_{a2}: Automatic features of student diagrams *can not* be used to predict students' novel written argumentation ability.

Student-produced argument diagrams and essays were collected from 178 students in Research Methods taught at the University of Pittsburgh. Those items were graded using a parallel rubric designed to focus on key components of the argument as well as its overall quality. Hypotheses $H_{h1} - H_{h4}$ were addressed via a pair of *reliability* and *validity* studies

discussed in Chapter 4. Likewise hypotheses H_{a1} & H_{a2} were addressed via statistical evaluations of graph features, discussed in Chapter 5, and regression model induction discussed in Chapter 7.

As I discuss in those chapters, all six of the null hypotheses were falsified with some caveats. Therefore, student-produced argument diagrams and essays can be graded *reliably* by domain experts and used as *valid* predictors of students' subsequent essay grades. Automatic graph grades can also be used as valid predictors of subsequent essay performance. Moreover, the automatic grading achieves levels of performance that are competitive with the human graders despite the fact that the graph grammars were limited to structural features of the diagrams and ignored the essay components.

Thus the answers to the question presented at the outset is *yes*.

While the above research was successful, it was not perfect. In the next section I will discuss some of the challenges of the grading model and focus on one graded pair that was problematic for the human graders. I will then discuss some of the immediate applications of these rules for education (see Section 8.3) and then conclude with a discussion of the applications of this work to various research fields along with future extensions (see Section 8.4 and Section 8.5).

8.2 GRADING CHALLENGES

Agreement between the two experimental graders was generally strong on the diagramming task with statistically- or marginally-significant agreement on all 14 diagram grades. For the essays it was weaker with statistically-significant agreement on 7 of the 14 grades including the *gestalt* grades *E.12 (Arg-Coherent)*, *E.13 (Arg-Convincing)*, and *E.14 (Arg-Quality)* (see Table 4.4 (pp. 53) in Chapter 4 (pp. 42)). Despite this general agreement, however, some difficult cases exist. One such example is highlighted here to illustrate future changes that may be required. The paired diagram shown in Figure 8.1 (pp. 139) and essay shown in Tables 8.2 (pp. 142) - 8.3 (pp. 143) were drawn from the graded dataset covered by both

graders. This pair was difficult for the human graders. In particular, the graders differed widely on the overall quality of the diagram (*G.14 (Arg-Quality)*) with the primary grader giving it a score of -2.5 and the reliability grader 3.5 on a scale of -5 to 5. They generally agreed, however, on the quality of the essay (*E.14*) with the primary grader assigning it a score of 0.5 and the reliability grader assigning a score of 0 on the same scale. The best fitted linear feature model (*Total-Trimmed-E.14*) of the set trained in Chapter 7, trained as it was on the grade features and the essay grade, predicted an essay score of 2.5.

8.2.1 Diagram Analysis

As noted above the diagram posed a problem for reliability. The primary grader spotted a number of issues with the diagram and generally graded it lower than the reliability grader across the board. The spread between the two graders is shown in Table 8.1. As the table shows the graders differed by as much as 1.5 points (on a scale of -2 to 2) on most of the questions.

As the table shows the graders disagreed about the appropriate use of citations in the diagram and the role that they played in the argument (questions *G.03 (RQ-Support)*, *G.06 (Cite-Conclusions)*, *G.07 (Cite-Reasons)*, & *G.08 (Claim-Support)*). They also disagreed on issues of novelty and openness (*G.10 (Hyp-Open)*, *G.11 (Study-Novel)*) as well as issues of gestalt quality (*G.12 (Arg-Coherent)*, *G.13 (Arg-Convincing)*, & *G.14 (Arg-Quality)*). Interestingly the graders agreed on important components of the argument such as the quality of the research question and the testability of the hypothesis (*G.01 (RQ-Quality)*, & *G.04 (Hyp-Testable)*).

This disagreement may be due to differences in training or to ambiguities in the structure of the diagram. While citation #33 (see Figure 8.2) is described correctly and, based upon the semantic content, appears to have been used to provide a definition, the student has linked it incorrectly with a supporting arc from the claim to the citation. Citation #5 by contrast is provided with a short description and linked to the remaining components via supporting arcs, but it is not clear whether it is providing a definition for the claims being made or a form of support as the arcs give small exemplary notes but provide no

Table 8.1: Per-grader grade scores and value spread for the example diagram.

Grade	Range		Grades	
	Min	Max	Primary	Reliability
G.01 (RQ-Quality)	-2	2	-2	-2
G.02 (RQ-Link)	-2.5	2	-2.5	-2.5
G.03 (RQ-Support)	-2	2	0	2
G.04 (Hyp-Testable)	-2	2	1	1.5
G.05 (Hyp-Link)	-2.5	2	-2.5	-2.5
G.06 (Cite-Conclusions)	-2.5	2	-0.5	1.5
G.07 (Cite-Reasons)	-2.5	2	0	1.5
G.08 (Claim-Support)	-2.5	2	-1.5	0.5
G.09 (RQ-Open)	-2.5	2	-2.5	-2.5
G.10 (Hyp-Open)	-2.5	2	-2	0
G.11 (Study-Novel)	-2.5	2	0	2
G.12 (Arg-Coherent)	-2	2	-0.5	1.5
G.13 (Arg-Convincing)	-2	2	0	1.5
G.14 (Arg-Quality)	-5	5	-2.5	3.5

argumentative content. Thus, while the diagrams reify important aspects of the argument they are not fixed and it is possible that this ambiguity was read differently by the graders.

More serious issues exist, however, with citation #30 and citations #27 & #5 (see Figure 8.2). Citation #30 is, apparently, being used to provide a qualification to the central claim by noting that people engage in unintentional mimicry. This is an important point and clearly relevant to their central research question (and subsequent hypotheses though no such link is drawn). However, the student drew a comparison arc to connect the two. As such he or she has chosen to express the disagreement in a way that violates our argument model. Ironically the opposite issue is apparent with citations #27 & #5 where the diagram author(s) included conflicting citations but used an opposing arc to state a basis of disagreement. The author(s) also failed to connect citation #27 directly to the remainder of the argument, thus leaving unclear how the distinction will play out.

The students' use of arcs was also somewhat inconsistent generally. Supporting arc #8 (see Figure 8.2) is being used to state a claim rather than an argumentative relationship, while the remainder of the arcs are used to characterize the functional role of the relationship (e.g. supporting arc 12 "Example of 1 study" (see Figure 8.3)). The student also failed to include arcs where arcs should be used such as to note the relationships between the unfounded claim #7 (see Figure 8.2) and any backing information.

Despite the features described above, however, there exists a chain of reasoning within the diagram. The bottom half of the diagram (claim #9 to hypothesis #24 Figure 8.3) is structured as a rhetorical chain or linear outline that moves from one citation through descriptions of the current study to the hypotheses. In that respect it is more akin to a classical textual outline than the nonlinear argumentation structure that was desired. In my own review of the diagrams, it appears that more than one student author similarly opted not to produce a structured representation of the argument but to generate what appears to be a semi-ordered representation of their written plan. More thorough study is required to determine if this is consistently true or if the ordering is itself indicative of the essay structure.

It will be necessary to follow up with the graders in order to determine why their interpretations of the argument differed. Nevertheless, the issues with the diagram pose some suggestions for future work and highlight some limitations. Specific rules designed to detect disconnected citations and improper use of comparison arcs already exist. Other rules may be defined, for example, to encourage additional text in the arcs and to flag direct connections between current study nodes.

These changes, however, are unlikely to address the full set of problems. Consider the semantic problems with supporting arc #8. While the structural isolation of claim #7 could be flagged, the fact remains that the content of the arc is framed as a claim and *does not* make clear how the claim relates argumentatively to the rest of the diagram. Similar problems exist with the content of other arcs and the framing of the citations. Problems of this type are semantic and are, ultimately, about the relationships between the texts within the boxes rather than the structural components of the argument and the diagrammatic relationships between them. As such they may require the application of NLP techniques or may be better solved by means of expert guidance or peer review or through advanced semantic analysis. If such analysis is automated however, it often requires substantial bodies of text in order to make comparisons. Such text may not always be available.

8.2.2 Essay Analysis

While the graders did agree on the overall quality of the essay shown in Tables 8.2 - 8.3, they disagreed on a number of the features. Focusing solely on the reliable items (see Section 5.5), they disagreed on *E.01 (RQ-Quality)* (0 vs. -2), *E.04 (Hyp-Testable)* (-1 vs. -2), *E.07 (Cite-Reasons)* (1.5 vs 0.5), & *E.10 (Hyp-Open)* (-2 vs. -2.5). Thus while the primary grader was more negative about the diagram she was more positive about the essay.

The two most disparate grades were *E.01* and *E.04* which focus on the research question and the hypothesis. While the associated diagram contained a single central claim, it contained no apparent research question. No such question was apparent in the essay text either. Thus the grades for *E.01* assigned by the reliability grader were consistent with my own assessment. The primary grader, however, was more generous. That generosity may be

explained by the fact that she had taught a section of the course and thus was more generous to the students on their framing.

While the level of the disagreement on *E.04* is relatively slight, it may be explained in part by the students' writing. In the diagram the author included two hypothesis nodes (#23 & #24 see Figure 8.3) which received relatively high scores from the graders. The nodes are odd in that they reflect both outcomes of the boolean variable but they are stated in clear language. The essay appears to include two similar statements, as well, at the end of paragraphs 3 and 4 respectively (see Table 8.3). This contradicts the assignment instructions that called for a single statement. Oddly enough this framing does not differ substantively from the framing used in the diagram, which suggests that the graders' disagreement stems from differing assumptions about the writing and a much higher standard for framing within the essay than within the diagram. Again, however, follow-ups with the graders would be required to confirm this.

8.2.3 Example Discussion

Ultimately this example illustrates some of the major challenges that should be addressed in future work. Despite the reification many of the issues with the diagrams are textual and semantic, not syntactic. This means that the framing serves to restrict the problems but does not solve them. Moreover the students' use of the diagrams, while inconsistent with our instructions, may not be wholly irrational. In this case, the student opted to use the diagram, somewhat consistently, to describe the order of his or her paper and spent less time considering the argumentative interrelationships. Thus we gave them a screwdriver and they, maddeningly, used it as a chisel, serviceable but still not desirable. In such cases, interventions should be carefully tailored not to reject such nonconforming activities entirely but to help students incorporate the desirable argumentative structures required.

Another issue faced by this work lies simply with the essays and with consistency. The graders were generally in agreement but differed on this particular case. The nature of their assumptions differed, however, from diagrams to essays and was likely informed by their experience. The primary grader had taught the course and worked with both diagrams and

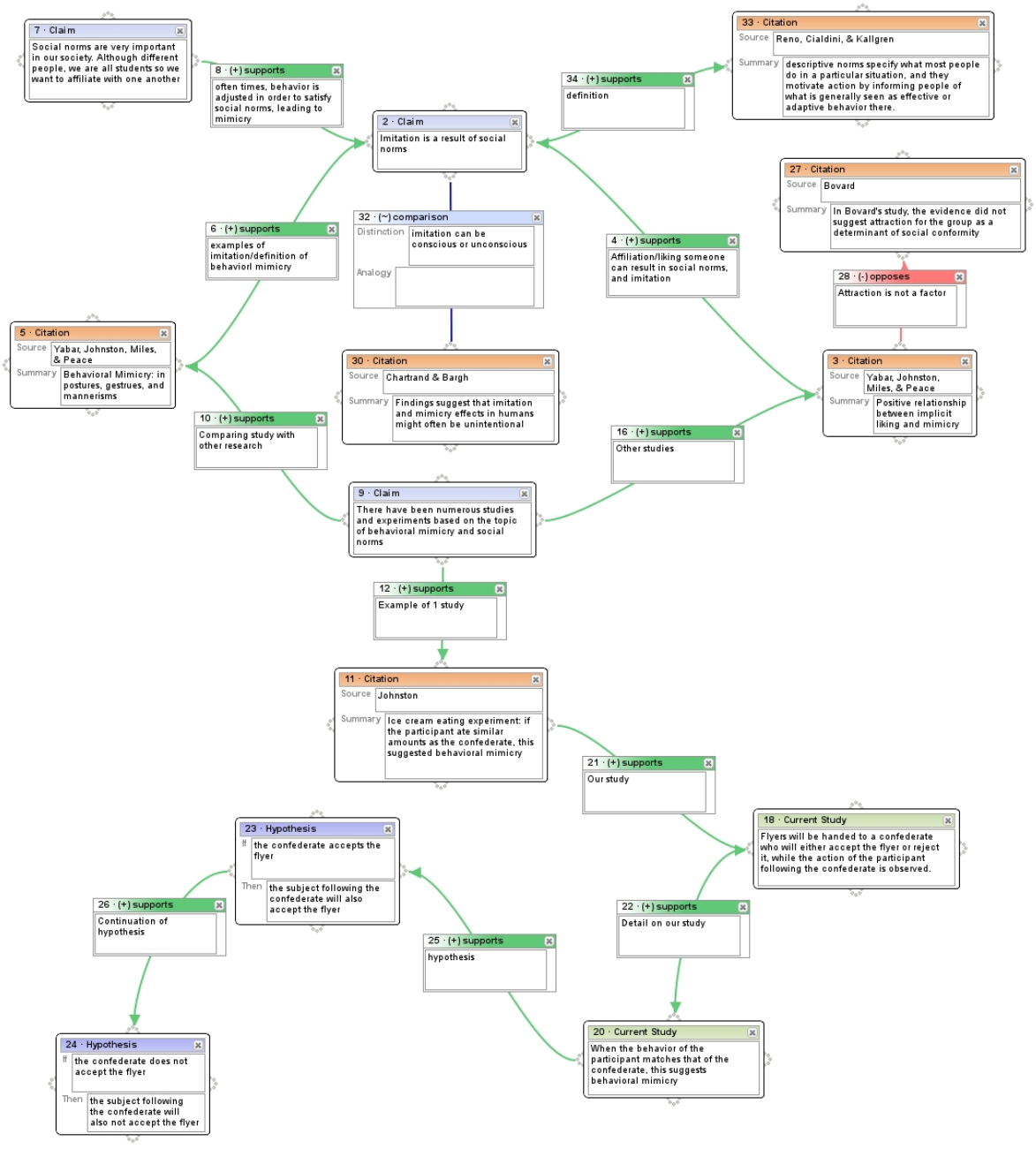


Figure 8.1: A difficult-to-grade diagram drawn from the graded dataset. This diagram is paired with the sample essay shown in Tables 8.2 (pp. 142) - 8.3 (pp. 143)

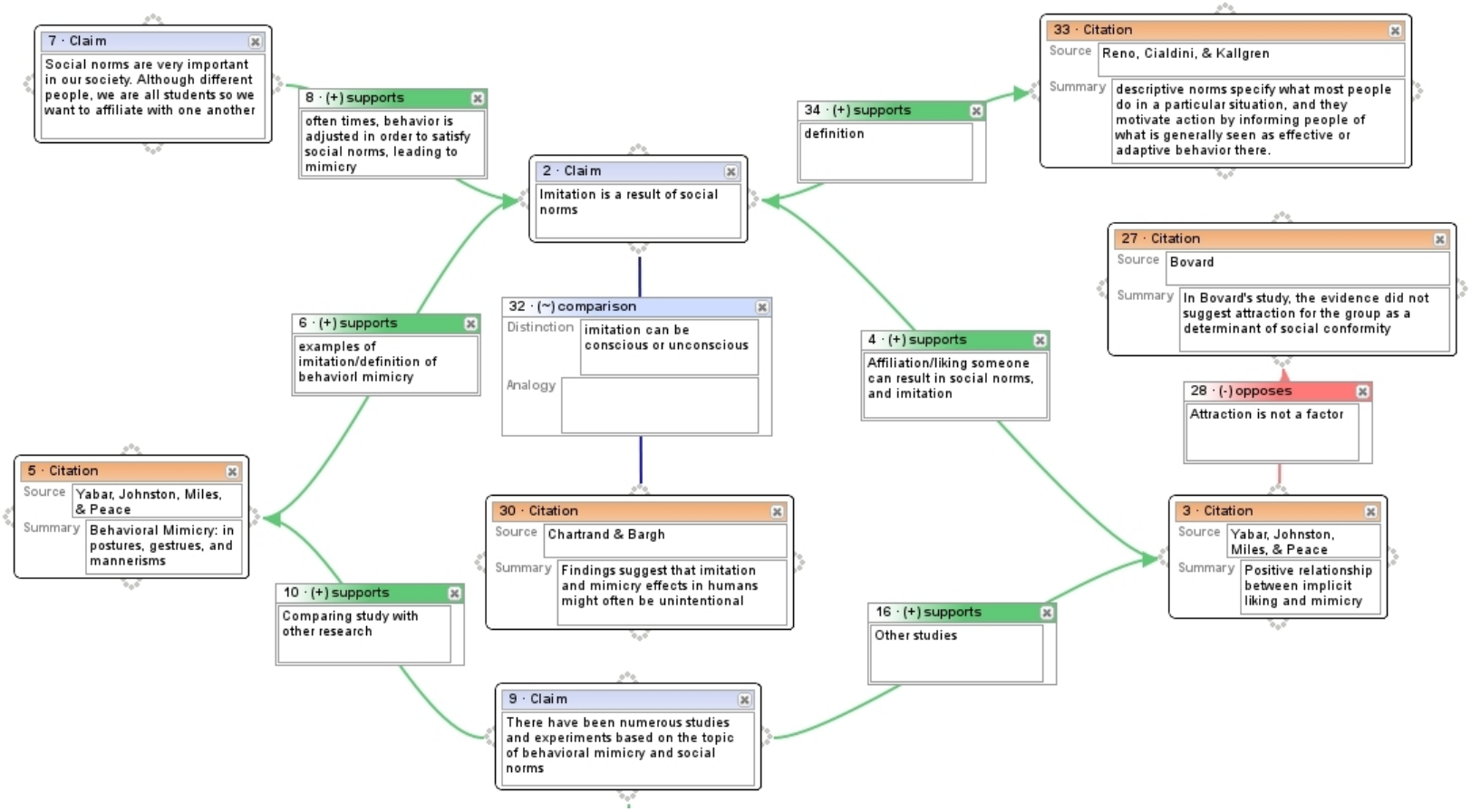


Figure 8.2: The top half of the difficult-to-grade diagram shown in Figure 8.1.

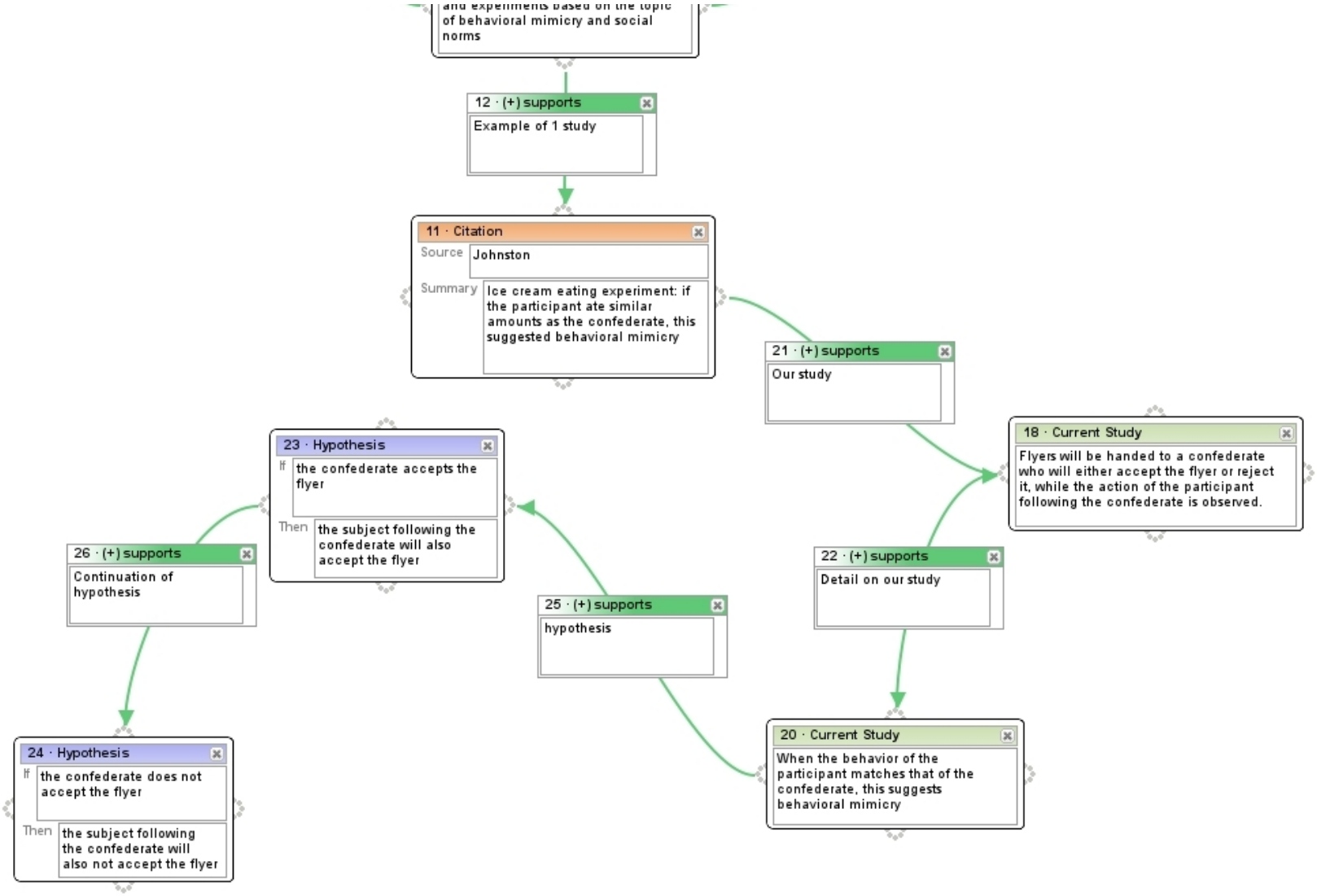


Figure 8.3: The bottom half of the difficult-to-grade diagram shown in Figure 8.1.

Table 8.2: Part 1 of a sample complex essay drawn from the graded dataset. This was easy for grader agreement but difficult for prediction and is paired with the diagram shown in Figure 8.1 (pp. 139)

The Influence of a Confederate's Action on the Subsequent Response of Participants
The social issue of conformity is prevalent in many societies. It refers to the influence that an individual or group has on another. These influences can affect the behaviors, actions, and attitudes of those who are being conformed. The act of conforming can be either conscious or unconscious and is often a result of the desire for social acceptance. According to Johnston (2002), even the idea of another person merely practicing a specific behavior makes an individual more inclined to mimic that same behavior. Conformity and imitation go hand in hand as a result of the inevitable impact of mimicry. This increase in mimicry can develop commonalities in social behavior, leading to conformity. Finally, this results in the initiation of social norms that become particular to a community. Social norms define the adaptive behavior of individuals in certain situations based on what is approved or disapproved in that particular situation (Reno, Cialdini, & Kallgren, 1993).

The act of physically imitating another individual is often referred to as behavioral mimicry. This includes similar hand gestures, facial expressions, postures, and other mannerisms that are observed (Yabar, Johnston, Miles, & Peace, 2006). Previous research has suggested that there is higher tendency to imitate those with whom an individual more frequently associates himself or herself (Bovard, 1953). The reasoning behind this may be attributed to the desire to feel more united with the general population.

Table 8.3: Part 2 of a sample complex essay drawn from the graded dataset. This was easy for grader agreement but difficult for prediction and is paired with the diagram shown in Figure 8.1 (pp. 139)

The Influence of a Confederate's Action on the Subsequent Response of Participants (cont).

With the size of a University campus, it is not possible for each student to have a defined relationship with one another. However, the population has a connection based on their role as students. The effect of association was demonstrated in previous studies. It was found that mimicry was greater among members of the same group (Yabar, Johnston, Miles, & Peace, 2006; Chartrand & Bargh, 1999). In the current study, the affiliation as students develops social norms within the campus, resulting in the tendency to act similarly to achieve social acceptance. In the class study, the target population was individuals in a University campus. The study observed the action of a student immediately following a confederate who either accepted or did not accept a flyer. If a student's action imitates that of the confederate, this may suggest conformity by means of behavioral mimicry.

Our study is a continuation of previous research that has tried to define reasons for conformity. By observing the response of students, the influence that peers have on one another in terms of behavioral characteristics can potentially be determined. This study aims to connect the behavior of a confederate with that of the subject immediately following, such that when a confederate accepts a flyer, the immediate subject will also accept a flyer.

---

essays while the reliability grader had trained with the finished diagrams but not used them in a classroom setting. As a consequence it seems that the primary grader was inclined to be far more negative about the diagrams and somewhat more positive about the essays. This varying experience may also explain why the level of agreement for the essays was so poor. Despite the fact that the research reports were the same and their training very similar, the graders may have fallen back on differing experiences to make their essay decisions. Any subsequent graded use therefore must address these problems in the design of appropriate training. These considerations also limit the generality of the conclusions drawn.

8.3 AUTOMATED ADVICE

Having shown the challenges of diagram analysis I will now discuss the direct applications of the rule induction. The diagram shown in Figure 8.4 (pp. 145) was previously discussed in Subsection 3.3.2 (pp. 37). As noted there the diagram was graded poorly. Interestingly, some of the salient problems with the diagram are amenable to automated analysis. For example, the diagram has an isolated hypothesis node which would trigger rules *R02a: NonHypo w/o outlink*, *R11a: Ungrounded Hypo*, and *R13: Disjoint Subgraphs*. The diagram has opposition but, due to the arc direction, has no *paired* or *chained counterarguments*. The diagram has no current study node which would trigger rule *R01n**.

Having shown that some of these rules are individually predictive in Chapter 6 they can then be used to provide direct advice to the student authors *or* to guide reviewers as I will discuss below. In the former case it would be necessary to augment the existing system with help messages associated with the rules and to provide these rules to the students as guidance. As will be discussed in Subsection 8.4.2 work of this type is already underway. Alternatively, these rules can be used to highlight key sections of the diagram for *expert instructors* or *peer reviewers* who will be tasked with providing direct advice to the diagram authors. As will be discussed in Subsection 8.4.2 this has distinct advantages and is also being tested presently.

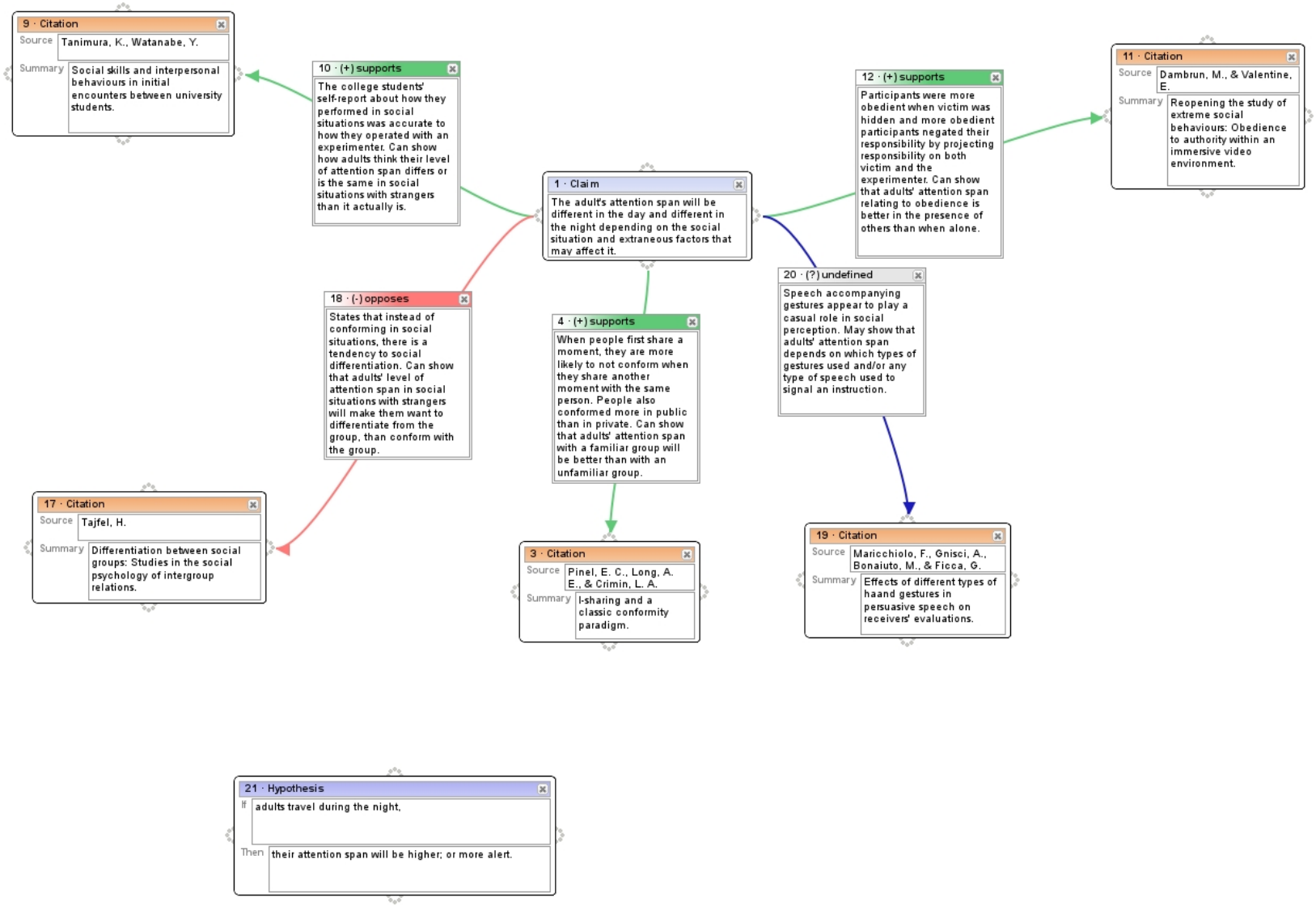


Figure 8.4: A sample diagram previously shown in Figure 3.4 suitable for automated advice.

Finally these rules can also, as noted previously, be used for automatic ranking. In a large-scale course it is not possible to provide detailed expert instruction to every student. However through automated rules of the type induced in Chapter 7 it may be possible to sort students for advice and then to present that information in an *instructor dashboard*. This would enable instructors to get a broad view of the course both by indicating the number of students succeeding and failing overall and by summarizing the most common errors. If, for example, most of the student diagrams trip rule *R11a: Ungrounded Hypo* then the instructor can focus their discussion on hypothesis statements and their connection to the literature. I will return to these points below.

8.4 CONTRIBUTIONS

This work makes research contributions to a number of domains including education, intelligent tutoring, and graph analysis. This work has also made technical contributions in the form of novel machine learning algorithms, and analysis tools. I will discuss these contributions individually below.

8.4.1 Education

The primary contributions of this thesis have direct relevance to education. In collaboration with Dr. Melissa Patchan and Dr. Chris Schunn, have developed a novel structural ontology and graphical representation for written scientific arguments. This representation, described in Section 3.2 (pp. 28), was specifically tuned to the types of arguments present in undergraduate research reports. This structure was encoded in the LASAD argument diagramming system and was used as a pre-writing exercise to support students in developing written research reports.

I have also shown that argument diagrams can be reliable and valid predictors of student performance on subsequent *authentic* tasks. As such they can be useful educational tools both as a vehicle for evaluation and intervention. Moreover, as the results of the reliability

analysis demonstrate (see Table 4.4) the diagrams can be graded *more reliably* than essays for argumentative features and the diagram grades are, in turn, predictive of the essay grades. The reification provided by argument diagrams is therefore beneficial for both students and instructors.

8.4.2 Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITSs) have shown success in many fields including physics, mathematics, and chemistry. To date ITSs have been less successful in more open-ended or ill-defined [73, 74] domains such as writing, design, or public policy. In recent years, however, there has been increased interest in these areas in work such as that of Easterday et al. [33], Roscoe et al. [99], and Chryssafidou & Sharples [20]. The present work also contributes to this research by demonstrating the utility of argument diagrams as pedagogical diagnostic tools. This work also highlighted a process for iterative evaluation of rules in an ill-defined domain through the use of exploratory data. This iterative evaluation is useful both to support automated diagram assessment, and to assess the *empirical validity* of the individual rules or constraints which allows us to assess the benefits of weak theory scaffolding. This procedure can be applied in other similar tutoring contexts, even those that do not use diagrams.

Argumentation is an open-ended domain and one that is problematic for expert systems. Rather than build a tutoring system with 'correct' rules predefined it is possible to first define a help-free system of the type used in the 2011 Study. Rules can then be articulated by domain experts and tested against the existing data as was done in Chapters 5 and 7. Expert-defined rules that have empirical validity can then be incorporated into subsequent versions of the system or used to refine existing assumptions. These sets can then, in turn, be used to collect additional data which are available for subsequent refinement. In this way we start with an "exploratory" dataset that can function as a testbed for iterative improvement. This is not the first time that this approach has been taken. A similar approach was taken in [17] where the authors sought to iteratively refine pedagogical policies in a physics tutoring system. This approach has also been used to generate procedural knowledge in an ill-defined

domain in [83] and to test and generate hints in [113]. Thus the present use contributes to this extant literature.

This kind of approach is well suited to online tutoring systems and large-scale systems. In domains of this type it is not always ideal or even possible to articulate expert-defined rules for every error. Nor can all problems be anticipated in an open and uncontrolled domain. Through iterative extraction, however, it is possible to update systems and even to validate expert assumptions about good behavior.

The challenges of this approach were highlighted in Chapter 5. The individual rules, despite being relevant in context, were not always significantly correlated with performance. This issue is, in part, a problem of data. While the amount of data collected from the course was sufficient for present purposes, it is relatively small by machine learning standards. Additional data will increase the sensitivity of the approach. The issue is also partially a problem of grading. While the four detailed reliable target grades (*E.01 (RQ-Quality)*, *E.04 (Hyp-Testable)*, *E.07 (Cite-Reasons)*, & *E.10 (Hyp-Open)* see Section 5.5) represented a range of features, they do not cover all of the pedagogically relevant aspects of the essays, and the single gestalt grade (*E.14 (Arg-Quality)*) used in may simply be insensitive to the problems addressed by an individual rule. Thus additional grading may be required.

When taking an approach of this type, however, it is also necessary to select an appropriate standard for empirical validity. For the present work I chose statistically- or marginally-significant correlation with the gestalt grade *E.14*. Lower standards, via higher p-values, may also be appropriate given the limitations of available data. It may also be appropriate to consider a role in a combined model of the type induced in Chapter 7. As noted, some of the graph features that formed the predictive model were not individually valid. While this work does not definitively settle the question of an appropriate standard, it does highlight the utility of such an approach.

This work also contributes to future work on peer-review by demonstrating the development of flexible ranking models that can be used to guide peer selection. In recent years there has been increased interest in the use of peer-review for writing education. Systems such as Comrade [35] and SWoRD [18] have been used in argumentation courses such as

Research Methods, History, and Human-Computer Interaction, to leverage peer comments in writing education. LASAD diagrams have been used for peer review both on Writing for Biology and in subsequent RM courses at the University of Pittsburgh. In the most recent studies the students were provided with automated advice based upon the complex features described in Subsection 5.4.2 (pp. 69). My colleagues and I are currently testing the impact of peer review on diagrams and subsequent essays.

Peer reviewers, like expert graders, can bring semantic understanding to bear on the analysis of diagrams allowing them to identify textual problems of the type described in Subsection 8.2.1. Peer reviewers, however, are not experienced graders and may face difficulties in identifying salient diagram issues or in providing constructive advice. Prior research at the grade-school level has shown that peer tutoring can be supported by expert systems that are trained to advise the advisors [127]. The models described in the present work can be used to support peer tutoring both by highlighting key features of the diagram that require grader attention and by ranking student diagrams to ensure that students are receiving advice from appropriate peers.

8.4.3 Educational Data Mining

This work makes three primary contributions to Educational Data Mining (EDM) in the area of applications. First and foremost it demonstrates that data mining and machine learning can be fruitfully applied in the semi-structured domain of argument diagrams and that said analysis can be used as a valid predictor of subsequent written essay performance. While educational data mining has been exploding in recent years, most of that growth has been in formal or well-defined domains such as math and physics or in classification tasks such as gaming detection. This work demonstrates that EDM can be applied in more open-ended areas such as writing and note-taking and in ill-defined domains where some scaffolding or *weak-theory structuring* has been done [74].

Secondly this work has demonstrated the potential of hybrid models that combine *a-priori* augmented graph grammars and linear regression. Little EDM work has been done to date on argument mining (see [70]). This work, therefore, serves both to introduce Aug-

mented Graph Grammars to the EDM community and to demonstrate their utility in realistic educational applications.

Finally, the iterative refinement via empirical validation discussed above is relevant both to ITSs and EDM. In the context of an intelligent tutoring system, empirical validation can be used to test individual rule additions and to confirm the utility of the advice. This approach is also ideally suited to large-scale class domains such as MOOCs where data is collected *en-masse*. Such domains can be viewed, if we choose to do so, as a single large ITS where student data from each iteration of the course is analyzed to refine the results for subsequent evaluation and automatic advice. While prior work on iterative refinement in EDM has been discussed above this is the first such work to focus on argumentation and writing.

8.4.4 Graph Analysis & Linear Regression

Graph analysis has always been a fundamental area of computer science (see [111, 11, 57]). In recent years graph analysis has become a major focus of analytical work with applications in social networks [78], biomedical analysis [84], educational data mining [113, 89], and others. Classic graph grammars of the type defined by Rekers & Schürr [97] have been instrumental to this work. As I noted previously, however (see Section 5.3) static graph grammars of this type have limitations when dealing with open-ended or complex data.

Current research problems, such as the analysis of flexible argument diagrams or interconnected document collections, are suited to more complex rules that incorporate both graphical connections and content analysis. The work described here contributes to research on graph analysis by highlighting one such problem, argument diagrams, and demonstrating that augmented graph grammars can be fruitfully applied to it. This work also extends the literature on graph grammars by presenting the augmented graph grammar library in Appendix F and showing a novel application of augmented graph grammars to open-ended argument mining.

While the immediate focus of this work has been on the use and success of augmented graph grammars for hand-tooled rules, the AGG library can readily serve as a basis for

subsequent work on grammar induction. While prior work has been done on induction of more restrictive grammars [24, 60, 59, 58] no work has been done to date on the induction of complex grammars of the type used here. As such this work represents a contribution to the generality of graph grammar techniques.

This work has also contributed techniques to machine learning for linear regression. Chapter 7 presents two greedy algorithms for the induction of linear regression models (*greedyLM* Algorithm 7.1) and non-multicollinear sets (*greedyTol* Algorithm 7.2). While neither of these algorithms is wholly novel they are themselves important extensions of prior work, particularly the heuristic approach taken in *GeedyLM*.

More to the point, in designing these algorithms I addressed issues of the assumptions encoded in linear regression, in particular the problem of multicollinearity. As discussed in Subsection 7.5.3 (pp. 116), multicollinear models are unstable and prone to overfitting under linear regression. This led me to hypothesize that multicollinear datasets with greedy induction would produce more error prone models than non-multicollinear sets. As noted in Subsection 7.5.3, however, that was not the case. Rather the raw data had lower or comparable RMSE and CMSE scores than the trimmed datasets suggesting that the use of RMSE as a greedy induction target may, at least in this case, obviate the need for subsequent reductions.

8.4.5 Technical Contributions

In addition to the research contributions discussed above this work has also made three technical contributions. In addition to the induction algorithms *greedyLM* (Algorithm 7.1) and (*greedyTol* (Algorithm 7.2). This work has led to the development of the AGG augmented graph-grammar library (see Appendix F (pp. 223)), and the SNG grading toolkit (see Appendix E (pp. 215)). These tools were instrumental to the completion of this thesis and will form the basis for future work in rule induction and online evaluation. I also plan additional analyses of the collected data.

8.5 FUTURE WORK

In addition to the concrete contributions described above this work both raises interesting research questions in several domains and will form the basis for future work.

8.5.1 Education

This work highlights a role for argument diagrams in education, even in absence of additional AI support or tutoring. More work remains to be done however. Two immediate extensions exist for the present work. The first goal is to test the connection between argument diagrams and long-term performance. While this work has shown validity for the diagrams, it has not yet been shown whether students' argument diagrams predict their performance on subsequent essays or exams. The distinction is a crucial one. If the diagram grades can only be used to predict the students' current essay grades, then they are useful as a gage of argument quality but not necessarily a gage of long-term argument comprehension or written argumentation ability. If, however, the grades are reliable predictors over a longer term then they may serve as more robust measures and even as independently-graded activities rather than assignment-specific interventions. Some data of this type has been collected from the 2011 course and analyses will take place in future work.

The second task is to assess the impact of argument diagramming on essay structure. While Chryssafidou & Sharples [20] and Carr [14] reported some qualitative differences between diagramming and non-diagramming conditions, no systematic analysis has yet been made. Ideally such comparisons would be made within a course using a within-subjects design or other balanced assignment that allows us to compare the effect of diagramming on students while controlling for course, instructor, and time. However such ideal conditions are not always feasible with real students. As an alternative approach it may be possible to compare essays gathered from the present study to that of other, non-diagram years, or to perform the same cross-section comparison within a course, say from Assignment 2. A sampling of papers from 2010 has already been collected but not yet been graded. I have also collected papers and some grades from the second course assignment where some sec-

tions opted out. I am examining this data to assess its suitability. In either case, while the comparisons would suffer from confounds due to course variation, the analysis would still be informative.

8.5.2 Intelligent Tutoring Systems

The first and most immediate extension of this work is to conduct a subsequent diagramming study in the same course context where rule-based help is available. Such a study would function both as an additional empirical validation of the rules, and the advice based upon them, and as a vehicle for additional data collection. It would allow us to collect additional diagram data that is conditioned on the existing rules which would, in turn, support subsequent analysis, as with the bootstrapping approach discussed above.

Research that builds upon this study is already being conducted at the University of Pittsburgh. In subsequent studies, my colleagues and I have modified the diagramming ontology to respond to instructor feedback, added automatic advice based upon the rules tested here, and subjected diagrams to peer review adding an additional round of user feedback. These studies have been designed to test the impact of the ontology features on student performance and to collect additional data that now incorporates the empirically-validated rules. Additional work is also being done to link the diagram structure to the feedback offered, thus allowing us to localize some of the most relevant features.

In the long run it would be worthwhile to integrate this kind of diagramming component into a writing tool. While the present line of work focuses on guiding students to produce the diagram and then the essay, the fact of the matter is that students will alter their argument as they write and may be more comfortable producing text in parallel with the argument. A number of the study participants commented on their desire for a method to map text easily from one form to the other.

Providing students with both a diagrammatic and written format may aid them in drafting and editing their arguments. However, the use of a dual format should be controlled lest one overloads the students. If, for example, users are able to produce an argumentative diagram and then to link said diagram to an essay, this would help them to focus on smaller

structured outlines and then to fill in the text as needed. This dual-representation format would also allow the tutoring system to leverage both representations when providing guidance, potentially allowing the techniques described here to be efficiently combined with the approaches taken by [99]. These issues were addressed in part in [68] as were features of the translation but more work remains to be done.

8.5.3 Graph Analysis & Linear Regression

While the augmented graph-grammars proved useful in the present work they face some limitations of efficiency and flexibility. Rules such as *R13_Disjoint_Subgraphs* are inefficient as implemented and the present structure of rules such as paths do not take full advantage of existing graph theory. Moreover the present rules do not test the full limitations of the field semantics as they engage in limited textual analysis. Future work with the graph grammars will focus on: improving the efficiency of the implementation; testing theoretical concerns regarding the assumptions; exercising the limits of the textual analysis; and automatically inducing predictive subgraphs rather than relying solely on *a-priori* definition. Given the demands of data-intensive machine learning, and the comparatively large search space, however, this work would likely require both the collection of additional data and continued involvement of domain experts.

In future work I will also test the generality of the machine learning results and of the conclusions described in Subsection 8.4.4 regarding the relationship between multicollinearity and RMSE. In order to test the generality of these results, however, they should be tested in other datasets including some artificial datasets where the level of multicollinearity can be controlled and where subsequent independent validation can take place above and beyond cross-validation.

8.6 CLOSING

The process of this research has been, at times, as fast as an ice age and felt as easy as balancing on top of a quail's egg. But ultimately the results are both timely and promising. Argument diagrams have a long history some of which, perhaps, has been spent as a solution looking for a problem. However ultimately they are pedagogically useful. While this study does not show that diagramming improves writing, it does show that diagrams can serve as a reliable vehicle to critique writing and to predict future performance, a vehicle that allows for both automatic assessment and expert evaluation. More remains to be done including determining whether or not the automatic and expert assessment can complement each other.

APPENDIX A

LASAD MATERIALS

This appendix includes introductory materials to the LASAD system that were provided to the instructors, students, and graders. The first document was a written howto document provided in paper form to all study participants. The document was also used as a reference point for system help. The slides were presented as part of an in-class introduction to the students.

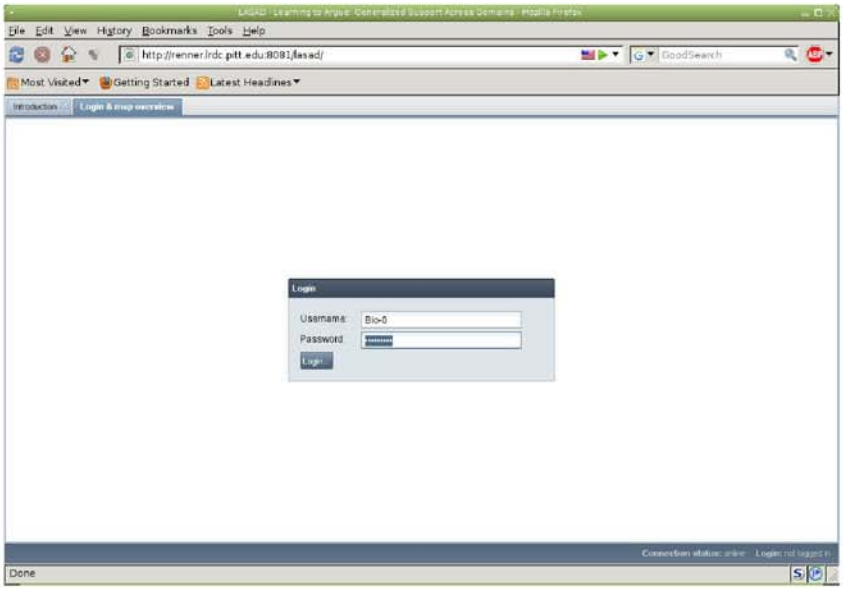
Introduction to the LASAD System.

Collin Lynch. (collinl@pitt.edu)
02/16/11

As you were shown in your classes LASAD is an on-line tool designed to facilitate the production of argument diagrams both as annotations of existing texts and when planning your own writing. You will receive specific assignments from your instructor or via a separate e-mail. In this document you will be shown the basics of LASAD for argument diagramming including logging in, creating diagrams, and annotating cases.

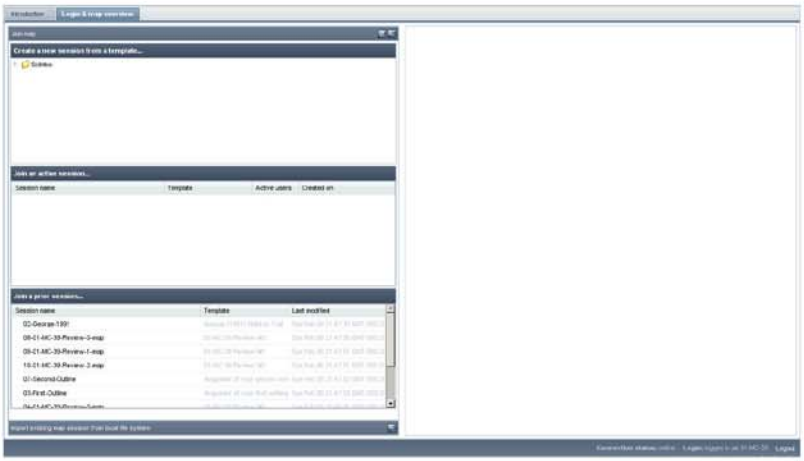
Logging in:

To use LASAD open any standard web browser and go to the URL provided in your class. There you will get the login screen shown below.

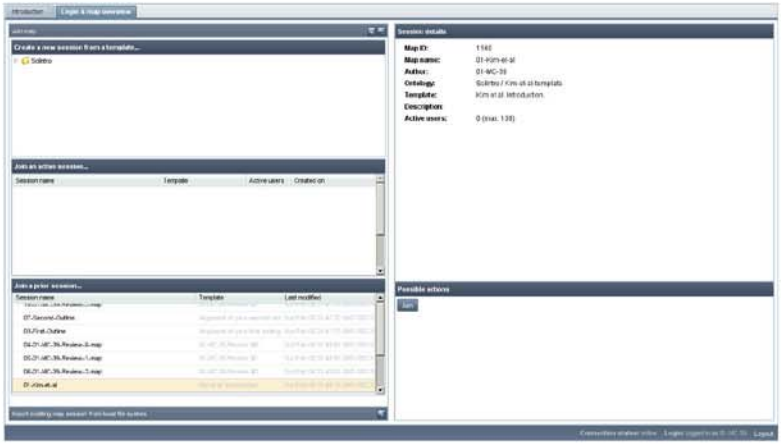


Go ahead and enter the user-name and password that you were provided into the browser as shown, and hit return to login. If you enter incorrect login information a message will pop up in the lower right corner of the screen informing you of the error.

Once you login to LASAD you will be provided with the following browser panel:



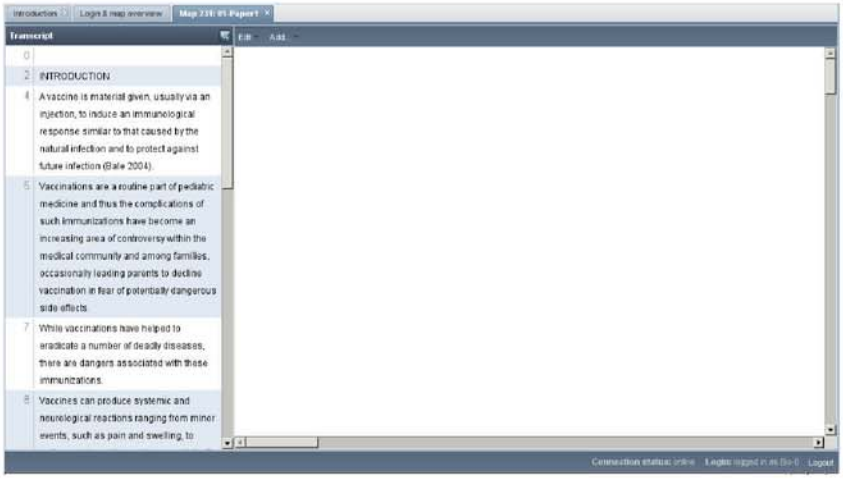
For the purposes of the study you should select a map from the *"Join a prior session..."* panel at the bottom left-hand corner of the screen. These sessions are user-specific so you will be working independently. For the present work you should not opt to *"Join an active session..."* or *"Create a new session from a template..."* Picking a map will cause the Map Details to appear on the right-hand side of the window. The maps are assignment-specific and your instructor will tell you what map to work on for each assignment.



Once you pick a map click on the "Join" button at the bottom of the right-hand side to start work.

Annotation tasks:

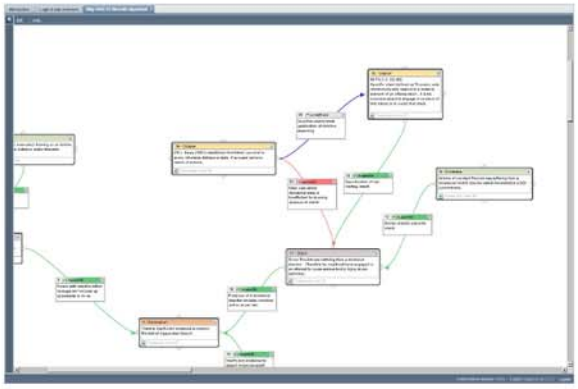
The bulk of your homework assignments will be *annotation tasks* in these tasks you will open up a map that has a set transcript associated with it. When you login to a map of this type you will see the transcript on the left-hand side of the screen and have a workspace on the right as shown below.



The transcript is a series of numbered lines each of which corresponds to a single sentence in the initial text. Both the workspace and the transcript can be scrolled independently to provide space.

Author Diagrams:

Some of your assignments will involve author diagrams where you use the tool to plan your diagramming tasks. For these tasks you will have a screen with no transcript as shown below. As described below you will create nodes without a transcript connection by right-clicking on the diagram panel.



Argument Diagrams.

Diagrammatic models of argument are graphical representations of arguments. These range from very simple "box and line" models, which solely represent text and immediate relationships between them, to more complex formalisms with specific construction rules. Here we employ a relatively simplified model that focuses attention on key parts of the arguments made in empirical work. Here we have four node types:

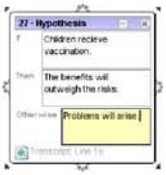
Claims: represent general research questions raised by or claims made by the author of the argument such as those shown at right. Claims can be used to encapsulate any rhetorical point that is important for later discussion.

As shown at right the claim node consists of a single text block and an optional link to the transcript. I will discuss the linking in more detail below.

Citations: represent literature references or links to other related work. This includes references to scholarly works such as published papers as well as other sources such as personal communication.

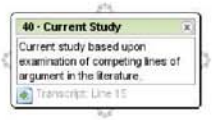
Again see the image at right. As shown here citations contain two text fields, one for the actual citation name, and the other for a summary of the citation if one is provided. It is good form when citing a source to provide at least a brief description of it including essential features such as the methodology employed so that your readers can judge the fitness of the citation in your work. Citations, as with all other nodes may also be linked to the transcript.

Hypothesis: Hypothesis nodes represent a formal empirical hypothesis that is or can be tested by some empirical study. The level of formality desired by your instructor is task-specific but the hypothesis itself is a pseudo-logical node as shown at right.

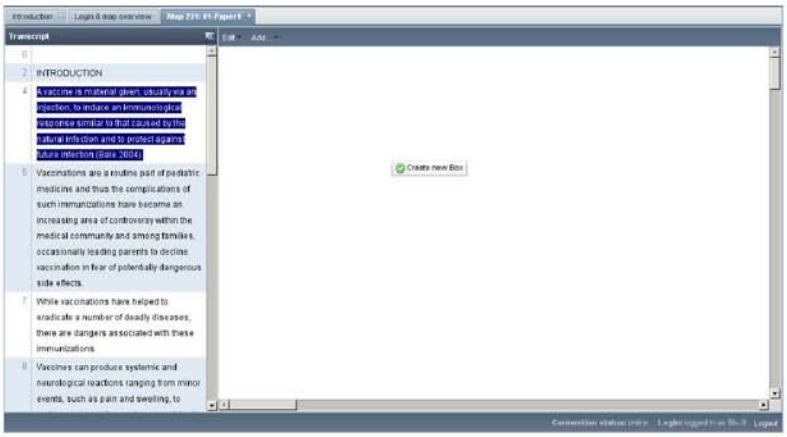


Hypotheses are expressed as if-then rules with the hypothesis node containing fields for 'If' and 'Then' clauses. The third "Otherwise" clause is optional and may be added or removed by selecting the [+] button in the node's titlebar. It can be removed by the [-] button.

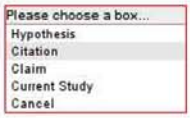
Current Study: The current study nodes offer an opportunity to make short comments about the structure of the current study. This is not about stating your methods section but more about referencing important characteristics of the current study when comparing the planned study to prior literature. Like the claim nodes a current study node consists of a single text field and optional transcript link.



Nodes can be added to the transcript in three distinct ways. First, and foremost, a node may be generated by highlighting a portion of the transcript as shown below. The user can then click on the highlight and by "dragging" it into the diagram workspace they will be asked if they want to generate a new node as shown in the image below.



Releasing the mouse will bring up a selection menu from which you can choose the type of node to be created. The resulting node will be linked to the specified portion of the transcript as represented by the line reference at the bottom of the node. Transcript links can be used to scroll automatically to the selected line.



Unlinked nodes can be added by right-clicking on the diagram space to generate a new node. On Macs this can be accomplished by using a two-finger click or via the *Add* menu at the top of the panel. This brings up the same addition menu as shown above. Unlinked nodes behave the same as others save that they do not contain a reference to the transcript and one cannot be added.

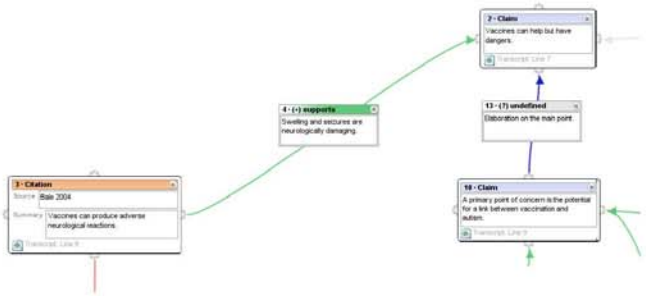
During annotation tasks you should use the highlighting method to maintain the connection to the text. During the authoring tasks, right-clicking is the best method for node addition.

Once nodes are created you can draw arcs between them. The Intro system defines four types of arcs, each representing a different type of relationship:

(+) Supporting Arcs: indicate a supportive relationship with one node (the tail node) supporting the root node. An example of this relationship drawn from an argument diagram is shown in the figure below. The green arc drawn from citation node #3 in the lower-left to claim node #2 in the upper right represents the citation #3 being used to support claim #2.

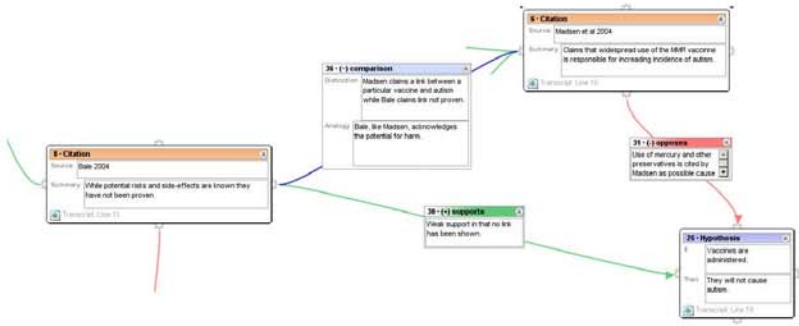
At the center of the green supporting arc is a text field indicating the reason why #3 supports #2. When forming your argument it is important to think not just about what facts you have on your side but why those facts or sources support the claims being made and to make that clear to the readers. That is the function of this field.

(?) Undefined Arcs: represent a factual or topical relationship between two nodes that is neither positive or negative. In the figure below such an arc is shown connecting from claim #10 to claim #2 on the right-hand side. Here the rationale in the text-box still expresses *why* the relationship exists but with no polarity. During annotation tasks the content of the arcs should be drawn from the text itself while during authoring tasks you are free to specify as desired.



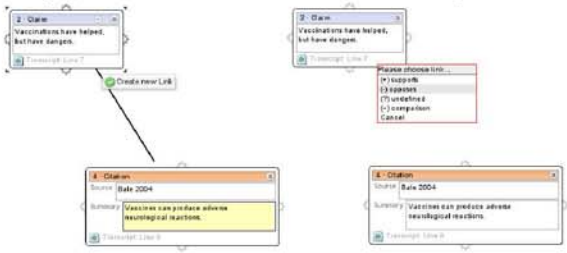
(-) Opposing Arcs: represent the opposite polar case from support. Here the tail node is opposing or undercutting the content of the head node. An example of this is shown in the figure below where citation #6 opposes hypothesis #25 citing the fact that the materials used are dangerous thus reducing the likelihood of the hypothesis holding true.

(~) *Comparison Arcs:* represent a means of drawing analogies and distinctions between two nodes, particularly between two citations or between citations and descriptions of the present study. When presenting literature in your introductory section, particularly literature that relates differently to shared neighbors as is the case here (citation #8 support the hypothesis while citation #6 opposes it). In this case we draw a comparison between the citations. Comparison arcs contain, at minimum, two text fields one for analogies or similarities between the studies, and the other representing distinctions between them. Additional fields can be added through use of the [+] button on the titlebar of the box located at the center of the arc. This button will become visible when the mouse is over the box.



In the example shown above the two studies are similar in that both acknowledge the potential danger of mercury and other materials. Yet they differ in that Madsen claims to draw a specific link between the MMR vaccine and autism while Bale states that no such link has been proven. Identifying such comparisons and presenting them in your literature review helps to ensure that your readers see why you cite apparent conflicts and how you plan to resolve them.

Drawing Arcs: in order to draw an arc between two boxes you should click on one of the four anchor nubs located on the sides of your designated tail node. Then, holding down the mouse you can drag the resulting line until it impacts your target node. Once that is done you will be presented with an arc-type menu (shown at right) similar to the node selection menu. Selecting the arc sets the type.



You can reverse the direction of an arc. with the reversal button located on the titlebar of the arc box. This button will appear when the mouse hovers over the titlebar of the arc. The button is shown below:

The [+] button which appears on the arcs when the mouse hovers over them is shown below. This is a general addition feature that can also be used to add optional features to the hypothesis node though a similar "mouse-aware" button.

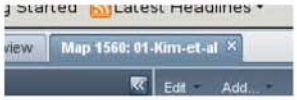
Deleting contributions: all items in the space can be removed by clicking on the [x] button located in the upper-right hand corner of each node or arc title pane.

Saving Diagrams:

Your work on LASAD is saved as you work so you do not need to take any special steps to save it. As long as you login to the same server each time your work will remain.

Closing LASAD:

When you are done working with a map you can close it by clicking on the [x] at the top of the map tab as shown below. This will bring up a popup asking you if you wish to close the map. Clicking yes will close it. Your work is saved automatically to the server as you work and can be reloaded by following the same initial loading steps.



In order to close the system clock the [Logout] button located on the bottom right-hand corner of the window as shown below.

System Information.

LASAD is an entirely web-driven application and, as such, stores your work on the server. You will not need to extract or print-out your diagrams for the course and you will be able to login multiple times using the user-name and password provided to retrieve them.

LASAD is written in JavaScript and thus is cross-platform compatible. However prior students have reported that the system is slower when being accessed through Internet Explorer. We thus encourage you to consider Firefox or other alternate browsers.

Some users have reported that, when a node is deleted the "System Resources" panel will expand into the field of view. This panel does not serve a purpose in the present system and  thus can be ignored. It can be collapsed by clicking on the collapse button:

For further information you can contact me via e-mail: collinl@pitt.edu

Diagramming with LASAD

Collin Lynch

Intelligent Systems Program & LRDC
University of Pittsburgh,
Pittsburgh, Pennsylvania.
02/2011

Scientific Argument

- Science is about communication.
  - State clear research questions.
  - Advance defensible answers to the questions.
- Structure:
  - Identify open research questions.
  - Identify relevant research hypotheses.
  - Make general research claims.
  - Defend those claims as being:
    - *appropriate,*
    - *relevant,*
    - *logically sound.*

Structure.

- Abstract: *Why read this paper.*
- Introduction: *What is in this paper.*
- Methods: *What will/did I do.*
- Results: *What did I find.*
- Conclusions: *What do I think about that.*

Introduction.

- State your general research questions & claims.
- Cite the relevant background literature.
- Describe your work (at a high level).
- State your research hypotheses.
- Draw connections between them.

- Paper Order: *Claims; Citations; Work; Hypotheses.*
- Work Order: *Claims; Hypotheses; Citations; Revise; Work.*

Introduction (Argument)

- Your work is *relevant* to the world.
- Your research question is *novel*, open or unanswered.
- Your hypotheses are *appropriate* for the question.
- Your hypotheses are *testable*.
- Your methodology is *sound*.
- Drawing *analogies* and *distinctions* between your work and the work of others.

"Yes I'll buy that." Reading, & Writing.

Example: Kim et al.

PARENTING AND PRESCHOOLERS' SYMPTOMS AS A FUNCTION OF CHILD GENDER AND SES
Hyun-Jeong Kim; David H. Arnold; Paige H. Fisher; Alexandra Zeljo

- Claims & Questions:
  - *Gender differences in psychopathology patterns are well-documented.*
  - *Research on early childhood might provide important information regarding initial contributions of various factors involved in these gender differences.*
  - *However, existing studies have mostly focused on school-age children and adolescents, with little study of younger children.*

Example: Kim et al.

- Citations:
  - *Boys show more frequent externalizing disorders than girls beginning at a fairly early age; studies suggest that gender differences in externalizing problems emerge sometime during the preschool years, eventually becoming two to three times more common in males than females (Keenan & Shaw, 1997).*
  - *This hypothesis is consistent with previous findings (Arnold et al., 1993; Baumrind, 1966).*
  - *Keenan and Shaw (1997) hypothesized that girls' internalizing behaviors are, in part, socialized through parental, teacher, and peer influences based on female stereotypes.*

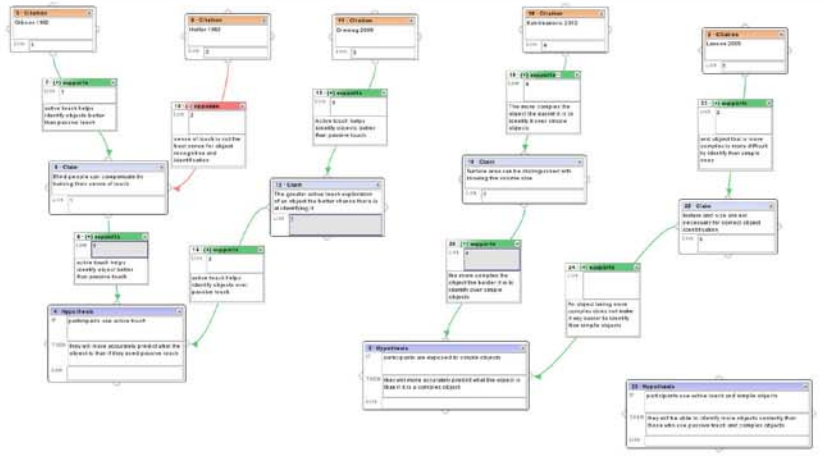
Example: Kim et al.

- Discussion of the study:
  - *In the present study we investigated the relation of lax and overreactive parenting to psychopathology in girls and boys.*
  - *We assume that externalizing behaviors of boys are more congruent with parents' gender stereotypes, so that parents would tend to be lax in response to boys' externalizing behaviors.*

Example: Kim et al.

- Hypotheses:
  - *First, we predicted a relation between externalizing behaviors and lax parenting of boys.*
  - *Second, for similar reasons, we predicted that internalizing behaviors would be related to lax parenting of girls.*

LASAD

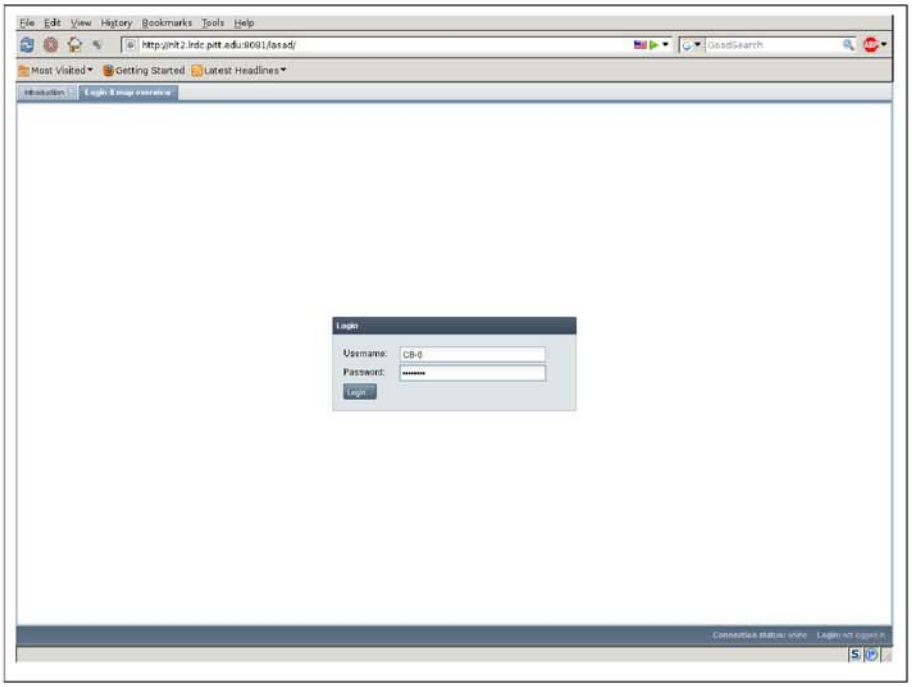


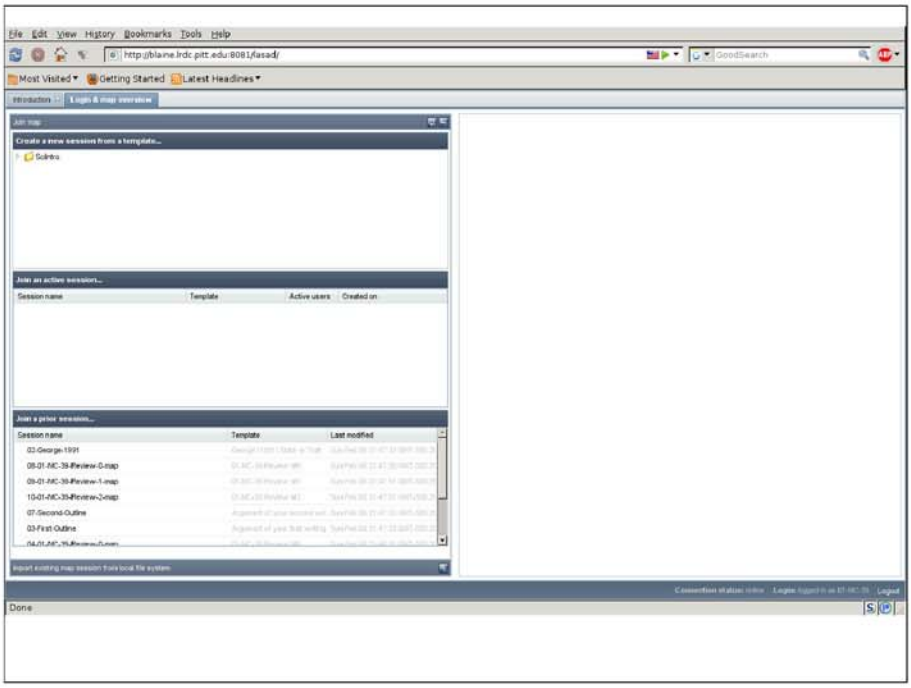
LASAD

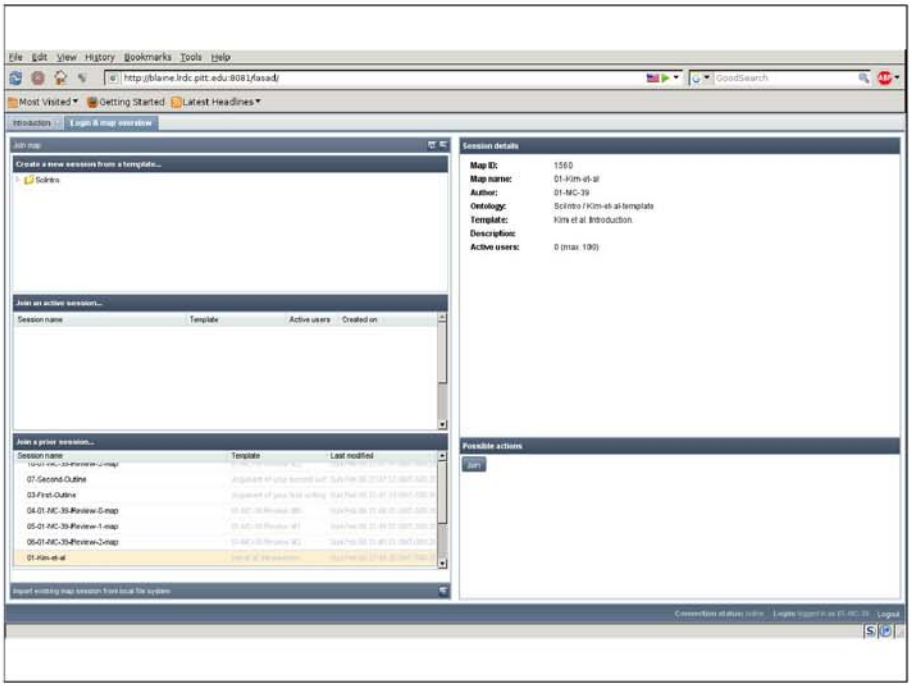
Reading: Read and annotate the argument structure in existing papers.

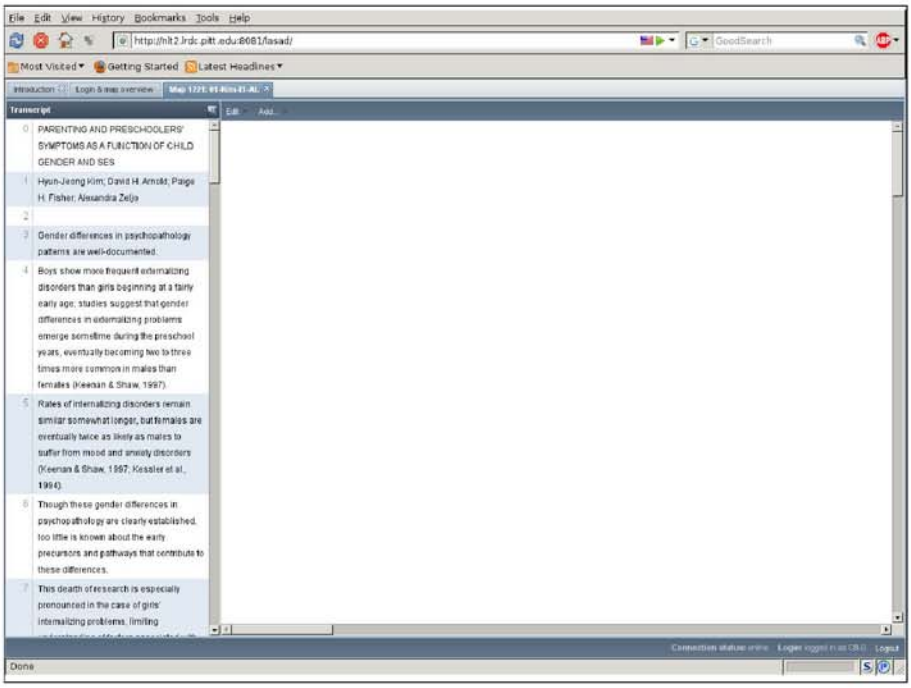
Outlining: Prepare your introductory argument *before* writing it.

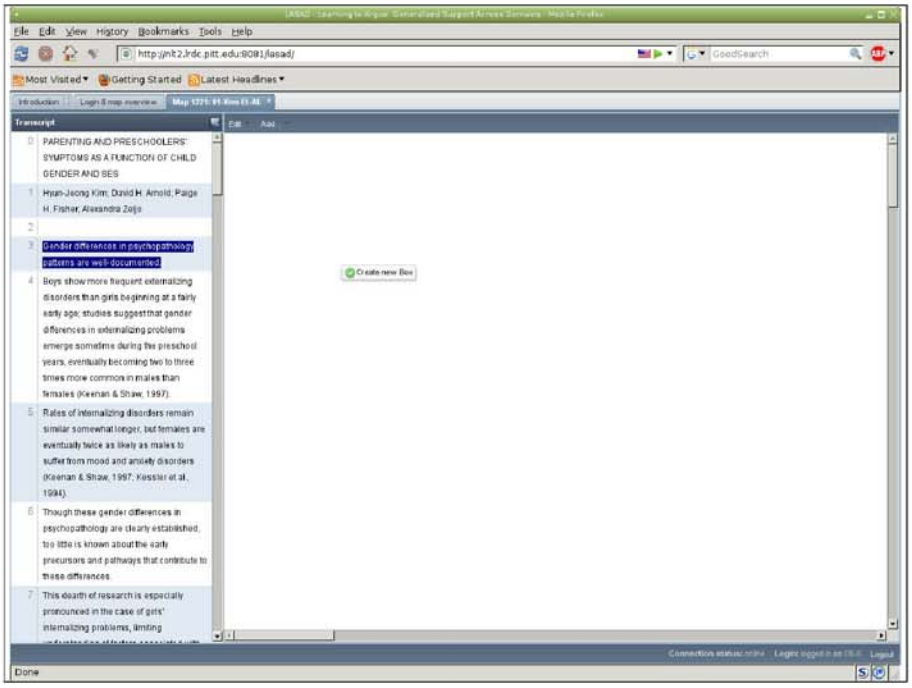
Reviewing: Annotating the argument in your peers' work.

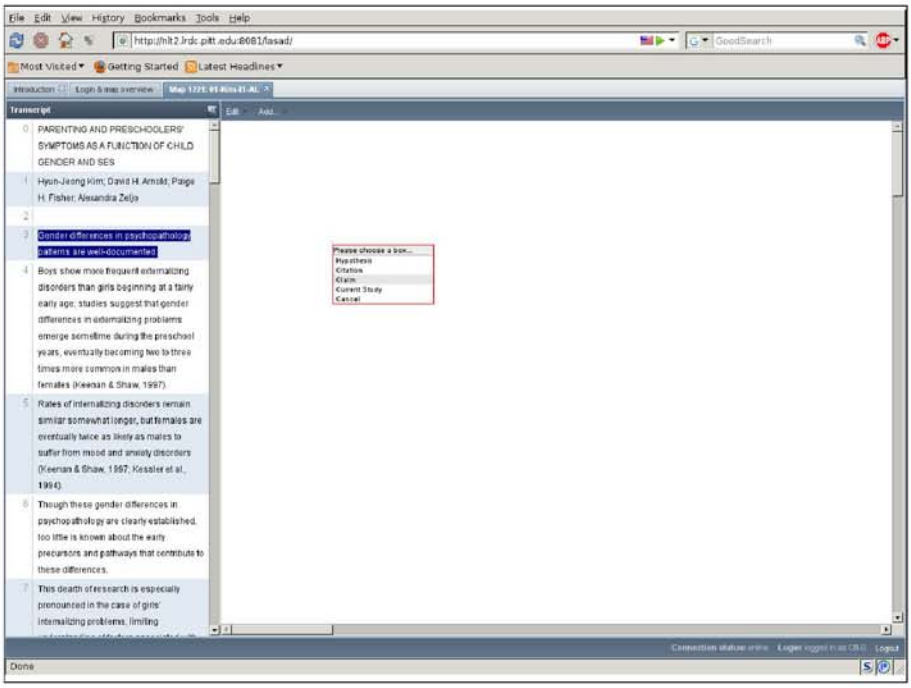


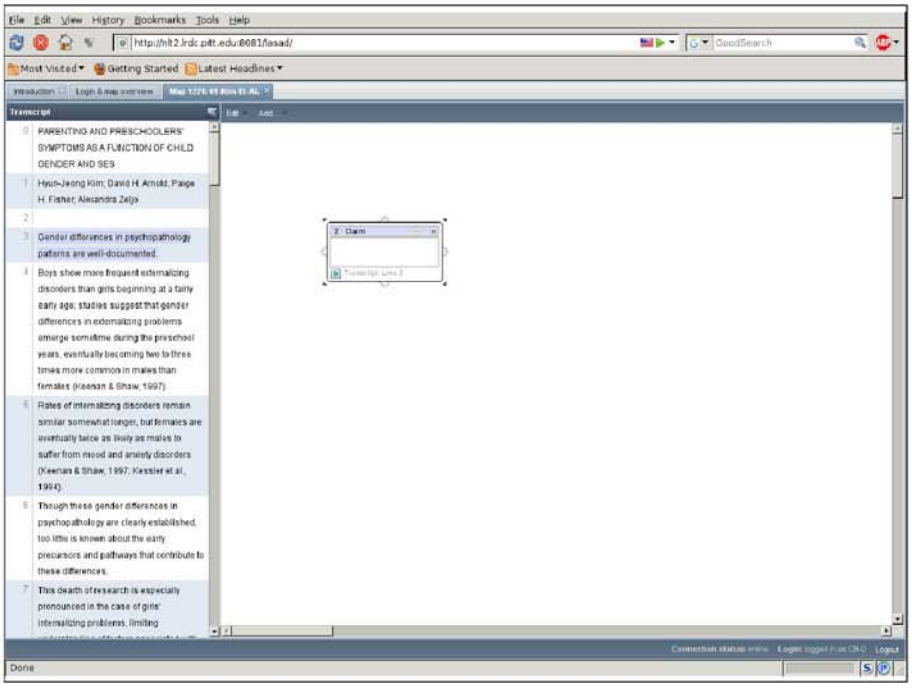

File Edit View History Bookmarks Tools Help
http://blaine.lrdc.pitt.edu:8081/fasad/
Most Visited
Getting Started
Latest Headlines
Introduction
Login & map overview
Join map
Create a new session from a template...
SciIntro
Join an active session...
Session name
Template
Active users
Created on
Join a prior session...
Session name
Template
Last modified
02-George-1991
08-01-NC-39-Review-0-map
09-01-NC-39-Review-1-map
10-01-NC-39-Review-2-map
07-Second-Outline
03-First-Outline
Import existing map session from local file system
Done

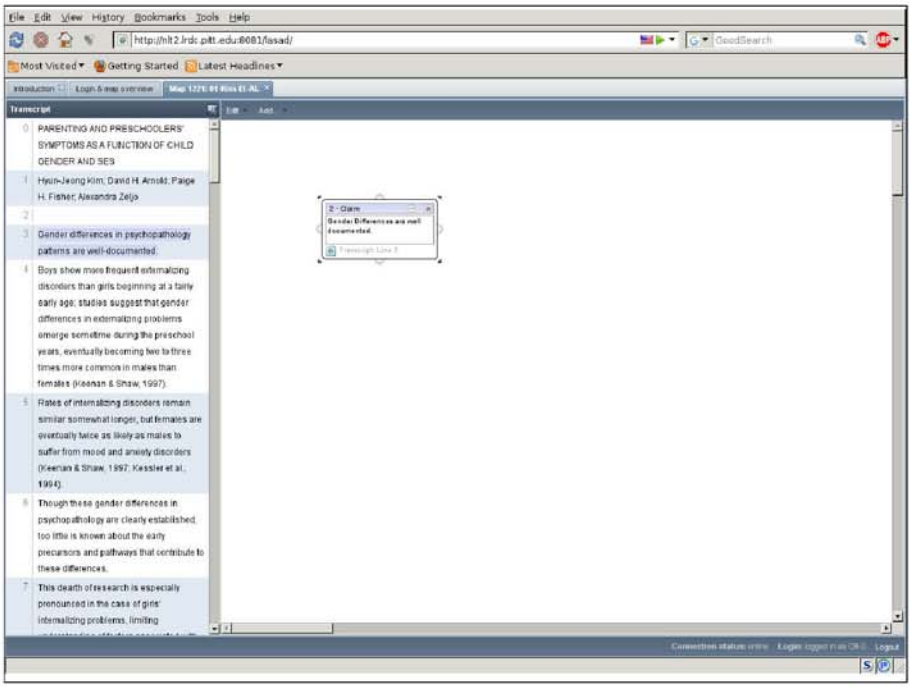

File Edit View History Bookmarks Tools Help
http://blaine.lrdc.pitt.edu:8081/fasad/
Most Visited
Getting Started
Latest Headlines
Introduction
Login & map overview
Join map
Create a new session from a template...
SciIntro
Join an active session...
Session name
Template
Active users
Created on
Join a prior session...
Session name
Template
Last modified
07-Second-Outline
03-First-Outline
04-01-NC-39-Review-0-map
05-01-NC-39-Review-1-map
06-01-NC-39-Review-2-map
01-Kim-et-al
Import existing map session from local file system
Session details
Map ID: 1560
Map name: 01-Kim-et-al
Author: 01-NC-39
Ontology: SciIntro / Kim-et-al-template
Template: Kim et al. Introduction
Description:
Active users: 0 (max. 100)
Possible actions
Join

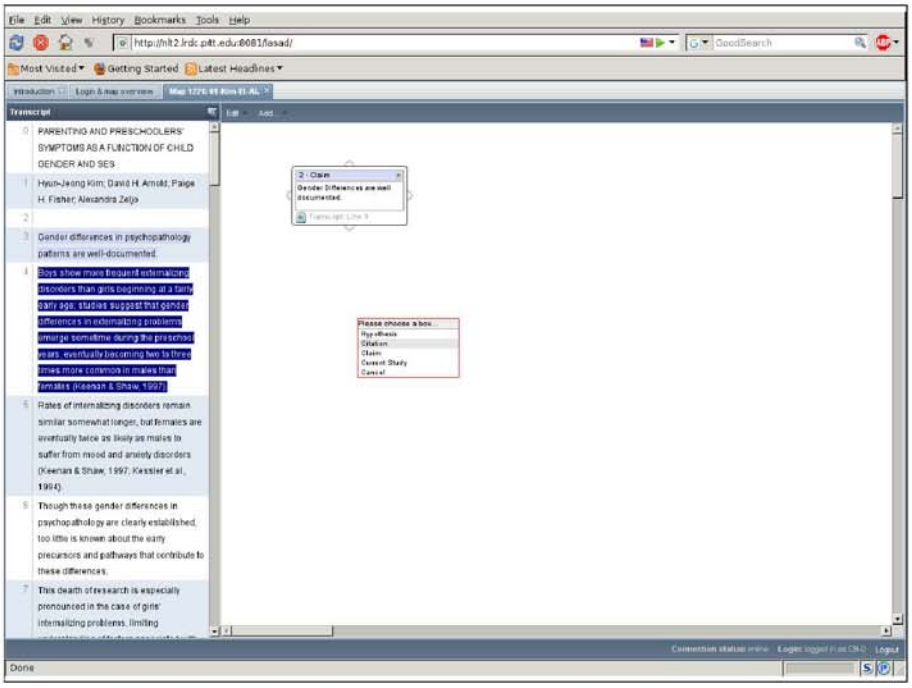

File Edit View History Bookmarks Tools Help
http://nlt2.lrdc.pitt.edu:8081/lasad/
Most Visited
Getting Started
Latest Headlines
Transcript
0 PARENTING AND PRESCHOOLERS' SYMPTOMS AS A FUNCTION OF CHILD GENDER AND SES
1 Hyun-Jeong Kim; David H. Arnold; Paige H. Fisher; Alexandra Zeljo
2
3 Gender differences in psychopathology patterns are well-documented.
4 Boys show more frequent externalizing disorders than girls beginning at a fairly early age; studies suggest that gender differences in externalizing problems emerge sometime during the preschool years, eventually becoming two to three times more common in males than females (Keenan & Shaw, 1997).
5 Rates of internalizing disorders remain similar somewhat longer, but females are eventually twice as likely as males to suffer from mood and anxiety disorders (Keenan & Shaw, 1997; Kessler et al., 1994).
6 Though these gender differences in psychopathology are clearly established, too little is known about the early precursors and pathways that contribute to these differences.
7 This dearth of research is especially pronounced in the case of girls' internalizing problems, limiting
Done

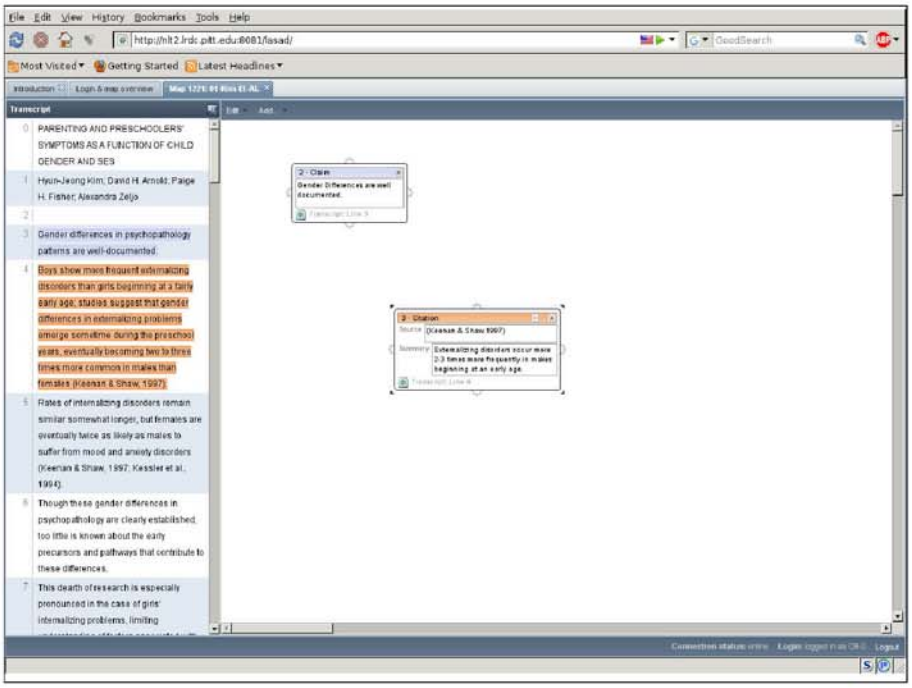

File Edit View History Bookmarks Tools Help
http://nlt2.lrdc.pitt.edu:8081/lasad/
Most Visited
Getting Started
Latest Headlines
Transcript
0 PARENTING AND PRESCHOOLERS' SYMPTOMS AS A FUNCTION OF CHILD GENDER AND SES
1 Hyun-Jeong Kim; David H. Arnold; Paige H. Fisher; Alexandra Zeljo
2
3 Gender differences in psychopathology patterns are well-documented.
Create new Box
4 Boys show more frequent externalizing disorders than girls beginning at a fairly early age; studies suggest that gender differences in externalizing problems emerge sometime during the preschool years, eventually becoming two to three times more common in males than females (Keenan & Shaw, 1997).
5 Rates of internalizing disorders remain similar somewhat longer, but females are eventually twice as likely as males to suffer from mood and anxiety disorders (Keenan & Shaw, 1997; Kessler et al., 1994).
6 Though these gender differences in psychopathology are clearly established, too little is known about the early precursors and pathways that contribute to these differences.
7 This dearth of research is especially pronounced in the case of girls' internalizing problems, limiting
Done

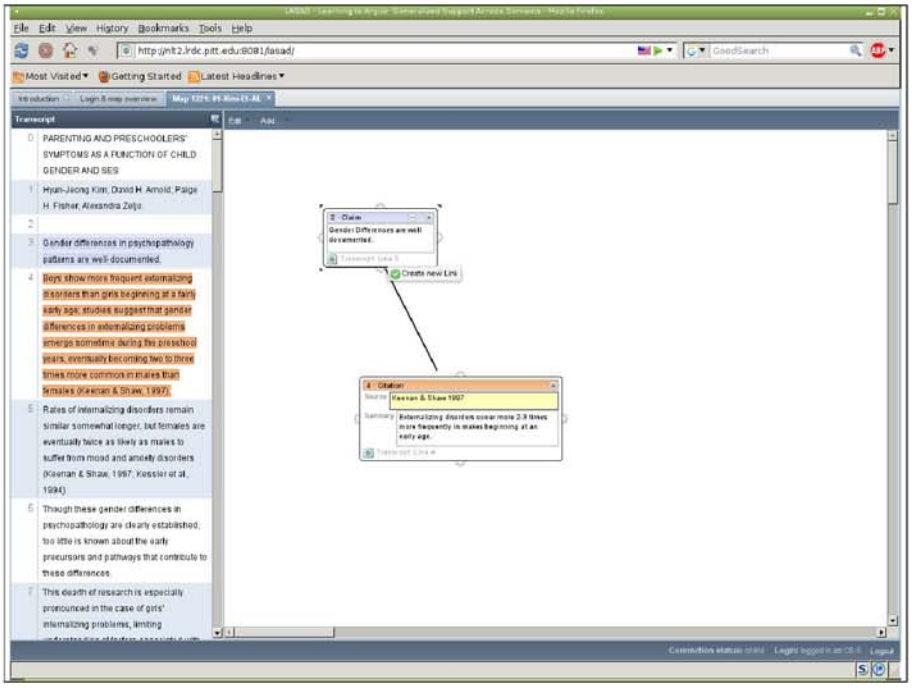

File Edit View History Bookmarks Tools Help
http://nlt2.lrdc.pitt.edu:8081/lasad/
Most Visited Getting Started Latest Headlines
Transcript
PARENTING AND PRESCHOOLERS' SYMPTOMS AS A FUNCTION OF CHILD GENDER AND SES
Hyun-Jeong Kim; David H. Arnold; Paige H. Fisher; Alexandra Zelj
Gender differences in psychopathology patterns are well-documented.
Boys show more frequent externalizing disorders than girls beginning at a fairly early age; studies suggest that gender differences in externalizing problems emerge sometime during the preschool years, eventually becoming two to three times more common in males than females (Keenan & Shaw, 1997).
Rates of internalizing disorders remain similar somewhat longer, but females are eventually twice as likely as males to suffer from mood and anxiety disorders (Keenan & Shaw, 1997; Kessler et al., 1994).
Though these gender differences in psychopathology are clearly established, too little is known about the early precursors and pathways that contribute to these differences.
This dearth of research is especially pronounced in the case of girls' internalizing problems, limiting
Please choose a box...
Hypothesis
Citation
Claim
Current Study
Cancel
Done

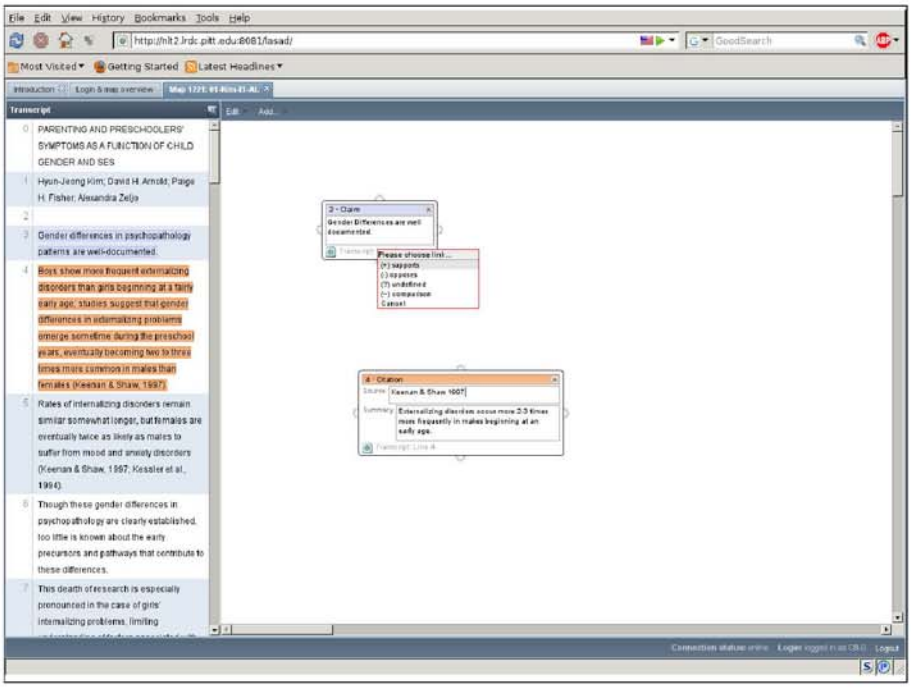

File Edit View History Bookmarks Tools Help
http://nlt2.lrdc.pitt.edu:8081/lasad/
Most Visited Getting Started Latest Headlines
Transcript
PARENTING AND PRESCHOOLERS' SYMPTOMS AS A FUNCTION OF CHILD GENDER AND SES
Hyun-Jeong Kim; David H. Arnold; Paige H. Fisher; Alexandra Zelj
Gender differences in psychopathology patterns are well-documented.
Boys show more frequent externalizing disorders than girls beginning at a fairly early age; studies suggest that gender differences in externalizing problems emerge sometime during the preschool years, eventually becoming two to three times more common in males than females (Keenan & Shaw, 1997).
Rates of internalizing disorders remain similar somewhat longer, but females are eventually twice as likely as males to suffer from mood and anxiety disorders (Keenan & Shaw, 1997; Kessler et al., 1994).
Though these gender differences in psychopathology are clearly established, too little is known about the early precursors and pathways that contribute to these differences.
This dearth of research is especially pronounced in the case of girls' internalizing problems, limiting
2 Claim
Done

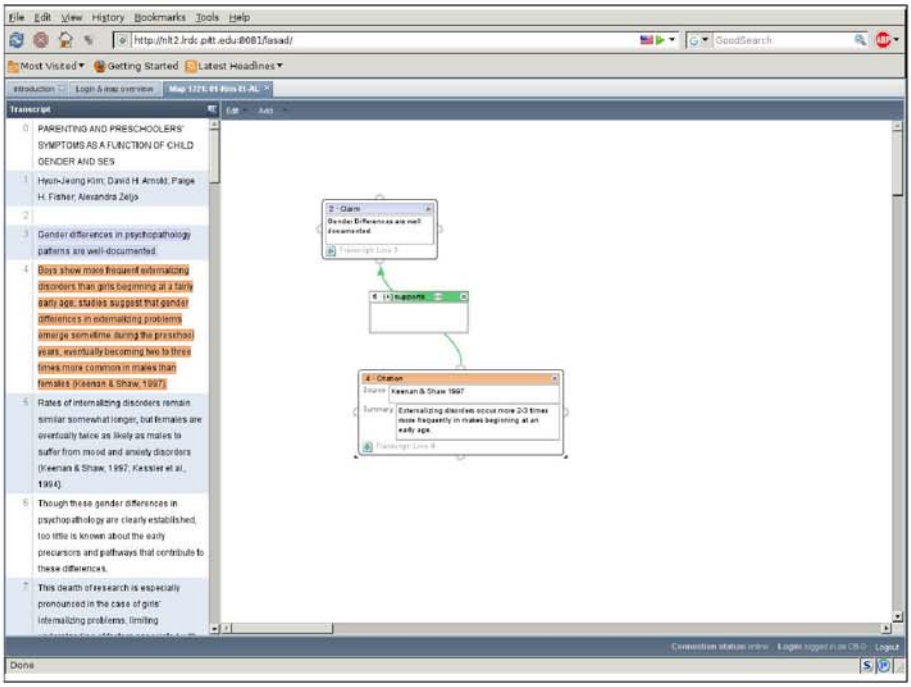

File Edit View History Bookmarks Tools Help
http://nlt2.lrdc.pitt.edu:8081/lasad/
Most Visited Getting Started Latest Headlines
Transcript
0 PARENTING AND PRESCHOOLERS' SYMPTOMS AS A FUNCTION OF CHILD GENDER AND SES
1 Hyun-Jeong Kim; David H. Arnold; Paige H. Fisher; Alexandra Zeljo
2
3 Gender differences in psychopathology patterns are well-documented.
4 Boys show more frequent externalizing disorders than girls beginning at a fairly early age; studies suggest that gender differences in externalizing problems emerge sometime during the preschool years, eventually becoming two to three times more common in males than females (Keenan & Shaw, 1997).
5 Rates of internalizing disorders remain similar somewhat longer, but females are eventually twice as likely as males to suffer from mood and anxiety disorders (Keenan & Shaw, 1997; Kessler et al., 1994).
6 Though these gender differences in psychopathology are clearly established, too little is known about the early precursors and pathways that contribute to these differences.
7 This dearth of research is especially pronounced in the case of girls' internalizing problems, limiting
2 - Claim
Gender Differences are well documented.
Connection status: online Login: logged in as CR-0 Logout

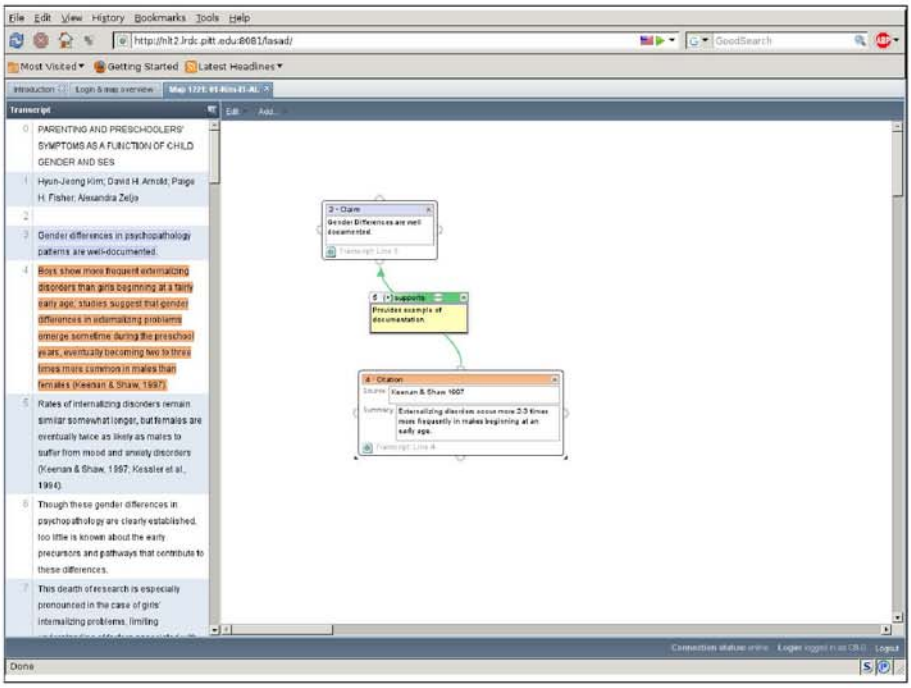

File Edit View History Bookmarks Tools Help
http://nlt2.lrdc.pitt.edu:8081/lasad/
Most Visited Getting Started Latest Headlines
Transcript
0 PARENTING AND PRESCHOOLERS' SYMPTOMS AS A FUNCTION OF CHILD GENDER AND SES
1 Hyun-Jeong Kim; David H. Arnold; Paige H. Fisher; Alexandra Zeljo
2
3 Gender differences in psychopathology patterns are well-documented.
4 Boys show more frequent externalizing disorders than girls beginning at a fairly early age; studies suggest that gender differences in externalizing problems emerge sometime during the preschool years, eventually becoming two to three times more common in males than females (Keenan & Shaw, 1997).
5 Rates of internalizing disorders remain similar somewhat longer, but females are eventually twice as likely as males to suffer from mood and anxiety disorders (Keenan & Shaw, 1997; Kessler et al., 1994).
6 Though these gender differences in psychopathology are clearly established, too little is known about the early precursors and pathways that contribute to these differences.
7 This dearth of research is especially pronounced in the case of girls' internalizing problems, limiting
2 - Claim
Gender Differences are well documented.
Please choose a box...
Hypothesis
Citation
Claim
Current Study
Connection status: online Login: logged in as CR-0 Logout
Done

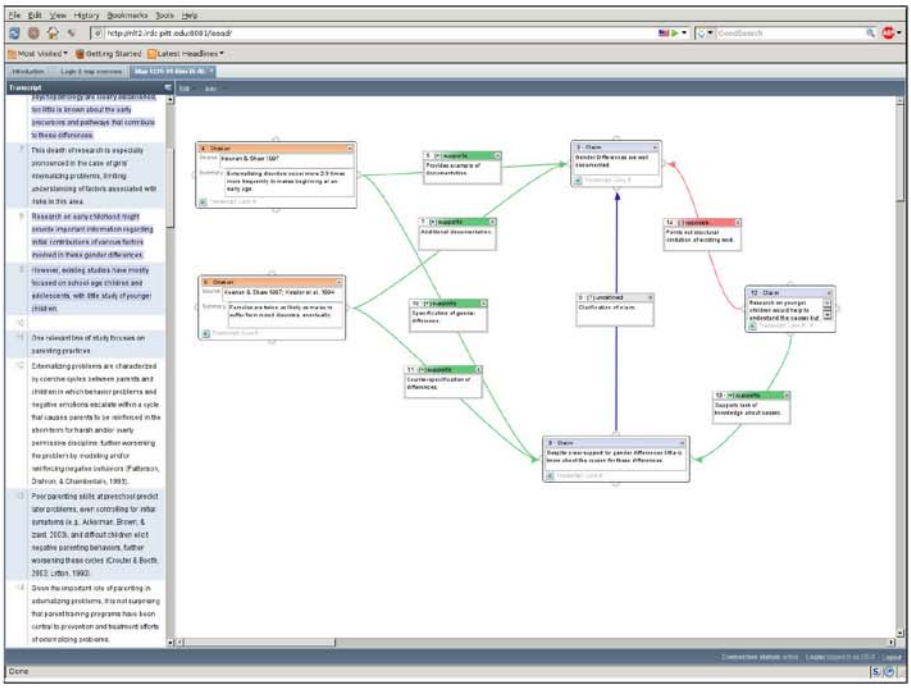

File Edit View History Bookmarks Tools Help
http://nlt2.lrdc.pitt.edu:8081/lasad/
Most Visited
Getting Started
Latest Headlines
Transcript
PARENTING AND PRESCHOOLERS' SYMPTOMS AS A FUNCTION OF CHILD GENDER AND SES
Hyun-Jeong Kim; David H. Arnold; Paige H. Fisher; Alexandra Zeljo
Gender differences in psychopathology patterns are well-documented.
Boys show more frequent externalizing disorders than girls beginning at a fairly early age; studies suggest that gender differences in externalizing problems emerge sometime during the preschool years, eventually becoming two to three times more common in males than females (Keenan & Shaw, 1997).
Rates of internalizing disorders remain similar somewhat longer, but females are eventually twice as likely as males to suffer from mood and anxiety disorders (Keenan & Shaw, 1997; Kessler et al., 1994).
Though these gender differences in psychopathology are clearly established, too little is known about the early precursors and pathways that contribute to these differences.
This dearth of research is especially pronounced in the case of girls' internalizing problems, limiting

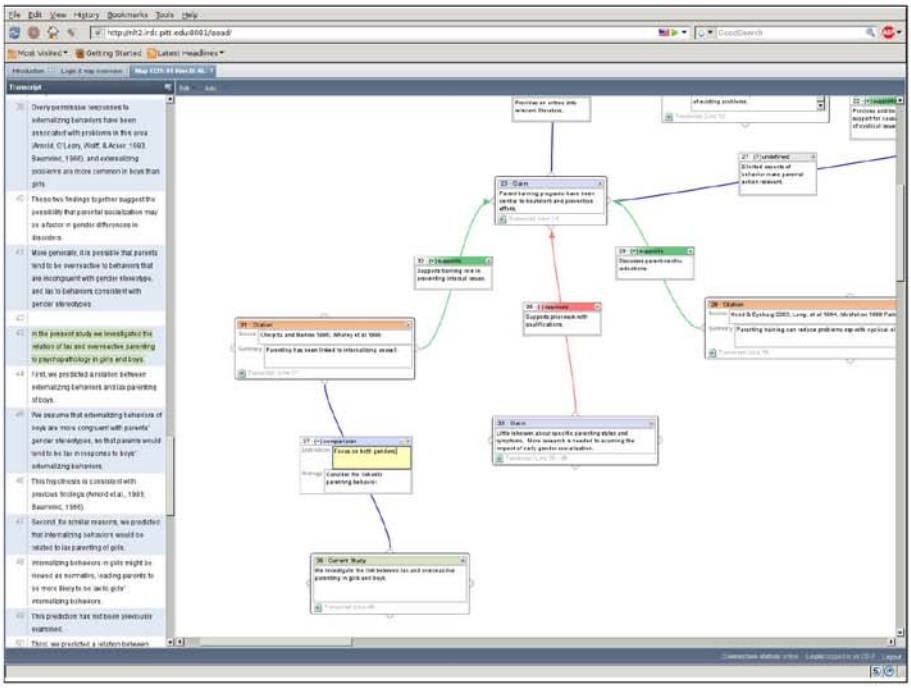

File Edit View History Bookmarks Tools Help
http://nlt2.lrdc.pitt.edu:8081/lasad/
Most Visited
Getting Started
Latest Headlines
Transcript
PARENTING AND PRESCHOOLERS' SYMPTOMS AS A FUNCTION OF CHILD GENDER AND SES
Hyun-Jeong Kim; David H. Arnold; Paige H. Fisher; Alexandra Zeljo
Gender differences in psychopathology patterns are well-documented.
Boys show more frequent externalizing disorders than girls beginning at a fairly early age; studies suggest that gender differences in externalizing problems emerge sometime during the preschool years, eventually becoming two to three times more common in males than females (Keenan & Shaw, 1997).
Rates of internalizing disorders remain similar somewhat longer, but females are eventually twice as likely as males to suffer from mood and anxiety disorders (Keenan & Shaw, 1997; Kessler et al., 1994).
Though these gender differences in psychopathology are clearly established, too little is known about the early precursors and pathways that contribute to these differences.
This dearth of research is especially pronounced in the case of girls' internalizing problems, limiting
Create new Link

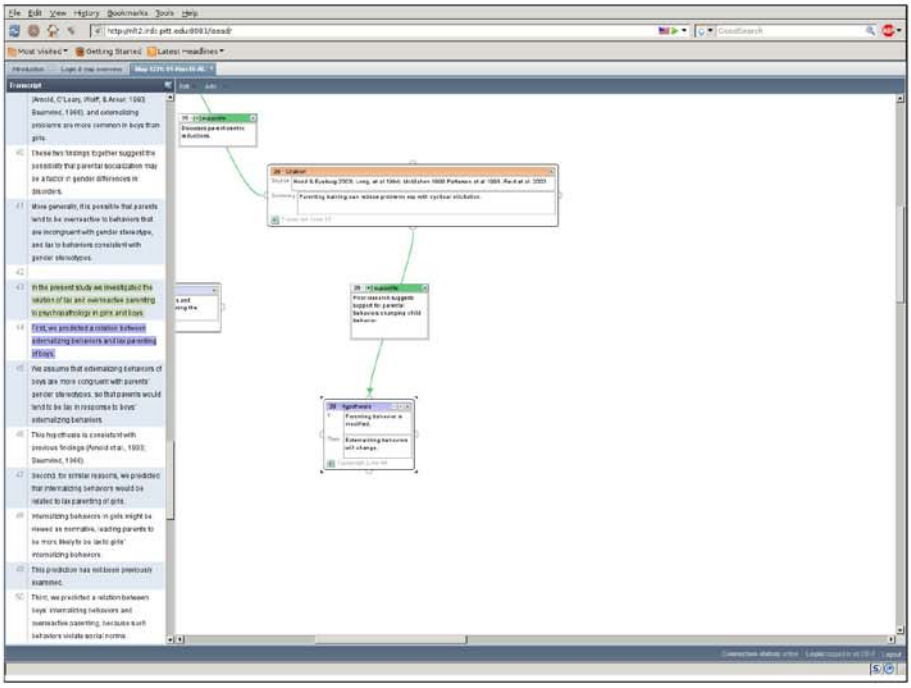

File Edit View History Bookmarks Tools Help
http://nlt2.lrdc.pitt.edu:8081/lasad/
GoodSearch
Most Visited
Getting Started
Latest Headlines
Transcript
PARENTING AND PRESCHOOLERS' SYMPTOMS AS A FUNCTION OF CHILD GENDER AND SES
Hyun-Jeong Kim; David H. Arnold; Paige H. Fisher; Alexandra Zelju
Gender differences in psychopathology patterns are well-documented.
Boys show more frequent externalizing disorders than girls beginning at a fairly early age; studies suggest that gender differences in externalizing problems emerge sometime during the preschool years, eventually becoming two to three times more common in males than females (Keenan & Shaw, 1997).
Rates of internalizing disorders remain similar somewhat longer, but females are eventually twice as likely as males to suffer from mood and anxiety disorders (Keenan & Shaw, 1997; Kessler et al., 1994).
Though these gender differences in psychopathology are clearly established, too little is known about the early precursors and pathways that contribute to these differences.
This dearth of research is especially pronounced in the case of girls' internalizing problems, limiting
2 - Claim
Gender Differences are well documented.
Please choose link ...
(+) supports
(-) opposes
(?) undefined
(~) comparison
Cancel
4 - Citation
Source
Keenan & Shaw 1997
Summary
Externalizing disorders occur more 2-3 times more frequently in males beginning at an early age.
Transcript Line 4
Connection status: online
Logged in as CB-0
Logout


File Edit View History Bookmarks Tools Help
http://nlt2.lrdc.pitt.edu:8081/lasad/
GoodSearch
Most Visited
Getting Started
Latest Headlines
Transcript
PARENTING AND PRESCHOOLERS' SYMPTOMS AS A FUNCTION OF CHILD GENDER AND SES
Hyun-Jeong Kim; David H. Arnold; Paige H. Fisher; Alexandra Zelju
Gender differences in psychopathology patterns are well-documented.
Boys show more frequent externalizing disorders than girls beginning at a fairly early age; studies suggest that gender differences in externalizing problems emerge sometime during the preschool years, eventually becoming two to three times more common in males than females (Keenan & Shaw, 1997).
Rates of internalizing disorders remain similar somewhat longer, but females are eventually twice as likely as males to suffer from mood and anxiety disorders (Keenan & Shaw, 1997; Kessler et al., 1994).
Though these gender differences in psychopathology are clearly established, too little is known about the early precursors and pathways that contribute to these differences.
This dearth of research is especially pronounced in the case of girls' internalizing problems, limiting
2 - Claim
Gender Differences are well documented.
4 - (+) supports
4 - Citation
Source
Keenan & Shaw 1997
Summary
Externalizing disorders occur more 2-3 times more frequently in males beginning at an early age.
Transcript Line 4
Connection status: online
Logged in as CB-0
Logout
Done


File Edit View History Bookmarks Tools Help
Most Visited Getting Started Latest Headlines
Transcript
PARENTING AND PRESCHOOLERS' SYMPTOMS AS A FUNCTION OF CHILD GENDER AND SES
Hyun-Jeong Kim; David M. Arnold; Paige H. Fisher; Alexandra Zelio
Gender differences in psychopathology patterns are well-documented.
Boys show more frequent externalizing disorders than girls beginning at a fairly early age; studies suggest that gender differences in externalizing problems emerge sometime during the preschool years, eventually becoming two to three times more common in males than females (Keenan & Shaw, 1997).
Rates of internalizing disorders remain similar somewhat longer, but females are eventually twice as likely as males to suffer from mood and anxiety disorders (Keenan & Shaw, 1997; Kessler et al., 1994).
Though these gender differences in psychopathology are clearly established, too little is known about the early precursors and pathways that contribute to these differences.
This dearth of research is especially pronounced in the case of girls' internalizing problems, limiting
Done


File Edit View History Bookmarks Tools Help
Transcript
Done





APPENDIX B

CLASS ASSIGNMENT

This chapter contains a representative copy of the class assignment provided to the students. This focuss solely on the essay criteria and includes additional information above and beyond the contents of the introduction.

Assignment 7: Observational Study Paper
General Overview of the Assignment

ASSIGNMENT 7 DUE:

Draft = 20 points

LENGTH: Approximately 8-10 pages

MUST INCLUDE (in this order):

Title Page
Abstract
Introduction
Method
Results
Discussion
References
Table
Figure Caption & Figure

OTHER REQUIREMENTS:

- Your paper must have at least **five (5)** references from PsycINFO. *You may use instructor-provided references for two (2) of your references, but you will need to find three (3) more references on your own. You need to turn in complete copies of the articles or chapters that you use for references with your paper. If you use a book you just need to turn in the relevant pages (e.g., a copy of a chapter) and the title page.*

- *Discuss at least one study or theoretical position that conflicts with a hypothesis. Explain why some prior work supports and other work opposes this hypothesis.*

- You should **NOT** wait until the entire study has been completed to begin writing — you should be working on sections of the paper throughout the course of this project. Your lab instructor will provide guidelines for submitting drafts.

- You must submit an electronic version of your paper to SWoRD.

- Late papers will not be accepted.

Grading Rubric

Element of Paper		*Max. Possible Points*	*Points Earned*
Abstract: All required information included? 150 words or less; concise, specific, and accurate? Appropriate level of detail?		**1**	
Introduction: Central topic introduced and background information provided? Brief high-level overview of study design and clear statement of hypotheses? Appropriate integration of conflicting research findings into a convincing argument for at least one hypothesis?		**4**	
Method: Participants adequately and accurately described? Procedures presented accurately and clearly so study can be replicated? Appropriate level of detail that excludes inconsequential details		**2**	
Results: Descriptive statistics reported either in text or table/figure? Statistical tests reported completely and accurately? Tables/figures correctly referenced in text? Results worded so they're clearly linked to hypotheses/research questions?		**2**	
Discussion: Main findings summarized? Results clearly and accurately interpreted? Current study put into context in relation to previous work? Strengths/weaknesses, alternative explanations, implications, suggestions for future research discussed as needed?		**4**	
Global Writing: Writing clear and concise, not wordy or confusing? Ideas well organized, part of a coherent argument, flow together well? Tone appropriate for readership of professional psych journal?		**3**	
Technical Writing: Sentences complete and grammatically correct? Paper carefully proof-read and spell-checked?		**2**	
APA Style: Is APA style used correctly for the following? Numbers Statistics In-text citations Paper header Abbreviations Section headings Etc.	Are the following elements formatted according to APA style? Title page Abstract Introduction Method Results Discussion References Table/Figure	**2**	
Total For Draft		**20**	

Grading Criteria

18 - 20 Points: The paper demonstrates a sophisticated and insightful understanding of the assignment. The content of all sections of the paper is complete and accurate. The writing is clear, well organized, and grammatical, and the tone of the language is appropriate for the given audience. It has been carefully spell-checked and there are few if any typographical errors. Careful attention has been paid to the use of APA style throughout the paper. Overall, an outstanding effort.

16-17 Points: The paper demonstrates a clear understanding of the assignment. The content of the sections of the paper for the most part is complete and accurate. The writing is usually well organized and grammatical. It has been spell-checked and there are few typographical errors. Attention has been paid to the use of APA style throughout the paper, although there are occasional errors. Overall, a solid, above-average effort.

14-15 Points: The paper demonstrates a basic understanding of the assignment. The content of several sections of the paper is incomplete and inaccurate. The writing is somewhat organized and grammatical. It has been spell-checked but typographical errors are relatively common. Some attention has been paid to the use of APA style throughout the paper, but errors are frequent. Overall, an adequate but unspectacular effort.

12-13 Points: The paper demonstrates only a superficial understanding of the assignment. The content of several sections of the paper are seriously incomplete and inaccurate; sections of the paper may be missing entirely. The writing is disorganized and ungrammatical. It has not been spell-checked. Typographical errors are so common as to interfere with a basic understanding the writing. Little attention has been paid to the use of APA style. Overall, a substandard effort.

Less than 12 Points: The paper demonstrates a complete misunderstanding of the assignment AND/OR the author clearly did not spend enough time on the paper. Significant portions of the paper are missing. The writing lacks any kind of organization. No attempt was made to proofread or spell-check the paper. No attention has been paid to APA style. An unsatisfactory effort.

TENTATIVE SCHEDULE:

Note that this order may change slightly depending on your lab's work pace, such as how long your data collection takes.

1. Begin a literature search on your class' topic (Assignment 2)
2. Finalize a general hypothesis and your study design.
3. Narrow literature search, *create an argument diagram* justifying your hypotheses (Assignment 4)
4. Gather data, write your **Method** section
5. *Review the argument diagrams* submitted by three peers (Assignment 5)
6. Provide back-evaluations about the helpfulness of your peers' reviews of your diagram and then *revise your argument diagram* based on their feedback (Assignment 6).
7. Conduct data analysis, discuss results of study, and create **Tables** and **Figures**.
8. Write **Title Page, Introduction**, **Results**, **Discussion**, and **Abstract** (Assignment 7)
9. Review the papers written by three peers (Assignment 8)
10. Provide back-evaluations about the helpfulness of your peers' reviews of your paper and then Revise your paper based on the feedback provided by your peers (Assignment 9)

Instructions for using SWoRD to submit your first draft

Save your paper when you are finished

1) When you are done with your paper, name the file so it can be uploaded. Also make sure you know where your file is saved and that is has the appropriate name.
2) The name of the file should start with the last name of your TF, a hyphen, the code "P1D1" without the quotation marks, followed by your Peoplesoft number as in: **Jones-P1D1-9999999.doc**
3) Note: .doc, .docx, txt, and .rtf are all adequate file types

Use SWoRD to submit your diagram.

1) Go to the main SWoRD web page**:** http://sword.lrdc.pitt.edu/sword/
2) Login using the password you created. If you forgot your password, click on the "Login Problems" link to reset your password.
3) From your account Home page click on the Course Name and this will take you to the Assignment List page. In the My Submission column click on Upload for the "First Paper Draft One."
4) Enter a Paper Short Name and click on the Browse button to choose the appropriate file. Click the Upload File button. The Assignment List page will come up showing that your paper was successfully submitted.
5) A warning about the **filename** and **Paper Short Name**: Please make sure that your **Paper Short Name and the file name** do not contain any special characters. SWoRD will not recognize them and you'll get an error when you try to submit your diagram.

APPENDIX C

GRADING RUBRIC

The grading Rubrics contain a total of 14 questions each covering specific features of the argument and argumentative skills, such as stating a research question and distinguishing cited works, and the overall coherence and quality. Each question in the rubric has a number of the form $(E|G).*$ which denotes the rubric it is in (either Essay or Graph, and its index. It also has a long name and a short descriptive name (e.g. *Arg-Coherent*). Because the rubrics are parallel the paired questions are isomorphic. Thus they share the same short name. When referring to individual questions then I will use both their qualified number (e.g. *E.13*) and their short name (e.g. *Arg-Convincing*). When referring to pairs of questions I will use a combined name (e.g. *G/E.11 (Study-Novel)*).

All questions were graded using the web-based SNG grading tool and contained three values:

Value The score which was a scale graded on a range of (*-2 – 2*) in $\frac{1}{2}$ point steps with the optional score of *N/A* for some questions. For the summative questions *E.14* and *G.14 (Arg-Quality)* the score range was (*-5 – 5*) again in $\frac{1}{2}$ point steps.

Free Text For some questions the grader was asked to supply a free-text answer specifying possible fixes or explaining their score.

Selections For questions with an identification component (e.g. pick the hypothesis) the grader was asked to select the relevant text or note the individual node and arc IDs.

The question pairs are listed below. The *E.01 (RQ-Quality) – 14 (Arg-Quality)* questions

were applied to the essays while the *G.01 (RQ-Quality) – 14 (Arg-Convincing)* questions were applied to the diagrams. For both rubrics the questions can be divided into two classes. Questions *G/E.01 (RQ-Quality) – G/E.11 (Study-Novel)* are *feature questions* that focus on key features of the argument such as the framing of the hypothesis or the support provided to the research question. Questions *G/E.12 (Arg-Coherent)*, *G/E.13 (Arg-Convincing)*, and *G/E.14 (Arg-Quality)* by contrast, are *gestalt questions* that ask about the argument as a whole.

E.01/G.01: Research Question Quality (*RQ-Quality*)

Range : -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.01 : Did the author clearly state a central research question early in the essay? Please highlight any research questions found in the text. Could the question be improved in any way? If so please describe it below.

(-2) No research question was stated.

(0) A research question was stated but not clearly.

(2) The author presents a clearly-framed research question at the start of the essay.

G.01 : Did the author include a central research question in the diagram represented by a root claim node with its contents framed as a question to which other subquestions or claims are connected? Can the question be improved in any way? Please add the node number(s) of the research question nodes below and describe any potential corrections or improvements.

Please cite the node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(-2) No claim node framed as a research question was included.

(0) A claim node exists that is framed as a question but the question, and its relation to the rest of the diagram, is not clear. Or the author has included single central node that is being used in lieu of a research question (such as a central Current Study node) but it is not a claim node or is not framed as a question.

(2) The author includes a single root claim node clearly framed as a research question.

E.02/G.02: Research Question Link (*RQ-Link*)

Range : N/A -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.02 : Was the research question, if stated, clearly relevant to, and integrated into the essay? That is, did the author reference it, link subsequent claims and hypotheses to it, or reference its content in the remainder of the essay.

(N/A) No research question stated.

(-2) The research question was disconnected from, irrelevant to, or ignored in the remainder of the argument.

(0) While the research question was relevant to the rest of the argument the author did not clearly connect their question to the hypotheses or other parts of the essays or reference it.

(2) The author stated a clear research question that is relevant to their argument and draws clear rhetorical and thematic links between the research question and their subsequent claims and hypotheses.

G.02 : Was the research question, if included, clearly relevant to, connected to the argument diagram? That is, did the author draw links between the research question node and other claims or subclaims in the diagram? Were clear reasons provided for those links on the arcs themselves? And was the content of the question relevant to the rest of the argument diagram?

Please cite the node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(N/A) No research question node was stated.

(-2) The research question node was disconnected from or wholly irrelevant to the remainder of the argument.

(0) While the research question was linked to other parts of the argument diagram the author did not provide reasons for the links or did not make clear why the question was relevant to the remainder of the argument.

(2) The author stated a clear and relevant research question, drew links between that question node and other important subclaims, and provided clear reasons for the relationships.

E.03/G.03: Research Question Support (*RQ-Support*)

Range : -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.03 : Did the author cite relevant background literature related to the research question? Is there any literature that is missing that should be included? If so please describe it below.

(-2) No literature is cited, the literature is unrelated, or the citations make it appear to be unrelated.

(0) The cited literature appears to be relevant *based upon the content of the essay alone* but the citation text does not adequately explain why the author *believes* it is relevant.

(2) The cited literature is clearly relevant and the citation text makes the relevance explicit.

G.03 : Did the author include citation nodes that describe relevant background literature related to the research question based upon the *content* of the citation nodes? Is there any literature that is missing that should be included? Or should any of the summaries be re-framed? If so please describe it below citing the specific node numbers of interest. Please cite the node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(-2) No citation nodes are included in the diagram, or the citations appear to be unrelated based upon their descriptions.

(0) The cited literature appears to be relevant *based upon the content of the diagram* but the citation summaries do not adequately explain why the author *believes* it is relevant.

(2) The cited literature is clearly relevant based upon the summaries presented in the citation nodes and the summary text makes this relevance explicit.

E.04/G.04: Testable Hypothesis (*Hyp-Testable*)

Range : -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.04 : Did the author articulate a testable hypothesis or hypotheses? Please highlight any hypothesis found in the text. Should the hypotheses be re-framed in any way? If so please state any corrections below.

(-2) No hypotheses are stated, or the stated hypotheses are not articulated in a way that can be tested.

(0) It is unclear whether the stated hypotheses are testable.

(2) The stated hypotheses are clear, logical, and can be tested experimentally.

G.04 : Did the author include one or more hypothesis nodes in his or her diagram that are framed as testable hypotheses? Should the hypotheses be re-framed in any way? If so please describe how and cite specific node numbers for the hypotheses of interest.

Please cite the node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(-2) No hypothesis nodes are included or the hypothesis nodes included are not articulated in a testable way.

(0) It is unclear whether the stated hypothesis node(s) are testable.

(2) The hypothesis node(s) included are clear, logical, and can be tested experimentally.

E.05/G.05: Hyp Question Link (*Hyp-Link*)

Range : N/A -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.05 : Were the hypotheses relevant to and clearly connected to the research question? If the connections or relevance could be improved please state how.

(N/A) No research question or hypotheses are presented.

(-2) The author did not explain the connection between the research question and hypotheses in the essay.

(0) While the author linked the research question and hypotheses in the text, he or she does not make the basis for the connection clear.

(2) The research questions and hypotheses are clearly stated and the author explains in the text how they connect to one another.

G.05 : Were the hypothesis node(s) presented relevant to and connected to the research question node via a path in the diagram? If any improvements could be made either by the addition of new nodes or by a re-framing of the content please describe it below.

Please cite the node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(N/A) No research question or hypothesis nodes are included in the diagram.

(-2) The author did not draw any paths from the hypothesis nodes to the question node in the diagram or the existing paths point the wrong way from the question node to the hypothesis nodes.

(0) While the author drew a *path* from the research questions and the hypothesis nodes in the diagram, he or she did not provide an warrant for each arc in the path or the warrants were not stated clearly.

(2) The author drew *arcs* that connected *directly* from the hypothesis nodes to the research question node and provided a clear warrant for each arc in the diagram.

E.06/G.06: Cite Conclusions Stated (*Cite-Conclusions*)

Range : N/A -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.06 : When presenting citations did the author explicitly state the conclusions that he or she drew from the citation and how they related to the present work? Could any of the citations be improved by a restatement of the conclusions? If so please state them below.

(N/A) No prior work is cited.

(-2) The author did not state the conclusions that he or she drew from the cited work nor did he or she state how they related to the present work.

(0) While the author states the conclusions that he or she drew from the cited work the author did not do so clearly nor did the author make the relationship between the cited work and the present work explicit.

(2) The author explicitly states the conclusions that he or she drew from the cited work and explicitly states how they relate to the present work.

G.06 : When presenting citation nodes did the author state the conclusions that he or she drew from the cited work in the summary field and make clear how the cited work related to the present work? Could the statements be improved in any way? If so please discuss them or make other comments with appropriate node numbers below.

Please cite the node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(N/A) No citation nodes are included in the diagram.

(-2) The author did not state the conclusions that they drew from the cited work in the summary field of the citation node, nor did he or she explain the relationship between the cited work and the present work on a path *from* the citation node to the other nodes in the diagram.

(0) While the author stated the conclusions that he or she drew from the cited work in each citation node and drew a path between the citation nodes and other parts of their diagram their warrants for the relationships in the path were not stated clearly.

(2) The author explicitly stated the conclusions that he or she drew from the work cited in the citation nodes. He or she also made clear how the cited work relates to the present work by drawing arcs directly from the citation nodes to other parts of the diagram and providing explicit warrants for the relationships.

E.07/G.07: Cite Reasons Stated (*Cite-Reasons*)

Range : N/A -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.07 : Did the author clearly summarize the relevant content of and conclusions drawn in any cited work? Did they clearly state the reasons why the cited work supports, opposes, or informs upon their claims? And did they clearly distinguish between the act of summarizing existing work and advancing their own arguments?

(N/A) No prior work cited.

(-2) The author did not summarize prior work; did not clearly state how the cited work supports, opposes, or informs upon their work; nor did they distinguish between the reporting and editorializing tasks.

(0) The author summarizes cited works inconsistently; does not consistently state how or why the work relates to their own argument; or does not clearly distinguish between summary and argumentation.

(2) The author clearly summarizes the relevant content of each cited work; clearly explains the works' relationship to their own arguments; and draws a clear distinction between presenting summaries of others' arguments and articulating their own.

G.07 : Did the author present clear summaries of the cited work in their citation nodes? Did they clearly state the reasons why the cited work supports, opposes, or informs upon the neighboring nodes on the arcs themselves? And is there a clear difference between the content of the summaries and the content of the argumentative arcs?

Please cite the node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(N/A) No citation nodes used.

(-2) The author did not include any summaries on the citation nodes; did not link the citation nodes to the rest of the argument or did not include reasons for the argumentative relations on the arcs; or did not clearly distinguish between the two.

(0) The author is inconsistent in summarizing the cited works; does not consistently relate the cited works to other nodes in the argument diagram; does not consistently give reasons for the relationships; or does not consistently differentiate between summarizing others work and explaining how that work supports, opposes, or informs upon their arguments.

(2) The author present clear consistent summaries of the relevant features of each citation in the summary field; they consistently connect the cited works to the rest of the argument diagram and include reasons for the argumentative relations on each arc. Moreover the summaries and reasons and clearly distinct from one-another.

E.08/G.08: Claims Supported (*Claim-Support*)

Range : N/A -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.08 : Did the author adequately support his or her claims or subclaims via relevant citations? Should any citations be added, should any of the included citations be re-framed, or should the author add any text to make the supporting relationships clear? If so please explain below.

(N/A) The author cited no prior work or did not articulate any research questions or claims.

(-2) The author drew no connections between the claims or subclaims and the citations.

(0) While the author drew connections between the citations and claims he or she did not state clearly whether the cited work supports the claims or why he or she thinks that it would support them.

(2) The author presents his or her claims and supporting citations clearly and states explicitly how and why the citations support the claims.

G.08 : Did the author adequately support his or her claim nodes with relevant citation nodes by drawing supporting paths *from* citation nodes to the claim or subclaim nodes and providing warrants for the supporting arcs? For each claim node there should be a path with supporting arcs leading *from* some citation node to the claim node. One citation node may support multiple claims. Should any citations or relations be added or should any of the included citations be re-framed to make the supporting relationships clear? Should any supporting arcs be added or re-framed to make the supporting relationships explicit? Please explain below.

Please cite the node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(N/A) The author included no citation nodes or claim nodes in his or her diagram.

(-2) The diagram contains no supporting paths *from* citation nodes to the claim nodes.

(0) While supporting arcs or supporting paths exist *from* citation nodes to the claim nodes the author does not provide an explanation for the supporting relations.

(2) The author drew supporting paths or arcs from the citation nodes to each of the claim nodes and provides an explanation for the supporting relationship on each intervening arc stating clearly how and why each source in a citation node supports the target.

E.09/G.09: Research Question Open (*RQ-Open*)

Range : N/A -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.09 : Did the author defend the open nature of his or her research question by citing similar work that disagrees about the question or comes to opposing conclusions regarding it? That is, did the author cite work that both supports the research question or subquestions, and work that opposes the question or subquestions? Please highlight any citations that disagree in the text and describe that disagreement below.

(N/A) No prior work is cited or no research question is articulated.

(-2) No work is cited that disagrees about any part of the author's argument.

(0) The author cites work that appears to disagree regarding the central research question or subquestions based upon the authors' statements or the content of the citations *as summarized by the author*, but does not explain the implications of this disagreement and conclusions that he or she draws from the cited work, nor does he or she draw clear analogies and distinctions between the different citations to explain the apparent differences.

(2) The author cites two or more pieces of work that *clearly and explicitly* disagree about the author's central research question or subquestions and clearly identifies and addresses the implications of this disagreement for the author's proposed study.

G.09 : Did the author defend the open nature of his or her research question by including citation nodes that disagree about the research question or come to opposing conclusions regarding the question or subquestions of it? That is, did he or she include at least one citation node that is connected *to* the research question node or a subquestion node via a supporting path, and one that is connected *to* the same question node via an opposing path? Please describe the disagreements that you noted below with specific node citations.

Please cite the node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(N/A) No citation nodes are included or no research question is included in the diagram.

(-2) No pair of citation nodes are presented that disagree about any part of the author's argument.

(0) The author includes citation nodes that disagree about his or her central research question or the subquestion nodes. However the author does not include a clear explanation of the supporting and opposing relationships by providing clear warrants on the arcs in the paths, nor do they attempt to explain the disagreement by drawing a comparison arc between the disagreeing nodes.

(2) The author includes two or more citation nodes that disagree about the author's central research question or subquestions. The citation nodes are also directly connected to the question node by including clear warrants on each of the supporting and opposing arcs. And, the author clearly explains the disagreement by drawing a comparison arc between the citation nodes with clearly specified analogies and or distinctions noted on it.

E.10/G.10: Hypothesis Open (*Hyp-Open*)

Range : N/A -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.10 : Did the author argue that his or her hypotheses are open questions by citing work that disagrees about the hypotheses or draws differing conclusions about them? Please highlight any citations that disagree in the text and describe that disagreement below.

(N/A) No prior work is cited or no hypotheses are articulated.

(-2) No work is cited that disagrees about the author's hypotheses or the author fails to make clear the relationship between the cited work and the hypotheses.

(0) The author cites work that appears to disagree about the validity of the stated hypothesis or hypotheses but the connection between the cited work and the hypothesis or hypotheses is not clear or adequately explained.

(2) The author cites work that clearly disagrees about the validity of the stated hypothesis or hypotheses and explains the implications of the conflicting citations.

G.10 : Did the author argue that his or her hypothesis nodes are open questions by: including citation nodes that disagree about them; drawing conflicting paths from the citation nodes to the hypothesis nodes with clear warrants on the arcs; and drawing a comparison arc between the citation nodes to explain the disagreement? Please describe any disagreements below with specific node citations.

Please cite the node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(N/A) No citation nodes are included in the diagram, or no hypothesis nodes are included.

(-2) No pair of citation nodes are presented that disagree about any part of the author's argument.

(0) The author includes citation nodes that disagree about his or her hypothesis node(s). However the author does not clearly explain the supporting or opposing relations by providing a warrant on the arcs, nor do they attempt to explain the disagreement by drawing a comparison arc between the disagreeing nodes.

(2) The author includes two or more citation nodes that disagree about the author's hypothesis node(s). The citation nodes are also directly connected to the hypothesis node(s) via clearly-explained supporting and opposing arcs. And, the author clearly explains the disagreement by drawing a comparison arc between the citation nodes with clearly specified analogies and or distinctions noted on it.

E.11/G.11: Current Study Novel (*Study-Novel*)

Range : N/A -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.11 : Did the author show the novelty of his or her proposed study by drawing explicit comparisons and contrasts between it and prior work that they had cited? Please highlight the portion of the text where they discuss the novelty of the work and describe the distinctions below.

(N/A) The author cited no prior work or does not describe their study.

(-2) The author does not describe his or her own study clearly or draw any connections between the cited work and their own study.

(0) The author describes his or her proposed study at a high level and cites prior work but does not draw any clear comparisons between the the proposed study and prior work.

(2) The author clearly describes his or her proposed study and draws clear and explicit distinctions between the proposed study and prior work that he or she cites. The author goes on to explain the similarities and differences between the cited work and the author's own proposed study.

G.11 : Did the author show the novelty of his or her work by including at least one current study node that summarizes his or her proposed study or key features of it and then drawing comparison arc(s) between those nodes and one or more citation nodes to highlight similarities to and differences from the prior work? Please indicate the nodes and arcs that show this comparison below and describe any improvements that should be made.

Please cite the node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(N/A) The author includes no citation nodes or no current study nodes.

(-2) The author does not frame the current study node clearly and does not draw any paths between the current study node and the citation nodes.

(0) The author includes one or more current study nodes and citation nodes but does not draw any comparison arcs between the current study and citation nodes or, if he or she does so, the analogies and distinctions are not stated clearly on the arcs.

(2) The author clearly describes the proposed study in one or more current study nodes. He or she also draws comparison arcs connecting the current study nodes to citation nodes describing similar cited work and clearly articulates the analogies and distinctions drawn between them on the arc.

E.12/G.12: Argument Coherence (*Arg-Coherent*)

Range : -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.12 : Did the author present a coherent argument overall in which each element of the argument is related meaningfully to the others?

(-2) No, the author's argument is completely unclear and lacks any internal cohesion.

(0) While the author attempts to connect the elements of his or her argument, the argument lacks sufficient clarity and coherence.

(2) The author presented a clear and coherent argument in which the elements of the argument are meaningfully related to one-another.

G.12 : Did the author develop a single coherent argument diagram in which each piece of the diagram is connected to its neighbors either directly or though a valid path where each arc in the path has a clear warrant?

Please cite any node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(-2) No, the author produced a series of disconnected argument diagrams or pieces of argument diagrams with no overall coherence.

(0) While the author presented a single coherent diagram he or she: drew incorrect or nonsensical relations; drew empty or nonsensical nodes; or did not include clear warrants for the relationships represented by the arcs.

(2) The author defined a single coherent argument diagram with every node well framed and all of the included arcs have clearly explained warrants.

E.13/G.13: Argument Convincing (*Arg-Convincing*)

Range : -2 -1.5 -1 -0.5 0 0.5 1 1.5 2

E.13 : Does the author present a convincing argument? That is, after reading the essay are you willing to accept the author's general argument in support of his or her proposed study?

(-2) The argument is wholly unconvincing.

(0) The argument is partially convincing but incomplete.

(2) The argument is complete and convincing.

G.13 : Does the author present a convincing argument in the diagram? That is, after examining the diagram are you willing to accept the general argument in support of his or her proposed study represented by the diagram?

Please cite any node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(-2) The argument represented by the diagram is wholly unconvincing.

(0) The argument represented by the diagram is partially convincing but incomplete.

(2) The argument represented by the diagram is complete and convincing.

E.14/G.14: Overall Quality (*Arg-Quality*)

Range : -5 -4.5 -4 -3.5 -3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

E.14 : Please rate the overall quality of the argument based upon the organization, coherence, and completeness.

(-5) Poor.

(0) Fair.

(5) Good.

G.14 : Please rate the overall quality of the argument diagram based upon its organization, coherence, and completeness.

Please cite any node and arc id numbers (located in the top-left of the box) by entering the number in brackets e.g.: [1], [5], or [6,7].

(-5) Poor.

(0) Fair.

(5) Good.

APPENDIX D

GRADING MATERIALS

The memo included here was provided to the graders along with the grading rubrics (see Appendix C (pp. 187)) as an introduction to the grading process. This was supplemented both by training on LASAD and direct discussion.

Grading Rubric Proposal.

Collin Lynch

June 29, 2013

1 Grading

As noted by Greene [1] outlining is important when planning a written item with multiple rhetorical goals like the research introduction. A structural or goal-based outline of the type discussed by Greene can help to order the material to be presented. By contrast a functional outline of the type used in the LASAD diagrams can not only help the reader order their material but highlight the, often complex, conceptual or structural relationships between the components.

In general there are three levels of grading for the types of arguments made in research reports: the overall *rhetorical goals* (i.e. did they get the point across or explain the work); the *structural criteria* (i.e. did they address appropriate grading criteria, is the discussion connected?); and the *basic logic* (i.e. did they present contradictions or otherwise inconsistent claims?). For the present task the grading rubric is focused on the rhetorical goals and the structural criteria. We will not be focusing on the basic logic.

The grading task involves two parallel rubrics one targeted to the essay grading task and the other targeted to diagrams. Each rubric consists of a set of 11 questions to be graded on a 5 point Likert scale from -2 to 2, and one overall question to be graded on a 11 point scale from -5 to 5. Ranging in general from -2 being "not at all" or "poorly" to 2 being "very well." For some of the questions, the binary choice of N/A is also available. Each question is presented below with a short guide to the anchor values of -2, 0, and 2. Intermediate scores should be treated as half-way between the states described on the anchor values. In the case of multiple items, such as the hypotheses or hypothesis nodes the grader should assign an average score.

1.1 Assignment

Before introducing the grading rubrics it is important to describe the general writing and outlining tasks and to refine our terms. The diagrams and essays to be graded were drawn from one of two assignments. In the diagramming assignment students were tasked with reading a published research report and diagramming the argument made in the introduction section. In the planning assignment students were tasked with planning an argument for their own paper using LASAD and then writing it. In both assignments the diagramming model and the essential argument instructions were the same.

In the introduction section to a paper the author seeks to present his or her overall argument. He or she wants to: explain what is in the paper; present their core research questions; make any general claims or subclaims; articulate testable hypotheses; cite relevant literature; and describe their study at a high level.

A Research Question is a general statement of scientific interest such as "Are people nicer to good-looking individuals?" Rhetorically these questions can break down into subquestions such as: "Are men nicer to good looking women?" The general research question is or should be stated at the start of the introduction while sub-questions are typically used to structure the text and organize paragraphs or subsections. Hypotheses by contrast are the specific predictions that will be tested by the study being described such as "Men are nicer than women when talking to ugly people."

A Claim is a general statement or assertion of fact such as "Gender differences influence social behavior." Claims can also be broken down into subclaims such as "Women are meaner than men." Claims and subclaims are used to state assumptions, frame the discussion, or make basic assertions relevant to the argument.

When presenting his or her own research question or subquestions the author should argue that the question is *open* or unanswered and *relevant* to the research community. He or she typically does this by citing prior studies that disagree or draw conflicting conclusions about the research question and the hypotheses. In the case above, for example, the author might cite a study which concludes that "...people let looks guide their social behavior..." and another that concludes "...looks do not matter." Broadly speaking these studies disagree about the research question with the former generally supporting it while the latter opposes it. This disagreement indicates that the question is open and relevant to the research community.

After noting the disagreement, the author should attempt to explain it by drawing analogies between the studies and highlighting key differences. Both studies are similar in that they focus on affective social behavior. However, he or she might point out, one study focused exclusively on public actions by college-age males on campus and the other on public behavior across ages and genders. The author should then go on to compare the cited work to his or her proposed study, again highlighting similarities and differences, to defend the *novelty* of the author's proposed study and explain how it would address these differences.

When making a claim, by contrast, the author is asserting that it is true. An author should defend the validity of the claim by citing prior work that supports the claim either theoretically or empirically.

2 Specific Instruction Caveats

During the Spring 2011 study students were instructed to use the "Current Study" node to note relevant features of the current study (e.g. sampling college students only; testing with warm and cold sensation). This node was then to be used both to provide support and opposition for claims or additional information. They were also instructed to draw analogies and distinctions between the current study node and cited nodes by means of the comparison arc. This arc has fields for citing the analogies or similarities between two nodes and the differences between them.

Rhetorically the intention was for students to, for example, cite a prior study on the level of haptic perception in individuals noting that the authors sampled a range of ages. They would then draw a current study node noting that their study would focus on college age students and control for the temperature of the object being sensed. They would then draw a comparison arc noting that both studies would focus on haptic perception, an analogy, but further note that their study would be restricted to college-age students, and control the temperature of the item being sensed, distinctions.

Please also note that the questions below are meant to cover the whole essay or diagram. As the example scores should illustrate the authors should receive full scores for well-formed citations, for example, only if they do so consistently. They should also receive the minimum points if they fail consistently. Thus the points assigned for a question should cover the entire diagram.

3 Essay Grading

In the initial round the essays will be graded in paper form. When grading questions in the Rhetorical goals section we ask that the graders highlight some of the segments that correspond to the questions. For the bulk of the grading, however they will only be asked to assign grades. The grading task itself will proceed based upon the introduction sections provided to the graders.

When grading the essays the graders will use a form based on the E.* questions listed at the end of this document.

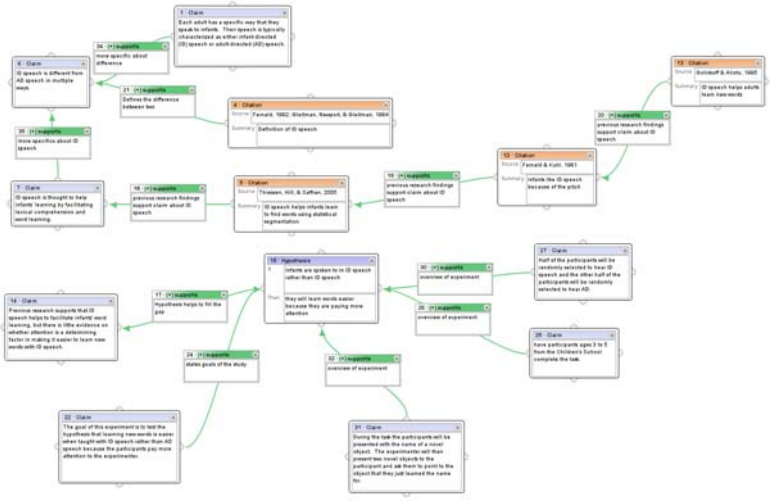


Figure 1: Example Diagram

4 Diagram Grading

Diagram grading will proceed with the student diagrams in the system. For the purposes of the rubric a few terms need to be specified. Firstly, in the diagramming language there is no specific node type for research questions. Both research questions and claims are represented in a claim node, the framing of the contents (as a question or statement) differentiates the two.

Secondly we should note that the supporting (+), opposing (-) and undefined (?) arcs are all directional. That is they point from a node A to B. Each of these arcs has a text field on them in which the student can enter a *warrant* that justifies the relationship. A warrant in argumentation theory is a reason or explanation for why the relationship holds. For example, when we claim that male students are more likely to help good looking female students we may cite a prior study that tested this hypothesis. The warrant in this case, the reason that the cited work supports our claim, is that the authors tested this claim in their own study and validated it.

Only the comparison arc is bidirectional and has no warrant, only specialized *analogy* and *distinction* fields. This arc is designed to be used to draw specific comparisons and contrasts between nodes such as noting the differences between a pair of cited works or analogizing prior citations to the present study. For some of the grading questions below the directionality of the arcs will not matter. If for example the question asks about a supporting arc from a citation node to a claim node for example then the students should lose points if the arcs run in the wrong direction. Similarly paths may also be directional in which case all arcs should be in the same direction from the source node to the target node.

In a graph a pair of nodes A and B are *directly connected* if there is an arc from one to the other in the diagram. In figure 1, for example, citation node 5 is directly connected to claim node 7 via supporting arc 18. The warrant for this citation provided on the arc is "Previous research findings support claim about ID speech." A *path* exists between two nodes if there is a route from one to the other through zero or more intermediate nodes. A *supporting path* is a path where each arc in the path is a supporting arc. Citation node 13, for example, is not directly connected to claim node 7. However there is a supporting path from

nodes 13 to 7 that includes citation nodes 12 and 5. Similarly there is a supporting path with no intervening nodes between hypothesis node 15 and claim node 14. No path, however exists between node 13 and any of the claim nodes in the lower part of the image, nor is there a path *from* claim node 14 to claim node 31 as the arc between nodes 14 and 15 points the wrong way from node 15 to node 14 thus violating the path.

Nodes may be *central* to a diagram or be *root* nodes if they play a central role in the argument. This is determined graphically based upon their position in the diagram and relationship to other nodes. In figure 1, claim node 6 is a root node for the diagram as all nodes in the upper half of the diagram are connected to it either directly or on a directed path. This is not an ideal structure, however, as node 1 contains the overriding research claim for the study and should therefore be the root node. Similarly in the lower-half of the diagram claim node 15 is a root node as all other nodes point to it via paths through node 15.

When grading the diagrams the graders will use a form based upon the G.* questions at the end of this document.

References

[1] Laurence Greene. *Writing in the Life Sciences*. Oxford University Press: New York, 2009.

APPENDIX E

SNG MANUAL

The attached manual describes the SNG grading toolkit. This is a web-based grading tool implemented in Javascript that I developed for the purposes of this thesis. The tool allows for graders to read and make rudimentary annotations of a text document while answering free-text and multiple-choice questions.

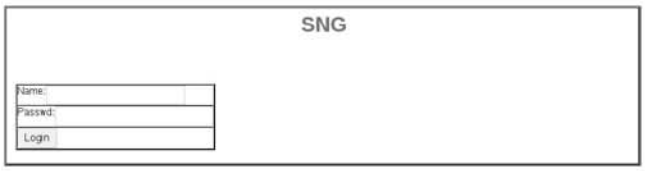
Using the SNG System.

Thank you again for agreeing to grade the essays and diagrams for the study. The grading, as we discussed, will take place in an online tool called SNG. This system has been deployed on a local server and will be used to read and grade the essays as well as the diagrams. The system has been setup to save your work as you go along and provides for regular backups of the data.

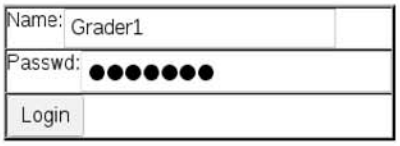
This User guide will provide a brief overview of how you work within the system and a guide to its components. As always with any questions just e-mail me (collinl@pitt.edu).

Login

To access SNG go to the url provided in your contact mail. You will see the following login screen:



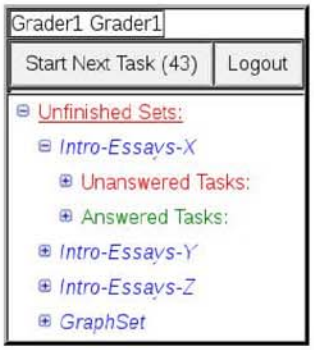
Enter the username and password provided in your e-mail and click on the login button:



The initial load time of the system can take a little while so please be patient.

Status Panel

Once the system loads you will see a Status Panel as shown below:



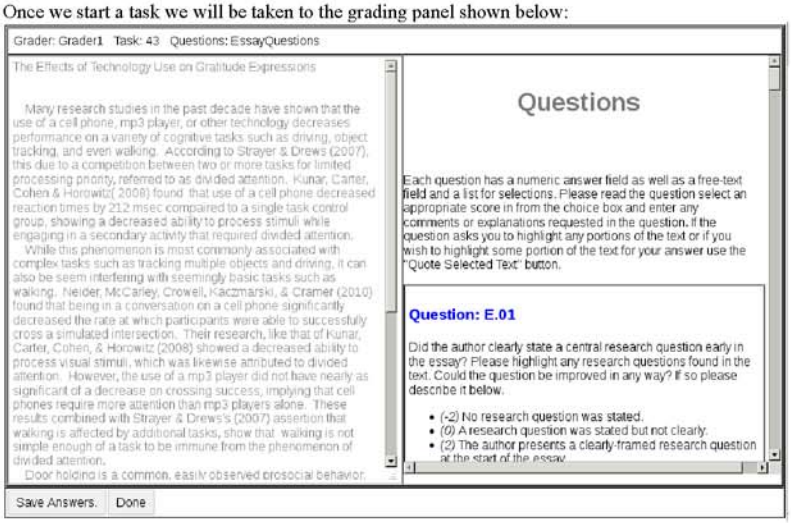
The status panel lists includes three features, the *Start Next Task* button, the *Logout* button, and the list of tasks. Clicking on the *Start Next Task* button will start the next uncompleted task in your list. The number indicated in the button is the ID of the next task to be completed. Once all assigned tasks are completed this button will be greyed out. Clicking on the *Logout* button will log you out of the system. You should use this button once you are done working.

The drop-down list provides a drop-down list of the task sets that have been assigned to you. It is color-coded. The top red element in the tree *Unfinished Sets* indicates that the task sets below it are not complete. As task sets are completed they will be moved to a new list, not shown, called *Finished Sets*.

The second layer of the tree, the blue items, are the set names. Each of these sets will include one or two subtrees labelled *Unanswered Tasks* (red) and *Answered Tasks* (green). Expanding these will show a list of buttons for the tasks as shown below. As you can see the topmost unfinished task is also the one numbered by the *Start Next Task* button. Clicking on any one of these buttons will start the specified task.

Grading Panel

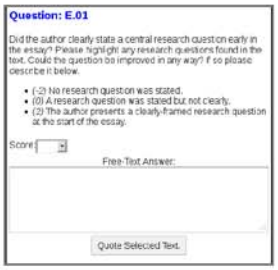
Once we start a task we will be taken to the grading panel shown below:



The panel contains header information across the top indicating the grader and the task being worked on. The essay being graded appears on the left hand side while the list of questions appear on the right. The two buttons at the bottom are used to save answers and close the grading panel respectively. I will discuss them in more detail below.

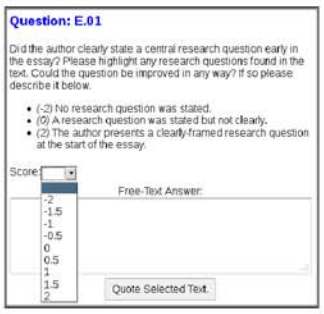
Questions:

The Questions Panel on the right hand side contains a header at the top describing how each question should be answered. As noted the questions all ask for a score value. Some also prompt you to highlight hypothesis statements or other pieces of the essay, and to enter an explanation for the scores. Each question is described in a question box of the type shown below:

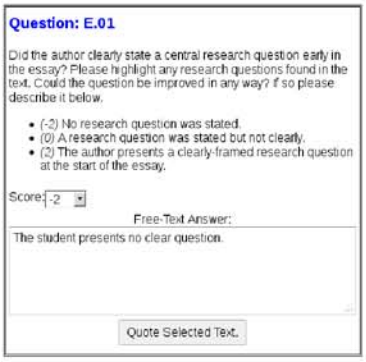


Reading from top to bottom, each question box contains: a question name (top in blue); the question text along with example score values; a pulldown menu for the scores; an answer field for the free-text answer; and a quotation button.

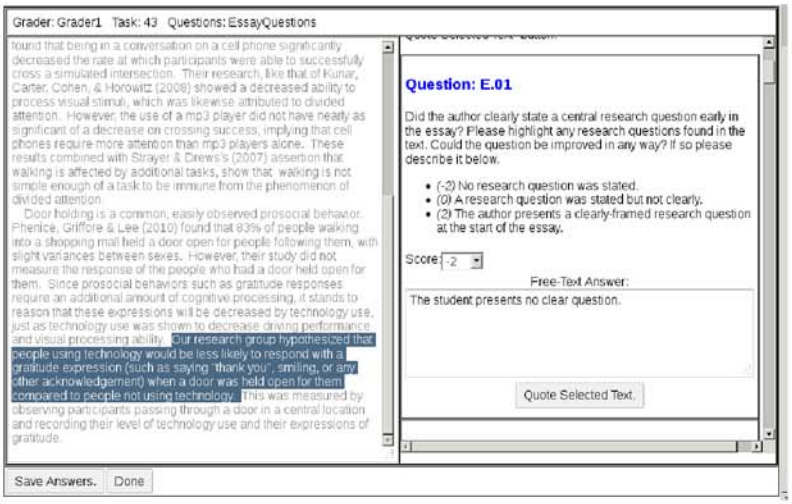
In order to set the question score you should click on the Score pulldown menu to select the appropriate value as shown below. For the purposes of grading a blank value indicates that you have not chosen a score value while a value of *N/A* means not applicable and is available for some questions.



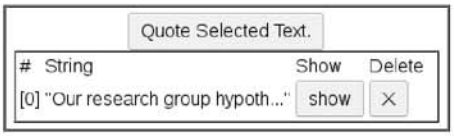
The free-text answer field is used for entering an explanation of your chosen score or descriptions as shown below:



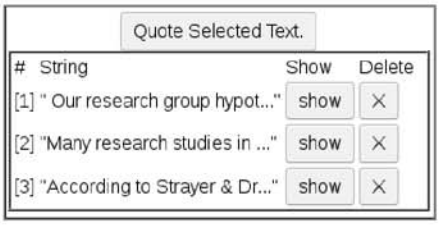
For some questions you are asked to highlight applicable portions of the text (such as the hypotheses) if any. You can also highlight other sections of the text that you wish to refer to in the essay. The *Quote Selected Text* button is provided for that purpose. In order to quote the text you should highlight a portion of the text in the essay panel. As shown below:



Then click on the *Quote Selected Text* button to produce a selection in the question box:



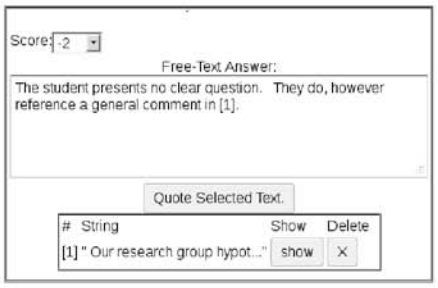
The table contains an ID in brackets that you can use in your free text answer when referencing the selection. It then contains the first 25 characters of the selection as a reference as well as *Show* and *Delete* buttons. You can add as many selections as you like to the table by making other highlights and clicking on the *Quote Selected Text* button again.



Clicking on the show button will cause the span of text that you selected to be highlighted in the essay field. Due to interface limitations only one span of text in the essay can show as highlighted at a time. Thus in order to view your selections you should use the *show* button. Clicking on the *X* button will remove this entry from the table.

For convenience each time that you add an entry to the table a reference to it will be appended to the free answer field using the same syntax as the ID. This is just added as a set of characters so you can freely insert them into your answer, change the text, or remove them if they are not necessary. When you delete entries from the table the text is not changed.

When referring to selections in your answer please use these references as you work as shown below:



Saving Answers:

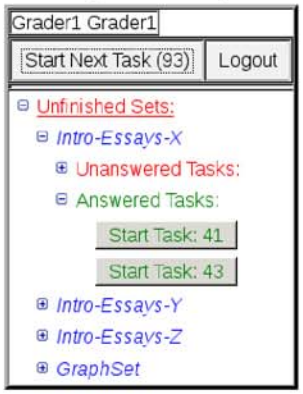
At the bottom left-hand corner of the grading screen there are two buttons: *Save Answers* and *Done*. Clicking the former will save your answers to the database but allow you to continue working. Clicking the latter will save your answers and, if successful, return you to the login screen.

In order to save the answers and close you must have provided a score for each question. If you click either button and you have not answered all of the questions you will see a message listing the questions that remain:

Once you have successfully answered all of the questions then you will see a success message:

Status Panel (Again)

Once you have completed all questions and return to the status menu you will see that the task has been added to the *Answered Tasks* sublist and you can move on to the next item. The *Start Next Task* button will also be updated so clicking on it will take you to the next undone task.



When you are done working for the time being just click logout and you can close your browser.

FAQ:

Q: Aaaahh! My browser closed! / The Internet died! Is my work gone?

A: If you close your browser when you are in the status panel your work has already been saved so no data will be lost. If, however you close your browser when you have a grading panel open then only your last saved work will be saved. In order to reduce excess load the system does not maintain a persistent connection and transmit an answer on every mouse-click only when you hit the *Save Answers* or *Done* buttons.

APPENDIX F

AUGMENTED GRAPH GRAMMARS

F.1 INTRODUCTION

Graph grammars are a formalism analogous to string grammars that are used to generate or modify graphs using existing grammar-based production rules. This library makes use of an augmented graph grammar formalism that I will define below. I will begin, however, with a short introduction to graph grammars and a citation of the relevant types.

F.2 BACKGROUND

Graph grammars were initially defined by Rekers and Schürr as a formal grammar whose atomic components are graphs, and where the rules or *productions* in the grammar transpose one graph to another [97]. More formally, they define a graph-grammar as:

> **Definition 3.6** A graph grammar GG is a tuple $(A; P)$, with A a nonempty initial graph (the axiom), and P a set of graph grammar productions. To simplify forthcoming definitions, the initial graph A will be treated as a special case of a production with an empty left-hand side. The set of all potential production instances of GG is abbreviated with $PI(GG)$.

And they define grammar productions as:

> **Definition 3.2** A (graph grammar) production p := (L; R) is a tuple of graphs over the same alphabets of vertex and edge labels LV and LE. Its left-hand side lhs(p) := L and

its right-hand side rhs(p) := R may have a common (context) subgraph K if the following restrictions are fulfilled:

- $\forall e \in E(K) \Rightarrow s(e) \in V(K) \wedge t(e) \in E(K)$ with $E(K) := E(L) \cap E(R)$ and $V(K) := V(L) \cap V(R)$ i.e. sources and targets of common edges are common verticies of L and R, too.
- $\forall x \in L \cap R \Rightarrow l_L(x) = l_R(x)$ i.e. common elements of L and R do not differ with respect to their labels in L and R.

Thus a graph grammar in their view is a system of production rules that, as with a formal grammar specifies a set of possible graph transformations. For efficiency reasons Rekers and Schürr restrict their attention to *layered graph-grammars* which imposes the additional restriction that the left hand side of each production rule define a smaller graph than its right-hand side thus avoiding cyclical derivations[1].

The formalism defined by Rekers and Schürr deals with standard context-sensitive grammar. This format assumes that nodes and arcs are drawn from a finite alphabet of types and contain no complex structure. A richer formalism that deals with extended graph structures is described by Pinkwart et al. in [89]. Here the author describes a grammar structure that treats nodes and arcs as complex structures containing both static types and complex fields. The authors further detail a set of set-theoretic functions over those fields used for constraint checking. In many respects the formalism that I describe here is consistent with that formalism save for syntactic changes and the addition of existential rules for paths. Some of these features can also be found in the existing LASAD AFEngine which makes use of a CLIPS rule approach [107].

F.3 AUGMENTED GRAPH GRAMMAR FORMALISM.

Roughly speaking the full AGG formalism is a method for defining first-order logic over graphs and graph components. Formally I view a graph as a set of nodes and arcs with optional added constraints. Full-fledged graph grammars allow for the construction of expressions of the form shown in Figures F1 (pp. 234) - F8 (pp. 238).

[1] For more on Formal grammars see: Sipser "Introduction to the Theory of Computation" [111].

As these figures illustrate, the graph grammars allow us to specify logical arguments over the graphs including existential and universal quantification, chaining, and recursive productions. In the subsequent sections I will explain each of the items in detail building up to the full grammar expressions. The above expression can also be written in a more textual syntax as shown below in Figure F4 (pp. 237). This textual format is designed for use with a grammar compiler which I will address later. In the subsections below I will describe both formats.

Where this formalism extends beyond the work above is that it allows for scoping and focuses on the addition of constraints for node features.

F.3.1 Constraints

Constraints represent individual bounds or limits on the node and arc features. Constraints such as $f.Type = Citation$ are used to specify features of the rich nodes and arcs. As shown in Figure F5 (pp. 237) these constraints can be *atomic* focusing on a single node; *paired* in which case they deal with a relationship between two elements; or *complex* where they combine the first two types.

Syntactically the constraints use a basic s-expression syntax with the individual components of the fields specified using a dot-successor format thus:

$$(<Var>.< Field >[.<Subfield>...]<Relation><Val>)... \tag{F.3}$$

Where *Var* is the variable name, *Fields* specifies the complex field being denoted, *Relation* is a binary relation over the items, and *Val* links to the other values. *Subfields* are optional specifiers for more complex access such as specific named fields as shown in fig 1. Syntactically this is akin to the access of methods in an OO language. Constraint expressions are specified the same in both formats.

Constraints can also be grouped into blocks as shown in Figure F6 (pp. 237). Block constraints designate a single variable scope for the subsequent expressions. This format merely provides additional syntactic sugar above and beyond the current form.

F.3.2 Graph Schema

A Graph Schema is a graph structure, represented diagrammatically or textually, that specifies a graph structure. This format includes named nodes and arcs which may in turn be directed, undirected, or unknown. The function of the graph schema is to form a matching structure and to designate variables for the nodes and arcs. As shown in Figure F7 (pp. 238) this can be done diagrammatically or with a textual listing of nodes and arcs.

The schema adopt the following syntax conventions:

1. **Atomic Elements** Lower-case node and arc variables designate *atomic* elements which are used in the current class (see below).
2. **Production Elements** Capitalized variables designate production items which are replaced by productions (see below).
3. **Negation** is specified for the individual items using ¬ symbol to denote explicitly missing items.
4. *** Endpoint** For syntactic reasons it is often necessary to specify an arc where we do not care about the specific node contents. While this can be handled by introducing a variable with no constraints we can also use * as a stand-in name.

Functionally the schema are, at least at present, subject to the following semantic constraints:

1. **Non-repeating** All item names are unique within a schema even if subsequent items match in other details.
2. **Isolated Negation** As I will detail below chained negation raises complex conditions. Therefore all negated terms must be *isolated* that is no negated arc may have a neighboring node that is itself negated, and no negated node may have a negated arc. This ensures that processing of negation is clearly scoped.
3. **Isolated Productions** (also known as *Grounded Context*) As with negation the presence of neighboring variable productions generates a challenge for the ordering of expansion. Thus variable nodes and arcs must also be isolated. In future I may consider alternate productions but that would require other search challenges.

4. **Partitioned Productions** As with string grammars each novel element in a production must map to a novel (as yet unmapped) node or arc. That is, excluding the context nodes that appear on the left-hand-side of a production (x & y in Figure F2 (pp. 235)) all other ground nodes appearing only on the right-hand-side of the production (s & o) must match new nodes that have not been mapped by the containing classes. This is an efficiency designed to make the mapping search tractable. [2]

5. **Open Matching** Schema are matched only against their context and do not assume outside elements.

The Schema themselves are composed of *atomic-* and *variable-nodes* denoted by the upper and lower-case letters above. They also admit four arc types illustrated in Figure F8 (pp. 238): *directed* (e.g. D); *unknown direction* (UD); *undirected* (N); and *undefined* or *unknown* (U). Formally speaking a directed arc in a schema will only match a directed arc in the underlying graph from node a to node b. An unknown-direction arc will match a directed graph arc either from c to d or from d to c but will *not* match an undirected arc between the two. Undirected schema arcs will only match undirected arcs between e and f. And an unknown arc will match any arc between g and h. While this proliferation of arc types can complicate the syntax somewhat it does lend itself to easy specification of a number of graph types.

F.3.3 Graph Classes

A graph class $G_i = S_i + C_i$ consists of a graph schema paired with a set of optional constraints. Taken together these specify a class of matching graphs and form the basic component of the subsequent expressions and productions. The graph classes are the basic unit of graphs that are dealt with in this formalism. An example of the graph class is shown in Figure F9 (pp. 239).

By default the graph classes are defined within a single graph context and fill the same syntactic and semantic requirements of the schema above. It is *not* necessary for all member

[2] Under other circumstances this might be relaxed to apply only to arcs but for the present it is not required.

nodes and arcs to be coupled with existing constraints. However in the absence of a specified constraint all possible values are acceptable.

F.3.4 Graph Ontology

In grounded graph grammar of the type used by Rekers and Schürr the space of possible node and arc types is specified by a language $\sum$ consisting of unique types. Here a more complex structure is required. *Ontologies* specify the range of possible node and arc types and, for each type, list the set of possible fields and the field types. The ontology also specifies the set of possible relations and comparisons however standard relations (e.g. =) are implied.

I will flesh this out later but in rough terms an ontology is defined as shown in Figure F10 (pp. 240).

Formally speaking a graph ontology must specify:

- The set of possible node and arc types.
- The available fields and subfields attached to each node and arc type.
- The type of each specified field and subfield.
- The set of field relations usable for constraints on each type.
- The set of operations that can be used on combined fields.

Structurally this means that the ontologies are responsible for complex operations and must include embedded functional code or draw from a standard set of codes and types. For the present a standard set dealing with string values will be implemented.

The available types must be drawn from predefined types available in the language or those added by subclassing. Thus they will be built in by lookup meaning that ultimately a full compiler will be needed but not yet.

F.3.5 Graph Production

A graph production $C_l \Rightarrow C_{r1}|C_{r2}...$ is a context-sensitive production rule that maps from one graph class containing a *production variable* to one or more alternate expansions. An example production rule is shown in Figure F11 (pp. 241).

Formally speaking the production rules are replacement operations which expand or otherwise replace one part of the graph class with other related classes. The rules are context-sensitive and can include optional non-production nodes and arcs (e.g. h, c, s, x, & y above) in the schema. Such variables are necessary iff the variable nodes and arcs are expanded in an existing schema. Thus given the production rule above, and an initial class shown in the figure below some mapped classes are shown in Figure F12 (pp. 242)

The graph production rules are subject to the following restrictions:

1. **Isolated Productions** As noted above production variables, like negation, must be isolated from one-another in the parent graph. This may also be called a *Grounded Neighborhood*.
2. **Expansive** The the left-hand graph class C_l must be a proper subset of each right-hand class C_{ri}. Thus All nodes and arc variables as well as the constraints must be present and cannot be rewritten. This is also called *LHS-Grounding*. In future versions of the formalism I may consider $\emptyset$ productions but not for the present.

Formally speaking there exist two general types of productions: *Node Productions* where the central variable is a node which is replaced by a given subgraph; and *Arc Productions* which replace a given arc. The former is illustrated in Figure F12 (pp. 242) while the latter is illustrated below in Figure F13 (pp. 243). At present the formalism is focused on arc productions and the mapping restrictions below apply differently.

F.3.5.1 Recursive Productions, and Scope The production rules can be recursive with each of the subgraphs providing a recursive subgraph for expansion. A simple recursive production for an opposing supporting path was shown in Figure F2 (pp. 235).

The grammars can also be *recursive* with subgraph expansion taking place with each item. This recursive expansion can be controlled by means of *scoping*. In the rules above an explicit scope for class C_{r2} of rule P_2 is shown requiring that the production expand at minimum zero times and at most twice. Thus the production imposes a maximum depth to the recursive call. This form of scoping is more complex to implement but provides nice syntactic gains while reducing the complexity of the ruleset.

F.3.6 Production Mapping

When applying a graph production we face an additional question about how to map the context class, and thus the context elements for subsequent expansion. In order for the class to be consistent we must consider all viable mappings from the subclass to the superclass. That is, given a graph class with an arc class or node class within it we must consider all possible classes. The rules for mapping are discussed below.

F.3.7 Arc Productions

Arc productions are defined by a context class consisting of a single directed arc with its attached nodes. One such production is shown in Figure F13 (pp. 243). Unlike ground arcs, variable arcs can only have one of two orientations. They can either be directed (pointing from one ground node to another) or of unknown direction (in which case no direction is specified). The context graphs, by contrast, must always use directed arcs to specify the variable arc. This is necessary so that we can establish the appropriate mapping if the parent class specifies a directional relation.

Now, when we consider the mapping process this allows for the following cases illustrated in Figure F14 (pp. 244). If we were mapping the production class $X_{context}$ from Figure F13 (pp. 243) to class C_A then we have two possible mappings for the context nodes: $\{a : w, b : z\}$ and $\{a : z, b : w\}$. While if we map it against C_B there is only one: $\{a : w, b : z\}$.

F.3.8 Graph Expression

The final component of the grammar language is the graph expressions represent chained quantified graph classes of the form:

$$\forall C_0 \quad | \quad \exists C_i \quad | \quad \neg\exists C_2 \quad | \quad ... \tag{F.4}$$

A sample expression was shown in Figure F1 (pp. 234). Each class is quantified using the universal or existential quantifiers and chained using a standard pipe which is read as "...such

that...". Thus in the example shown in the figure below we assert that for all subgraphs that match class C_0 there is no surrounding class matching C_1.

Formally the graph expressions are piped class tests linked by the use of shared variables. More formally an expression E_g must be expansive or *Right-Grounded*:

$$\forall C_{i>0} \in E : C_{i-1} \subseteq_g C_i \tag{F.5}$$

For the present Right-grounding also requires that the final class in the expression be an existential ($\exists$) test with optional negation. Thus for an expression containing $|E_g| = j$ classes the final class C_{j-1} can only be scoped as $\exists$ or $\neg\exists$. Thus the final graph class is a match that must be valid or not for the expression to hold true.

These requirements are a direct consequence of the use of the augmented graph grammar for graph matching. In order for the graph to hold it must be the case that the final expression can be matched.

Mapping for the expressions is a cascading process with some interesting caveats based upon the RHS. That is, given expressions of the form:

$\exists C_i | E_j$ This holds iff $\exists C_i$ mapping over G s.t. the expression E_j holds given the mapping.

$\neg\exists C_i | E_j$ This holds iff $\forall m(C_i, G)$ that is for all mappings of C_i over G none of them satisfies E_j.

$\forall C_i | E_j$ This holds iff $\forall m(C_i, G)$ it is the case that E_j holds given M.

$\neg\forall C_i | E_j$ This holds if there exists some mapping $m(C_i, G)$ s.t. E_j does not hold.

This fact of quantification means that finding a "mapping" over an expression means finding mappings only for the leftmost set of subexpressions that are $\exists$ or $\neg\forall$. The other quantifications scope over all cases.

Table F1 (pp. 245) shows the logical breakdown of the cases.

F.3.9 Text Syntax

Formally speaking a graph schema S is defined by a 2-tuple (S_N, S_E) of node and arc variables. While this is conventionally represented in the form of a diagram it can also be shown as a textual pair as shown in figure Figure F4 (pp. 237). Here lower-case variables are used to designate non-production nodes and arcs while upper-case items are used to represent production variables. In order to represent a class we expand this to a 3-tuple $C = (C_N, C_E, C_c)$ of nodes, arcs, and constraints.

In standard graph theory arcs or edges are denoted as $e(a, b)$ for undirected arcs and $e\overrightarrow{(n, m)}$ for directed arcs. Here I use a similar syntax however the arcs are designated using their variable name and directionality is indicated through the use of [notation. Thus:

- $a(b, C)$ denotes an open or unspecified arc connecting nodes b and C.
- $D(f, g>$ denotes a strictly directed arc labeled with the production variable D from nodes f to g. This is equivalent to $D = e\overrightarrow{(f, g)}$.
- $h[I, J]$ denotes a strictly undirected arc labeled with the non-production variable h connecting production nodes I and J. This is equivalent to $h = e(I, J)$.

I have previously shown how conditions are specified using standard textual notation above. That is continued here.

As shown in figure 1a productions are generated using the standard pipe syntax and can be denoted using variables for convenience as shown in the figure below. This is also true for expressions as shown in Figure F1 (pp. 234).

F.3.10 Future Alternatives

There exist a number of future alternatives that may be considered such as *ordered negation*. Here negation is scoped using the lexicographic ordering of the node and arc variables. While this is doable it could get out of hand with a schema such as the arc shown in Figure F4 (pp. 237) yielding Equation F.6

$$\exists A : \neg\exists b(a, y> : \neg\exists x(*, Y) : \neg\exists Y : ... \tag{F.6}$$

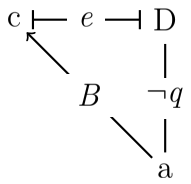
This this like complex expansion would require additional, and often undesirable, search.

Syntactically the requirements of right-grounding for expressions may impose some limits on the system that should be avoided. It may be useful to allow $\forall$ quantified classes at the right hand side but a clear functional semantics must be defined.

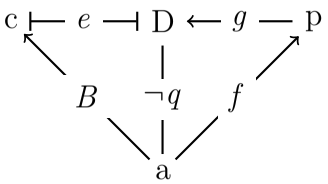
F.4 COMPILATION & EVALUATION

At a basic level classes are used to test for the existence of satisfactory items. Implicitly all classes represent existentially quantified claims over the transaction context of the form $\exists C_i$ or $\neg\exists C_j$. Therefore any compilation of the expression must perform a complete search in order to negate by failure. Similarly, universally quantified items represent exhaustive collections for evaluation. Thus in general the processes will be done in a brute-force manner as determined by the compiler.

Having said that some search efficiency is potentially useful. Search across the productions, for example, can proceed in a *depth-shy* or *context-driven* manner where we focus on indexing all possible non-production nodes before considering candidate expansions. When considering the class C_1 shown in Figure F1 (pp. 234), for example, we can first collect all possible claim nodes and then proceed from there with the search for citations and word-sets before proceeding with the production rules for B and D. Negation, of course, poses a challenge because while we can index non-matching nodes or arcs they make the most sense within the declared context.

(C_1) 

$$\left\{ \begin{array}{c} a.Type = \text{“}Claim'' \\ e.Type = \text{“}Claim'' \\ c \left\{ \begin{array}{c} field(\text{“}Words'') = \{Handedness, ...\} \\ Type = \text{“}Claim'' \end{array} \right\} \end{array} \right\}$$

(C_2) 

$$\left\{ \begin{array}{c} f.Type \in \{\text{“}Supporting'', \text{“}Undefined''\} \\ p.Type = \text{“}Claim'' \\ g.Type = f.type \end{array} \right\}$$

$$\forall C_1 \quad | \quad \neg\exists C_2 \tag{F.1}$$

Figure F1: An example graph expression with subclasses.

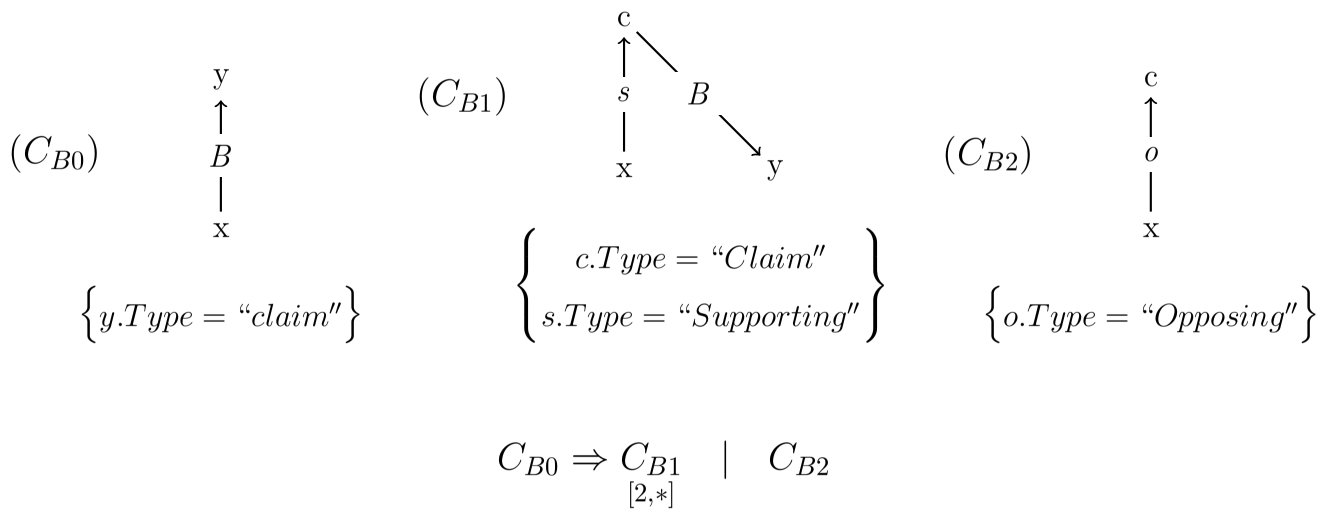


Figure F2: An example graph production with subclasses.

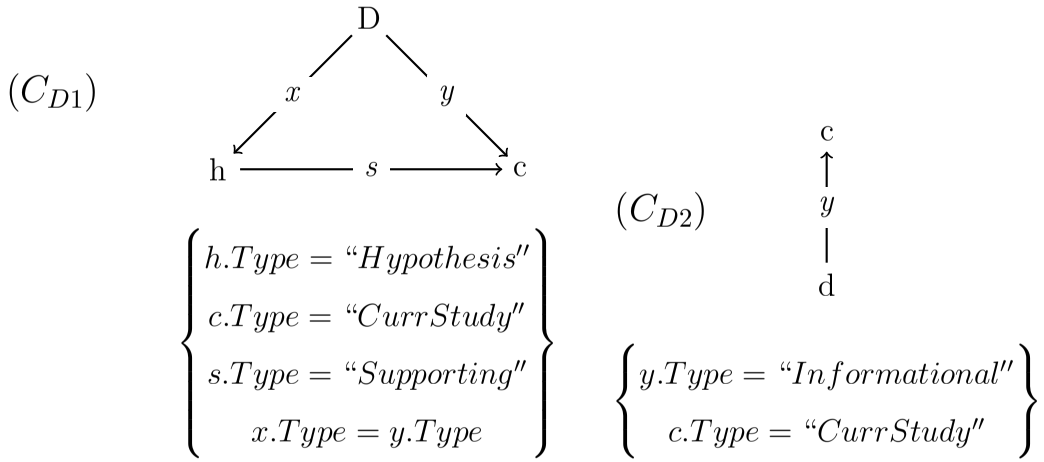


Figure F3: An example graph production with subclasses.

$$C_1 = \{(a, c, D, E),$$
$$(B(a, c>, q<a, E>, e[c, D], [(\neg,])qDa),$$
$$(a.Type = \text{“}Claim'', e.Type = \text{“}Claim'',$$
$$c : \{field(\text{“}Words'') = \{Handedness, ...\}, Type = \text{“}Claim''\})\} \quad \text{(F.2)}$$

Figure F4: Sample Class 1 represented in textual format.

$$\left\{ \begin{array}{c} a.Type = \text{“}Claim'' \\ e.Type = \text{“}Claim'' \\ c \left\{ \begin{array}{c} field(\text{“}Words'') = \{Handedness, ...\} \\ Type = \text{“}Claim'' \end{array} \right\} \\ x.Type = y.Type \end{array} \right\}$$

Figure F5: Constraint set example.

$$c : \{field(\text{“}Words'') = \{Handedness, ...\}, Type = \text{“}Claim'', x.Type = y.Type\}$$

Figure F6: Constraint group example.

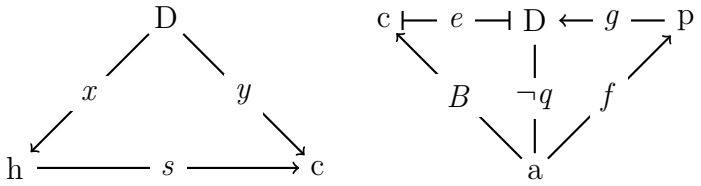


Figure F7: Graph Schema Examples.

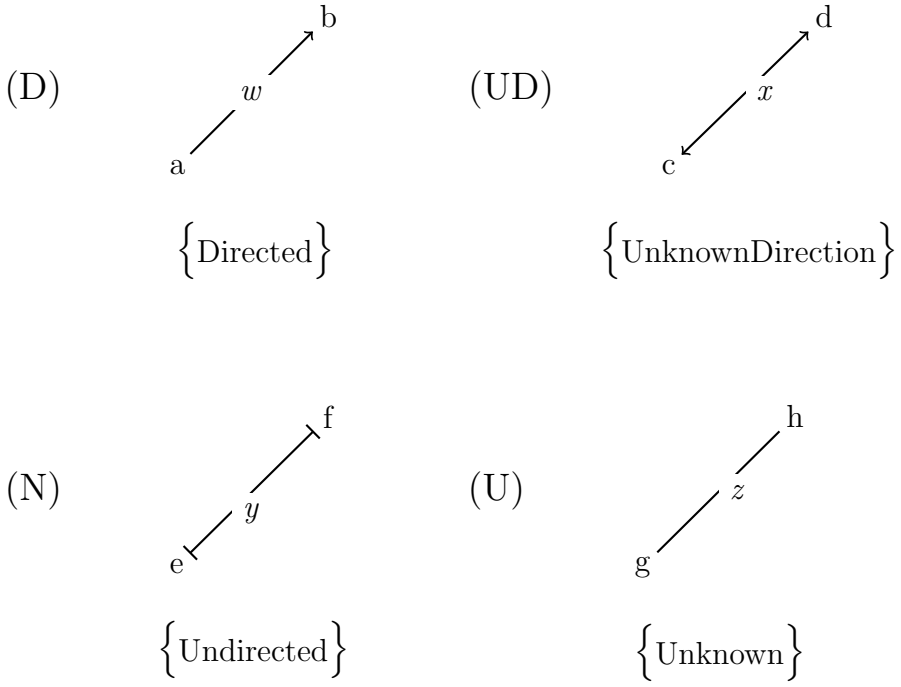
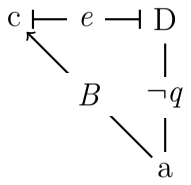
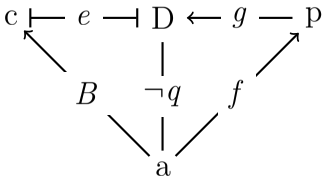


Figure F8: An example graph production with subclasses.

(C_1) 

$$\left\{ \begin{array}{c} a.Type = \text{“}Claim\text{”} \\ e.Type = \text{“}Claim\text{”} \\ c \left\{ \begin{array}{c} field(\text{“}Words\text{”}) = \{Handedness, ...\} \\ Type = \text{“}Claim\text{”} \end{array} \right\} \end{array} \right\}$$

(C_2) 

$$\left\{ \begin{array}{c} f.Type \in \{\text{“}Supporting\text{”}, \text{“}Undefined\text{”}\} \\ p.Type = \text{“}Claim\text{”} \\ g.Type = f.type \end{array} \right\}$$

Figure F9: Graph Class Examples.

```
{
 Nodes:{
   Claim:{
     Text(String)
     Text.Words(StringSet)
   }

   Hypothesis: {
     If(String)
     If.Words(StringSet)
     Then(String)
     Then.Words(StringSet)
   }
 }
Arcs:{
  Comparison: {
   ...

Types: { String, StringSet }
...
```

Figure F10: Sample Ontology Structure

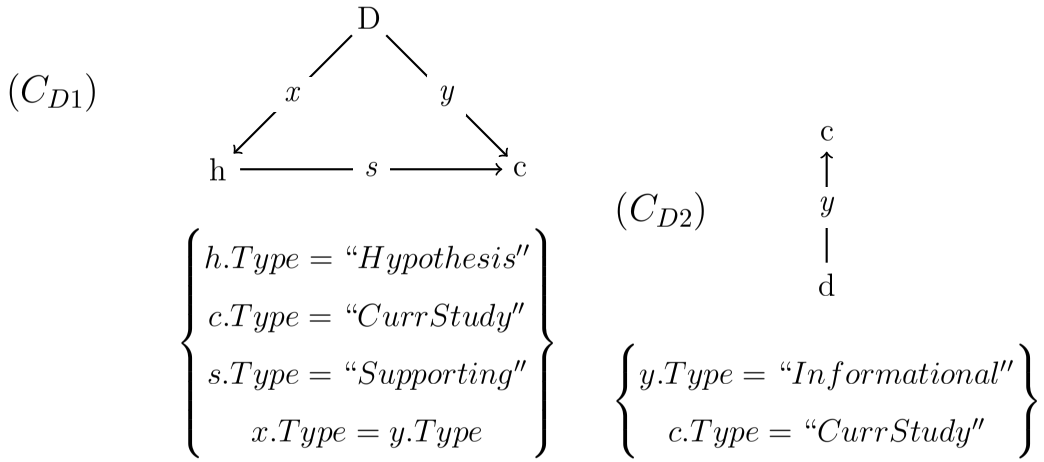


Figure F11: An example graph production with subclasses.

D

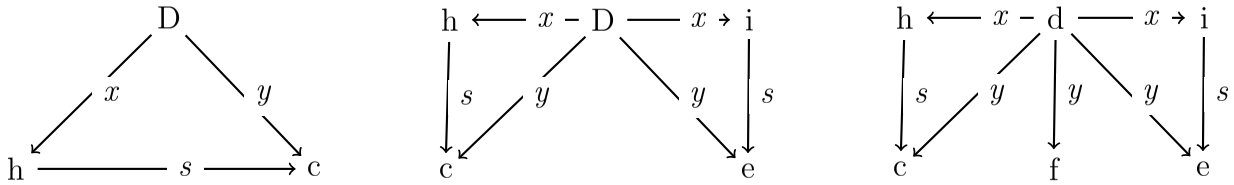


Figure F12: Graph Expansion Examples.

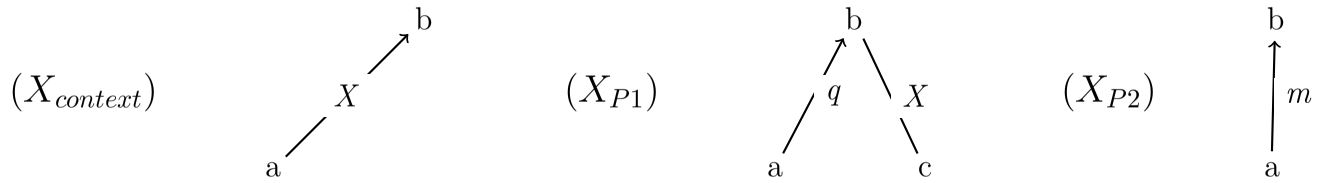


Figure F13: Arc production example.

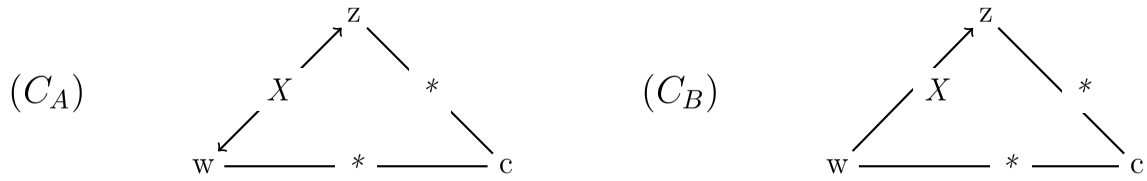


Figure F14: Variable arc mapping examples.

Table F1: Individual graph expression scope values. Here the outcome None means additional search is required.

Expression	$m(C_i, G)$?	E_j	Result	Note
$\exists C_i \vert E_j$	No	(T/F)	F	If none is found on the current try then no other exists.
	Yes	T	T	If a map is found then we are done.
	Yes	F	None	If false then we need to keep trying.
$\neg\exists C_i \vert E_j$	No	(T/F)	T	If none is found then none can be so True.
	Yes	T	F	If one is found then we fail.
	Yes	F	None	Else we should keep searching.
$\forall C_i \vert E_j$	No	(T/F)	T	It has held for the prior entries so it passes.
	Yes	T	None	It continues to hold so keep trying.
	Yes	F	F	This violates the rules so the search stops.
$\neg\forall C_i \vert E_j$	No	(T/F)	F	False as none so far have held negatively.
	Yes	T	None	Continue to search for the counterexample.
	Yes	F	T	Success as we have found the counterexample.

APPENDIX G

LINEAR REGRESSION

As noted previously in Subsection 7.2.1 (pp. 99) Linear Regression models are robust models common to many empirical domains. They are linear and additive models that represent variable relationships with a polynomial of the form:

$$y_i = \alpha + \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \epsilon_i \tag{G.1}$$

where α defines the *Intercept Point* or base value of the model; each β_k is a *coefficient* that defines the strength and the sign of the relationship between the independent variable x_k and the dependent variable y; and the *error term* ϵ_i defines the error term specific to the data term y_i.

In that subsection I noted that, according to the Gauss-Markov theorem, Least-Squares regression will produce an optimal model if the following conditions are met [34, 133]:

Linearity: The independent variables are linearly related to the dependent variable with constant individual effects represented by the β_k values. For generalized additive models (see [141, 54]) this assumption is relaxed to an assumption of *Additivity* (see: Section G.1 (pp. 247)).

Independence: The samples were taken independently and thus any error terms ϵ_i and $\epsilon_{j \neq i}$ are independent of one another as are the errors across the dependent variables y_j and y_k (see: Section G.2 (pp. 248))

Variability: The individual independent variables are non-constant (see: Section G.2 (pp. 248)).

Non-Multicollinearity: The independent variables are independent and not collinear. That is, no variable x_i is dependent on other variables $x_0, \ldots, x_{m \neq i}, \ldots, x_n$ (see (see: Section G.3 (pp. 248)).

Homoscedasticity: The variance of the error terms ϵ_i is constant for all observations. That is, the error variance is not affected by the independent variables (see Section G.4).

Normally-Distributed Errors: (aka *Normality*) The error terms of the model are normally-distributed: $\epsilon \sim N(0, \sigma_\epsilon^2)$ (see Section G.5).

Weak Exogeneity: The independent variables are error-free either because they are set by experimental condition (dividing students by age) or because they can be measured without error. As such they do not introduce a significant source of error into the model (see: Section G.6).

More specifically, if the assumptions of *Linearity*, *Independence*, and *Homoscedasticity* are met then the least-squares estimator will be the most efficient unbiased linear estimator available [34]. If the additional assumption of *Normally-Distributed Errors* is met then least squares regression will be the most efficient estimator over *all unbiased estimators,* even nonlinear ones. Crucially this list of assumptions does *not* include that of normally-distributed data which is generally, though incorrectly, assumed to be a requirement for least-squares regression. Distributional assumptions are only required if F-scores, p-values, or confidence intervals are being calculated [34, 133, 141, 54]. I describe each of these assumptions in detail below.

G.1 LINEARITY

The assumption that the independent variables are linearly related to the dependent variable is inherent to the structure of the models and is thus an *inductive bias* [82]. There are alternative regression models such as *generalized additive models* (see. [141, 54]) that relax this hard bias. However these models can themselves be computationally intensive and prone to over-fitting. As such I treat it as an assumption of the learning algorithm.

G.2 INDEPENDENCE & VARIABILITY

The assumptions of independence and variability are features of the data collection. As described in Chapters Chapter 3, 4, and 5 the diagrams and essays were produced by individual authors or unique teams while the grading of the individual items was conducted in a double-blind manner and the graph features were calculated automatically. Therefore the data are independent of one-another and the assumption of *independence* is satisfied. For the purposes of this analysis I also omitted all constant-valued features from consideration therefore the assumption of *variability* also holds.

G.3 NON-MULTICOLLINEARITY

Multicollinearity occurs when there are non-trivial relationships between the independent variables [36]. Thus given the standard linear model:

$$y_i = \alpha + \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \epsilon_i \tag{G.2}$$

It is generally assumed that there exists no *strong* linear relationship among the independent variables. That is, it is not possible to train a strong model of the form shown below:

$$x_j = \alpha' + \beta'_0 x_0 + \ldots + \beta'_{(j-1)} x_{(j-1)} + \beta'_{(j+1)} x_{(j+1)} + \ldots + \beta'_n x_{(n)} + \epsilon'_i \tag{G.3}$$

When such strong relationships exist the individual coefficients in the primary model $\beta_p, \ldots \beta_q$ will be sensitive to minor changes in the data, will be prone to over-fitting, and will thus be unstable under cross-validation. Such models are often characterized by: inflated variance, unexpected regression coefficient valences (e.g. a beneficial variable with a negative sign), and high overall performance coupled with low partial correlation coefficients (i.e. high overall predictiveness but with low partial success) [108, 77, 34, 36]. Therefore multicollinearity poses a threat to the reliability and generality of the models making it impossible to use them either to assess the relative impact of individual variables or to operate

on novel data.

A number of different methods have been proposed to detect multicollinearity, some of which are formally equivalent. These include: direct testing of individual collinear relationships via *Tolerance/VIF* tests [108]; calculation of a shared *condition number (K)* which estimates instability [34]; and application of exploratory *Principal Components Analysis (PCA)* or *Factor Analysis* [34]. In the current work I make use of the Tolerance/VIF tests to evaluate individual variables.

The *tolerance* of a variable x_i is defined as:

$$t(x_i) = 1 - R^2_{x_i} \tag{G.4}$$

where $R^2_{x_i}$ is the squared *Multiple Correlation* for the regression of x_i on the other independent variables $x_{j \neq i}$ [34, 108, 134]. Crucially: *variables with extremely low tolerance are strongly predicted by their neighbors.* While no formal threshold exists Schroeder and others propose a threshold of $t(x_i) \leq 0.01$ citing Afifi & Clark [1]. This is the corollary of another commonly used measure the *Variance Inflation Factor (VIF)* which is simply the inverse value:

$$VIF(x_i) = \frac{1}{1 - R^2_{x_i}} \tag{G.5}$$

Like tolerance, VIF is used for assessing which variables are multicollinear. VIF scores "in excess of 10" are treated as a sign of multicollinearity [108].

Both the tolerance and VIF thresholds are heuristic measurements that require a measure of judgment before changes are made. Moreover, while Tolerance and VIF provide variable-specific measurements, the values depend upon all other variables in the set. Thus two variables x_i and x_j with extremely low tolerance may, in fact be cross-correlated and cannot both be ruled out. As discussed in [36] the relationships between variables can be complicated with a set of independent variables containing multiple overlapping collinear subsets. Therefore there is no guarantee that a clean partitioning of the dataset can be defined and trimming individual variables will likely result in some loss of unique information.

It is important to note that Multicollinearity is a function of the independent variables and the dataset, not the dependent variables. As such remediation of multicollinearity is a subject of some debate. Proposals include: centering variables and accepting remaining problems [145]; pruning low tolerance variables [77]; application of biased methods such as *Ridge Regression* [108]; combination of collinear terms via PCA[1]; and collection of new data from scratch [34]. In each case the appropriateness of the proposals depends upon the goals of the analysis.

The goal of the present work is to induce parsimonious models from a larger dataset. As noted in Chapter 5 overlaps among the diagram features are expected; therefore individual variable pruning is appropriate. The greedy pruning algorithm used will be described in Section 7.3.

G.4 HOMOSCEDASTICITY

Homoscedasticity is the assumption that the error variance is constant across all observations thus:

$$\forall_i V(\epsilon_i | x_{i,0}, \ldots x_{i,n}) = \sigma_\epsilon^2 \tag{G.6}$$

where $V(a|b)$ is the variance of a given b [34].

As discussed above, this assumption is of primary concern for the efficiency of the least squares estimator as per the Gauss-Markov theorem. Both the assumptions of homoscedasticity and normally-distributed errors are frequently ignored in practice. I will discuss tests for Homoscedasticity in Subsection 7.4.3.

[1] Professor Chris Schunn, University of Pittsburgh, personal communication

G.5 NORMALLY-DISTRIBUTED ERRORS

The assumption of normally-distributed errors, frequently referred to by the confusing shorthand *normality*, is an assumption that the error terms are normally-distributed:

$$\epsilon \sim N(0, \sigma_\epsilon^2) \tag{G.7}$$

That is, we assume that the observed values y_i are normally distributed around the regression plane with a mean error of 0 and a fixed σ^2. As with Homoscedasticity the assumption of normally-distributed errors is required to show that the induced models are optimal but is *not* required to show that they are appropriate. I will return to this topic in the results section.

G.6 WEAK EXOGENEITY

Facially, weak exogeneity is an absolute assumption. In most real-world applications of linear models the measurement is never entirely free of error. Therefore the question is whether or not the information is sufficiently reliable for present purposes. In this chapter the focus will be on inducing models that match the automatic graph features described in Section 5.4 to the essay grades. The automatic features were calculated programmatically and as noted previously have 100% test-retest reliability. As such they introduce no salient measurement errors.

Similarly, the reliability of the manual grades was dealt with in Section 5.5. As noted there, all of the graph grades met standards of empirical reliability while five of the essay grades did so. Again that threshold is sufficient for the present analysis. Therefore, based upon the thresholds shown in Table 4.5 (pp. 56) I will focus on inducing models for *E.01 (RQ-Quality)*, *E.04 (Hyp-Testable)*, *E.07 (Cite-Reasons)*, *E.10 (Hyp-Open)*, and *E.14 (Arg-Quality)*.

G.7 NORMALLY-DISTRIBUTED DATA

One commonly-held assumption that was not discussed above is the assumption of normally-distributed data. It is commonly assumed that that linear models and least-squares regression require that both the independent and dependent variables be normally distributed:

$$\forall x_{*,m} : x_{*,m} \sim N(\mu_{x_{*,m}}, \sigma^2_{x_{*,m}}) \tag{G.8}$$

$$y \sim N(\mu_y, \sigma^2_y) \tag{G.9}$$

Strictly speaking, this is not the case. In order to perform hypothesis testing and to calculate appropriate p-values for a fitted linear model, the dependent variable(s) must be normally-distributed. This is not, however, required to fit models using least-squares regression or for the empirical evaluation of models using RMSE/CMSE as I do here. For more discussion on the role of distributional assumptions in linear models see the discussion of *Generalized Linear Models* in [34, 141].

APPENDIX H

INDUCED MODEL DETAILS

This appendix contains detailed tables showing the induced models produced during the feature induction process.

Table H1: Graph Grade Induced linear models for essay grades Model description and RMSE/CMSE scores.

Grade	RMSE	CMSE
E.01 (RQ-Open)	0.3106	0.3592
E.01 $\sim$ G.01		

E.04 (Hyp-Testable)	0.2322	0.2828
E.04 $\sim$ G.04		

E.07 (Cite-Reasons)	0.2477	0.2701
E.07 $\sim$ G.07 + G.05 + G.10 + G.12 + G.13		

E.10 (Hyp-Open)	0.334	0.333
E.10 $\sim$ G.10 + G.12 + G.06 + G.04 + G.13 + G.03 + G.14 + G.02 + G.05		

E.14 (Arg-Quality)	0.2062	0.2410
E.14 $\sim$ G.07 + G.05		

Table H2: Raw Feature Model RMSE and CMSE scores for *E.01 (RQ-Quality)*

Dataset	# Predictors	RMSE	CMSE
Intuitive-NoP	11	0.3108	0.3614
$E.01 \sim Rule\_R02c\_NonHypoRQwoOut\_log$			
Intuitive	18	0.3014	0.3396
$E.01 \sim Order\_Elt\_comparison\_bin$			
Intervention	47	0.2924	0.3248
$E.01 \sim Order\_Elt\_comparison\_bin$ $+Rule\_R01nb\_NoClaim\_bin$ $+Rule\_R01pb\_HasClaim\_log$ $+Rule\_R04a\_EmptyNodeFields$ $+Order\_OverlappingNodes$ $+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$			
Total	77	0.2902	0.3199
$E.01 \sim Order\_Elt\_comparison\_bin$ $+Rule\_R01nb\_NoClaim\_bin$ $+Order\_Elt\_claim\_log$ $+Rule\_R04a\_EmptyNodeFields$ $+Order\_MinDegree$ $+Rule\_R02c\_NonHypoRQwoOut\_log$ $+Rule\_R13\_DisjointSubgraphs\_log$ $+Rule\_R10d\_Hypothesis\_Comp\_bin$ $+Rule\_R02a\_NonHypowoOut\_log$			

Table H3: Raw Feature Model RMSE and CMSE scores for *E.04 (Hyp-Testable)*

Dataset	# Predictors	RMSE	CMSE
Intuitive-NoP	11	0.219	0.2726
$E.04 \sim Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$			
Intuitive	18	0.2194	0.2747
$E.04 \sim Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$ $+Rule\_R01pb\_HasClaim\_log$			
Intervention	47	0.2099	0.2411
$E.04 \sim Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$ $+Rule\_R10c\_Claim\_Comp\_bin$ $+Rule\_R05\_HypoSupportsCite$ $+Rule\_R01pa\_HasHypothesis\_log$ $+Rule\_R08\_Unopp\_Hypo\_log$			$+Rule\_R11ub\_Undef\_Unfounded\_Claim\_log$ $+Rule\_R05a\_HypoOpposesCite\_bin$ $+Rule\_R08\_Unsupp\_Hypo\_log$ $+Rule\_R11\_Ungrounded\_Hypo\_Claim\_log$ $+Rule\_R01nb\_NoClaim\_bin$
Total	77	0.207	0.2492
$E.04 \sim Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$ $+Rule\_R10c\_Claim\_Comp\_bin$ $+Rule\_R03\_NonCitewoIn\_bin$			$+Rule\_R11ub\_Undef\_Unfounded\_Claim\_log$ $+Rule\_R05a\_HypoOpposesCite\_bin$ $+Order\_MinChildren\_IgnoreEmpty\_log$ $+Rule\_R01nd\_NoRQ\_bin$

Table H4: Raw Feature Model RMSE and CMSE scores for *E.07 (Cite-Reasons)*

Dataset	# Predictors	RMSE	CMSE
Intuitive-NoP	11	0.2642	0.2979
$E.07 \sim Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$ $+Rule\_R02a\_NonHypowoOut\_log$ $+Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R08\_Unsupp\_Hypo\_log$			
Intuitive	18	0.2489	0.2752
$E.07 \sim Rule\_R01pa\_HasHypothesis\_log$ $+Rule\_R02a\_NonHypowoOut\_log$ $+Rule\_R01pc\_HasCite\_log$ $+Rule\_R01pb\_HasClaim\_log$ $+Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$ $+Order\_Elt\_comparison\_bin$ $+Rule\_R02b\_NonHypoClaimwoOut\_log$ $+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$			
Intervention	47	0.241	0.2627
$E.07 \sim Rule\_R01pc\_HasCite\_log$ $+Order\_Elt\_comparison\_bin$ $Order\_OverlappingNodes$ $+Rule\_R08\_Unopp\_Hypo\_log$ $+Rule\_R02a\_NonHypowoOut\_log+$ $+Rule\_R11b\_Unfounded\_Claim\_log$ $+Rule\_R01pa\_HasHypothesis\_log$ $+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$			
Total	77	0.227	0.2401
$E.07 \sim Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$ $+Rule\_R02a\_NonHypowoOut\_log$ $+Order\_OverlappingNodes$ $+Rule\_R02b\_NonHypoClaimwoOut\_log$ $+Rule\_R03b\_NonCiteOrCurrwoIn\_bin$ $+Order\_Order\_log$ $+Order\_Elt\_comparison\_bin$ $+Order\_AvgDegree$ $+Rule\_R01nb\_NoClaim\_bin$			

Table H5: Raw Feature Model RMSE and CMSE scores for *E.10 (Hyp-Open)*

Dataset	# Predictors	RMSE	CMSE
Intuitive-NoP	11	0.358	0.3401
$E.10 \sim Rule\_R06a\_Curr\_Uncompared\_w\_Cite$ $+Rule\_R03\_NonCitewoIn$			
Intuitive	18	0.3228	0.3168
$E.10 \sim Order\_PairedCounterarg$ $+Rule\_R08\_Unsupp\_Hypo\_log$ $+Rule\_R01pc\_HasCite\_log$ $+Order\_Elt\_supports\_log$ $+Rule\_R02c\_NonHypoRQwoOut\_log$ $+Rule\_R01pb\_HasClaim\_log$			
Intervention	47	0.3235	0.323
$E.10 \sim Order\_PairedCounterarg$ $+Rule\_R01pc\_HasCite\_log$ $+Order\_Elt\_supports\_log$ $+Rule\_R07\_UncomparedOpp\_bin$ $+Rule\_R08\_Unsupp\_Hypo\_log$ $+Rule\_R05\_HypoSupportsCite$ $+Rule\_R13\_DisjointSubgraphs\_log$			
Total	77	0.3217	0.3279
$E.10 \sim Order\_PairedCounterarg$ $+Rule\_R08\_Unsupp\_Hypo\_log$ $+Rule\_R07u\_Undef\_UncomparedOpp\_log$ $+Rule\_R10c\_Claim\_Comp\_bin$ $+Rule\_R01na\_NoHypothesis\_bin$ $+Order\_Minus\_citation\_AvgFieldSentLen\_log$ $+Order\_MaxParents\_log$ $+Rule\_R01nc\_NoCite\_bin$ $+Order\_AvgChildren\_IgnoreEmpty\_log$ $+Order\_MaxParents\_IgnoreEmpty\_log$ $+Rule\_R10d\_Hypothesis\_Comp\_bin$ $+Rule\_R01nb\_NoClaim\_bin$ $+Rule\_R02b\_NonHypoClaimwoOut\_log$ $+Rule\_R11a\_Ungrounded\_Hypo\_log$ $+Order\_MinParents\_IgnoreEmpty\_log$ $+Rule\_R07ub\_Undef\_UndistinguishedOpp\_log$ $+Rule\_R13\_DisjointSubgraphs\_log$			

Table H6: Raw Feature Model RMSE and CMSE scores for *E.14 (Arg-Quality)*

Dataset	# Predictors	RMSE	CMSE
Intuitive-NoP	11	0.2148	0.2609
$E.14 \sim Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$			
Intuitive	18	0.2126	0.2547
$E.14 \sim Order\_Elt\_comparison\_bin$ $+Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$			
Intervention	47	0.2122	0.2473
$E.14 \sim Order\_Elt\_comparison\_bin$ $+Rule\_R10c\_Claim\_Comp\_bin$ $+Rule\_R01pa\_HasHypothesis\_log$ $+Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$ $+Rule\_R13\_DisjointSubgraphs\_log$ $+Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R02a\_NonHypowoOut\_log$ $+Rule\_R01nc\_NoCite\_bin$			
Total	77	0.2079	0.2415
$E.14 \sim Order\_Elt\_comparison\_bin$ $+Order\_Elt\_hypothesis\_log$ $+Order\_MaxChildren\_IgnoreEmpty\_log$ $+Rule\_R01na\_NoHypothesis\_bin$ $+Order\_MaxParents\_log$ $+Rule\_R12\_UndefinedCiteClaim\_bin$ $+Rule\_R04a\_EmptyNodeFields$ $+Order\_MinChildren\_IgnoreEmpty\_log$ $+Rule\_R10a\_Hypo\_or\_Claim\_Comp\_bin$ $+Order\_MaxParents\_IgnoreEmpty\_log$ $+Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$ $+Order\_PairedCounterarg$ $+Order\_MaxChildren\_log$ $+Rule\_R01pa\_HasHypothesis\_log$ $+Rule\_R10c\_Claim\_Comp\_bin$ $+Order\_AvgChildren\_IgnoreEmpty\_log$ $+Rule\_R08\_Unopp\_Hypo\_log$			

Table H7: Trimmed Feature Model RMSE and CMSE scores for *E.01 (RQ-Quality)*

Dataset	# Predictors	RMSE	CMSE
Intuitive-NoP	11	0.3104	0.3535
$E.01 \sim Rule\_R03\_NonCitewoIn$			
Intuitive	18	0.3014	0.3396
$E.01 \sim Order\_Elt\_comparison\_bin$			
Intervention	47	0.2907	0.3208
$E.01 \sim Order\_Elt\_comparison\_bin$ $+Rule\_R01nb\_NoClaim\_bin$ $+Rule\_R01pb\_HasClaim\_log$ $+Rule\_R04a\_EmptyNodeFields$ $+Rule\_R10c\_Claim\_Comp\_bin$ $+Rule\_R10d\_Hypothesis\_Comp\_bin$ $+Rule\_R02a\_NonHypowoOut\_log$ $+Rule\_R02b\_NonHypoClaimwoOut\_log$			
Total	77	0.2896	0.3257
$E.01 \sim Order\_Elt\_comparison\_bin$ $+Rule\_R01nb\_NoClaim\_bin$ $+Rule\_R04a\_EmptyNodeFields$ $+Order\_MaxParents\_IgnoreEmpty\_log$ $+Rule\_R10c\_Claim\_Comp\_bin$ $+Rule\_R11ub\_Undef\_Unfounded\_Claim\_log$ $+Order\_Elt\_claim\_log$ $+Rule\_R10d\_Hypothesis\_Comp\_bin$ $+Order\_AvgParents\_IgnoreEmpty\_log$			

Table H8: Trimmed Feature Model RMSE and CMSE scores for *E.04 (Hyp-Testable)*

Dataset	# Predictors	RMSE	CMSE
Intuitive-NoP	11	0.2202	0.2747
$E.04 \sim Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R01n\_NoCurrstudy\_bin$		$+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$	
Intuitive	18	0.2194	0.2747
$E.04 \sim Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R01pb\_HasClaim\_log$		$+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$	
Intervention	47	0.211	0.2557
$E.04 \sim Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$ $+Rule\_R10c\_Claim\_Comp\_bin$ $+Order\_Elt\_supports\_log$ $+Rule\_R07u\_Undef\_UncomparedOpp\_log$		$+Rule\_R11ub\_Undef\_Unfounded\_Claim\_log$ $+Rule\_R05a\_HypoOpposesCite\_bin$ $+Rule\_R02\_NonCurrStudywoOut\_log$ $+Rule\_R08\_Unopp\_Hypo\_log$ $+Rule\_R01nc\_NoCite\_bin$	
Total	77	0.2119	0.254
$E.04 \sim Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$ $+Rule\_R02\_NonCurrStudywoOut\_log$ $+Order\_ChainedArgNodes\_log$		$+Rule\_R11ub\_Undef\_Unfounded\_Claim\_log$ $+Rule\_R05\_HypoSupportsCite$ $+Order\_MaxChildren\_IgnoreEmpty\_log$	

Table H9: Trimmed Feature Model RMSE and CMSE scores for *E.07 (Cite-Reasons)*

Dataset	# Predictors	RMSE	CMSE
Intuitive-NoP	11	0.2667	0.3042
$E.07 \sim Rule\_R02a\_NonHypowoOut\_log$ $+Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R02b\_NonHypoClaimwoOut\_log$			
Intuitive	18	0.2477	0.2736
$E.07 \sim Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$ $+Rule\_R01pa\_HasHypothesis\_log$ $+Order\_Elt\_comparison\_bin$ $+Rule\_R02a\_NonHypowoOut\_log$ $+Rule\_R01pc\_HasCite\_log$ $+Rule\_R01pb\_HasClaim\_log$ $+Rule\_R02b\_NonHypoClaimwoOut\_log$			
Intervention	47	0.2371	0.262
$E.07 \sim Rule\_R01pa\_HasHypothesis\_log$ $+Rule\_R02a\_NonHypowoOut\_log$ $+Order\_Elt\_comparison\_bin$ $+Rule\_R06\_Cite\_Uncompared\_w\_Curr\_log$ $+Order\_OverlappingNodes$ $+Rule\_R11b\_Unfounded\_Claim\_log$ $+Rule\_R08\_Unopp\_Hypo\_log$ $+Rule\_R01pb\_HasClaim\_log$ $+Rule\_R01nb\_NoClaim\_bin$ $+Rule\_R02b\_NonHypoClaimwoOut\_log$ $+Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$			
Total	77	0.2434	0.26
$E.07 \sim Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$ $+Order\_Elt\_hypothesis\_log$ $+Rule\_R08\_Unopp\_Hypo\_log$ $+Order\_OverlappingNodes$ $+Rule\_R06\_Cite\_Uncompared\_w\_Curr\_log$ $+Order\_Elt\_comparison\_bin$ $+Order\_MaxParents\_IgnoreEmpty\_log$ $+Order\_MinChildren\_IgnoreEmpty\_log$ $+Rule\_R11u\_Undef\_Ungrounded\_Hypo\_Claim\_log$ $+Order\_MaxFieldSentLen\_log$ $+Rule\_R01nc\_NoCite\_bin$ $+Rule\_R10c\_Claim\_Comp\_bin$ $+Rule\_R04a\_EmptyNodeFields$ $+Rule\_R02a\_NonHypowoOut\_log$ $+Rule\_R02b\_NonHypoClaimwoOut\_log$			

Table H10: Trimmed Feature Model RMSE and CMSE scores for *E.10 (Hyp-Open)*

Dataset	# Predictors	RMSE	CMSE
Intuitive-NoP	11	0.3582	0.3401
$E.10 \sim Rule\_R03\_NonCitewoIn$ $+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$			
Intuitive	18	0.3288	0.3284
$E.10 \sim Order\_PairedCounterarg$ $+Rule\_R01pc\_HasCite\_log$ $+Rule\_R08\_Unsupp\_Hypo\_log$ $+Order\_Elt\_supports\_log$ $+Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R01pa\_HasHypothesis\_log$ $+Rule\_R03b\_NonCiteOrCurrwoIn$ $+Rule\_R02a\_NonHypowoOut\_log$			
Intervention	47	0.3229	0.3121
$E.10 \sim Order\_PairedCounterarg$ $+Rule\_R08\_Unsupp\_Hypo\_log$ $+Rule\_R06a\_Curr\_Uncompared\_w\_Cite$ $+Rule\_R01p\_HasCurrstudy\_log$ $+Rule\_R03b\_NonCiteOrCurrwoIn\_bin$ $+Rule\_R05\_HypoSupportsCite$ $+Rule\_R10c\_Claim\_Comp\_bin$ $+Rule\_R03\_NonCitewoIn\_bin$ $+Rule\_R05a\_HypoOpposesCite\_bin$			
Total	77	0.3159	0.3282
$E.10 \sim Order\_PairedCounterarg$ $+Order\_AvgChildren\_log$ $+Order\_MaxChildren\_IgnoreEmpty\_log$ $+Order\_MinParents\_IgnoreEmpty\_log$ $+Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R05b\_HypoToCite\_log$ $+Rule\_R01nc\_NoCite\_bin$ $+Order\_Minus\_citation\_MinFieldSentLen\_bin$ $+Order\_MinFieldSentLen$ $+Rule\_R07\_UncomparedOpp\_bin$ $+Rule\_R01nb\_NoClaim\_bin$ $+Rule\_R11\_Ungrounded\_Hypo\_Claim\_log$ $+Rule\_R10c\_Claim\_Comp\_bin$ $+Rule\_R10a\_Hypo\_or\_Claim\_Comp\_bin$ $+Order\_MaxParents\_IgnoreEmpty\_log$ $+Rule\_R11b\_Unfounded\_Claim\_log$ $+Order\_MinDegree$ $+Rule\_R03b\_NonCiteOrCurrwoIn\_bin$			

Table H11: Trimmed Feature Model RMSE and CMSE scores for *E.14 (Arg-Quality)*

Dataset	# Predictors	RMSE	CMSE
Intuitive-NoP	11	0.2148	0.2609
$E.14 \sim Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$			
Intuitive	18	0.2143	0.2569
$E.14 \sim Order\_Elt\_comparison\_bin$ $+Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$ $+Rule\_R03b\_NonCiteOrCurrwoIn$			
Intervention	47	0.2126	0.251
$E.14 \sim Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$ $+Rule\_R01na\_NoHypothesis\_bin$ $+Order\_Elt\_comparison\_bin$ $+Rule\_R13\_DisjointSubgraphs\_log$ $+Rule\_R08\_Unopp\_Hypo\_log$ $+Rule\_R11ub\_Undef\_Unfounded\_Claim\_log$ $+Rule\_R02a\_NonHypowoOut\_log$			
Total	77	0.2065	0.2369
$E.14 \sim Order\_Elt\_comparison\_bin$ $+Rule\_R01na\_NoHypothesis\_bin$ $+Rule\_R11ua\_Undef\_Ungrounded\_Hypo\_bin$ $+Order\_MaxParents\_IgnoreEmpty\_log$ $+Rule\_R13\_DisjointSubgraphs\_log$ $+Rule\_R01nc\_NoCite\_bin$ $+Order\_MinChildren\_IgnoreEmpty\_log$ $+Order\_MinDegree$ $+Rule\_R10c\_Claim\_Comp\_bin$			

BIBLIOGRAPHY

[1] A. Afifi and V. Clark. *Computer-aided Multivariate Analysis*. Wadsworth, Belmont California, 1984.

[2] Ruggero J. Aldisert. *Logic for Lawyers a Guide to Clear Legal Thinking*. Clark Boardman Company Ltd., 1989.

[3] Vincent Aleven, Amy Ogan, Octav Popescu, Cristen Torrey, and Kenneth R. Koedinger. Evaluating the effectiveness of a tutorial dialogue system for self-explanation. In James C. Lester, Rosa Maria Vicari, and Fábio Paraguaçu, editors, *Intelligent Tutoring Systems*, volume 3220 of *Lecture Notes in Computer Science*, pages 443–454. Springer, 2004.

[4] Kevin Ashley. Teaching a process model of legal argument with hypotheticals. *Artificial Intelligence and Law*, 17:321–370, 2009.

[5] Kevin D. Ashley and Collin F. Lynch. Are argument diagrams of hypothetical reasoning diagnostic? In Mary Kay Stein and Linda Kucan, editors, *Instructional Explanations in the Disciplines.*, pages 171–188. New York: Springer, 2010.

[6] Kevin D. Ashley, Collin F. Lynch, Niels Pinkwart, and Vincent Aleven. A process model of legal argument with hypotheticals. In Enrico Francesconi, Giovanni Sartor, and Daniela Tiscornia, editors, *JURIX*, volume 189 of *Frontiers in Artificial Intelligence and Applications*, pages 1–10. IOS Press, 2008.

[7] R.S.J.d. Baker, A.T. Corbett, I. Roll, and K.R. Koedinger. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, 18(3):287–314, 2008.

[8] Wolfgang Banzhaf, Peter Nordin, Robert E. Keller, and Frank D. Francone. *Genetic Programming; an Introduction*. Morgan Kaufmann Publishers; San Francisco, 1998.

[9] John R. Beech. *How to Write in Psychology: A Student Guide*. Wiley-Blackwell: Malden Massachusetts, 2009.

[10] H. Blok. Estimating the reliability, validity and invalidity of essay ratings. *The Journal of Educational Measurement*, 22(1):41–52, 1985.

[11] Béla Bollobás. *Modern Graph Theory.* Springer Science+Business Media Inc. New York, New York, U.S.A., 1998.

[12] S.W. van den Braak, H. van Oostendorp, H. Prakken, and G.A.W. Vreeswijk. A critical review of argument visualization tools: Do users become better reasoners? In F. Grasso, R. Kibble, and C. Reed, editors, *Workshop Notes of the ECAI-2006 Workshop on Computational Models of Natural Argument (CMNA VI)*, pages 67–75, Riva del Garda, Italy, 2006.

[13] J. Brenner. Die nutzung des lasad-systems zur unterstützung juristischer argumentation. Diploma thesis, Clausthal University of Technology, Department of Informatics, 2010.

[14] Chad S. Carr. Using computer supported argument visualization to teach legal argumentation. pages 75–96. Springer-Verlag, London, UK, 2003.

[15] Michelene T. H. Chi, Nicholas de Leeuw, Mei-Hung Chiu, and Christian LaVancher. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477, 1994.

[16] Michelene T. H. Chi and Randi Daimon Koeske. Network representation of a child's dinosaur knowledge. *Developmental Psychology*, 1:29–39, 1983.

[17] Min Chi, Kurt VanLehn, Diane J. Litman, and Pamela W. Jordan. Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Model. User-Adapt. Interact.*, 21(1-2):137–180, 2011.

[18] Kwangsu Cho, Tingting Rachel Chung, William R. King, and Christian D. Schunn. Peer-based computer-supported knowledge refinement: an empirical investigation. *Commun. ACM*, 51(3):83–88, 2008.

[19] Evi Chryssafidou. Dialectic: Enhancing essay writing skills with computer-supported formulation of argumentation. In C. Stephanidis, editor, *Proceedings of the ERCIMWG UI4ALL one-day joint workshop with i3 Spring Days 2000 on "Interactive Learning Environments for ChildreN"*, March 2000.

[20] Evi Chryssafidou and Mike Sharples. Computer-supported planning of essay argument structure. In *Proceedings of the 5th International Conference of Argumentation*, June 2002.

[21] Christina Conati and Kurt VanLehn. Toward computer-based support and cristina conati and kurt vanlehn. *International Journal of Artificial Intelligence in Education*, 11:398–415, 2000.

[22] Jeff Conklin. *Dialogue Mapping: Building Shared Understanding of Wicked Problems.* Wiley, Chichester, England, 2006.

[23] Diane J. Cook and Lawrence B. Holder, editors. *Mining Graph Data*. John Wiley & Sons, 2006.

[24] Diane J. Cook, Lawrence B. Holder, and Nikhil Ketkar. Unsupervised and supervised pattern learning in graph data. In Cook and Holder [23], chapter 7, pages 159–81.

[25] Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition, 2001.

[26] J. G. Crammond. An analysis of argument structures in expert and student persuasive writing, 1997.

[27] J. G. Crammond. The uses and complexity of argument structures in expert and student persuasive writing. *Written Communication*, 15(2):230–268, 1998.

[28] Scott A. Crossley, Laura K. Varner, Rod D. Roscoe, and Danielle S. McNamara. Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In Lane et al. [62], pages 269–278.

[29] Peter Dalgaard. *Introductory Statistics with R*. Springer Verlag New York Inc., 2002.

[30] Semire Dikli. An overview of automated scoring of essays. *JTLA Journal of Technology Learning and Assessment*, 5(1), 2006.

[31] Vania Dimitrova, Riichiro Mizoguchi, Benedict du Boulay, and Arthur C. Graesser, editors. *Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling, Proceedings of the 14th International Conference on Artificial Intelligence in Education, AIED 2009, July 6-10, 2009, Brighton, UK*, volume 200 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2009.

[32] Matthew W. Easterday, Vincent Aleven, and Richard Scheines. 'tis better to construct than to receive? the effects of diagram tools on causal reasoning. In Luckin et al. [66], pages 93–100.

[33] Matthew W. Easterday, Vincent Aleven, Richard Scheines, and Sharon M. Carver. Constructing causal diagrams to learn deliberation. *International Journal of Artificial Intelligence in Education*, 19(4):425–445, 2009.

[34] John Fox. *Applied Regression Analysis and Generalized Linear Models*. Sage Publishing, Thousand Oaks California, 2008.

[35] Ilya Goldin. A focus on content: The use of rubrics in peer review to guide students and instructors., 2011.

[36] Robert A. Gordon. Issues in multiple regression. *American Journal of Sociology*, 73(5):592–616, March 1968.

[37] Thomas F. Gordon, Henry Prakken, and Douglas Walton. The carneades model of argument and burden of proof. *Artif. Intell.*, 171(10-15):875–896, 2007.

[38] A.C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. Coh-metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, 36:193–202, 2004.

[39] Nancy L. Green. Towards automated analysis of student arguments. In Lane et al. [62], pages 591–594.

[40] Laurence Greene. *Writing in the Life Sciences*. Oxford University Press: New York, 2009.

[41] Sebastian Gross, Bassam Mokbel, Barbara Hammer, and Niels Pinkwart. Feedback provision strategies in intelligent tutoring systems based on clustered solution spaces. In Jörg Desel, Jörg M. Haake, and Christian Spannagel, editors, *DeLFI*, volume 207 of *LNI*, pages 27–38. GI, 2012.

[42] Sebastian Gross, Bassam Mokbel, Barbara Hammer, and Niels Pinkwart. Towards providing feedback to students in absence of formalized domain models. In Lane et al. [62], pages 644–648.

[43] G. H. Hardy. *A Mathematician's Apology*. Cambridge University Press, 1996.

[44] Mara Harrell. Assessing the efficacy of argument diagramming to teach critical thinking skills in introduction to philosophy. *Inquiry*, 27(2):31–38, 2012.

[45] Maralee Harrell. Argument diagramming and critical thinking in introductory philosophy. *Higher Education Research & Development*, 30(3):371–385, 2011.

[46] Maralee Harrell and Danielle Wetzel. Improving first-year writing using argument diagramming. In Markus Knauff, Natalie Sebanz, Michael Pauen, and Ipke Wachsmuth, editors, *Proceedings of the* 35^{th} *Annual Conference of the Cognitive Science Society*, pages 2488–2493. The Cognitive Science Society: Austin Texas, U.S.A.

[47] Andreas Harrer, Rakheli Hever, and Sabrina Ziebarth. Empowering researchers to detect interaction patterns in e-collaboration. In Luckin et al. [66], pages 503–510.

[48] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag New York, 2001.

[49] Theil Henri. *Economic Forecasts and Policy*. Holland, Amsterdam: North, 1961.

[50] Ken Hyland. *Second Language Writing*. Cambridge Language Education. Cambridge University Press, 2003.

[51] Istvan Jonyer. Graph grammar learning. In Cook and Holder [23], chapter 8, pages 183–201.

[52] Istvan Jonyer, Lawrence B. Holder, and Diane J. Cook. Mdl-based context-free graph grammar induction and applications. *International Journal on Artificial Intelligence Tools*, 13(1):65–79, 2004.

[53] Istwan Jonyer, Lawrence B. Holder, and Diane J. Cook. Concept formation using graph grammars. In *Proceedings of the KDD Workshop on Multi-Relational Data Mining*, 2002.

[54] Luke Keele. *Semiparametric Regression for the Social Sciences*. John Wiley & Sons Ltd: The Atrium, Southern Gate, Chichester, 2008.

[55] Nikhil S. Ketkar, Lawrence B. Holder, and Diane J. Cook. Empirical comparison of graph classification algorithms. In *CIDM*, pages 259–266. IEEE, 2009.

[56] Donald E. Knuth. *The Art of Computer Programming: Sorting and Searching*, volume 3. Addison-Wesley, 2^{nd} edition, 1998.

[57] Donald E. Knuth. *The Art of Computer Programming: Combinatorial Algorithms, Part 1*, volume 4A. Addison-Wesley, 1^{st} edition, 2011.

[58] Jacek P. Kukluk, Lawrence B. Holder, and Diane J. Cook. Inference of edge replacement graph grammars. *International Journal on Artificial Intelligence Tools*, 17(3):539–554, 2008.

[59] Jacek P. Kukluk, Lawrence B. Holder, and Diane J. Cook. Inferring graph grammars by detecting overlap in frequent subgraphs. *Applied Mathematics and Computer Science*, 18(2):241–250, 2008.

[60] Michihiro Kuramochi and George Karypis. Finding topological frequent patterns from graph datasets. In Cook and Holder [23], chapter 6, pages 117–58.

[61] Imre Lakatos. *Proofs and Refutations*. Cambridge University Press, 1976.

[62] H. Chad Lane, Kalina Yacef, Jack Mostow, and Philip I. Pavlik, editors. *Artificial Intelligence in Education - 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings*, volume 7926 of *Lecture Notes in Computer Science*. Springer, 2013.

[63] K. N. Llewellyn. *The Bramble Brush; On our Law and it's study*. Oceana Publications Inc, Dobbs Ferry, New York, 1951.

[64] Frank Loll and Niels Pinkwart. Guiding the process of argumentation: the effects of ontology and collaboration. In *Proceedings of CSCL 2011 (in press)*, 2011.

[65] Frank Loll and Niels Pinkwart. Lasad: Flexible representations for computer-based collaborative argumentation. *Int. J. Hum.-Comput. Stud.*, 71(1):91–109, 2013.

[66] Rosemary Luckin, Kenneth R. Koedinger, and Jim E. Greer, editors. *Artificial Intelligence in Education, Building Technology Rich Learning Contexts That Work, Proceedings of the 13th International Conference on Artificial Intelligence in Education, AIED 2007, July 9-13, 2007, Los Angeles, California, USA*, volume 158 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2007.

[67] Collin F. Lynch and Kevin D. Ashley. Modeling student arguments in research reports. In Vince Duffy, editor, *Advances in Applied Human Modeling and Simulation Advances in Human Factors and Ergonomics 2012 14 Volume Set: Proceedings of the 4th AHFE Conference 21-25 July 2012*. CRC Press: Taylor and Francis Group, July 2012.

[68] Collin F. Lynch, Kevin D. Ashley, and Mohammad H. Falakmassir. Comparing argument diagrams. In Burkhard Schäfer, editor, *Legal Knowledge and Information Systems - JURIX 2012: The Twenty-Fifth Annual Conference, University of Amsterdam, The Netherlands, 17-19 December 2012*, volume 250 of *Frontiers in Artificial Intelligence and Applications*, pages 81–90. IOS Press, 2012.

[69] Collin F. Lynch, Kevin D. Ashley, Niels Pinkwart, and Vincent Aleven. Argument diagramming as focusing device: does it scaffold reading? In Vincent Aleven, Kevin D. Ashley, Collin F. Lynch, and Niels Pinkwart, editors, *Proceedings of the Workshop on AIED Applications for Ill-Defined Domains*, pages 51 – 60, 2007. Held at the 13th International Conference on Artificial Intelligence in Education.

[70] Collin F. Lynch, Kevin D. Ashley, Niels Pinkwart, and Vincent Aleven. Argument graph classification with genetic programming and c4.5. In Ryan Shaun Joazeiro de Baker, Tiffany Barnes, and Joseph E. Beck, editors, *EDM*, pages 137–146. www.educationaldatamining.org, 2008.

[71] Collin F. Lynch, Kevin D. Ashley, Niels Pinkwart, and Vincent Aleven. Argument diagramming and diagnostic reliability. In Guido Governatori, editor, *JURIX 2009: The Twenty-Second Annual Conference on Legal Knowledge and Information Systems, Rotterdam, The Netherlands, 16-18 December 2009*, volume 205 of *Frontiers in Artificial Intelligence and Applications*, pages 106–115. IOS Press, 2009.

[72] Collin F. Lynch, Kevin D. Ashley, Niels Pinkwart, and Vincent Aleven. Computational argument as a diagnostic tool: the role of reliability. Technical report, AAI Press, 2009. Presented at the AAAI Fall Symposium on "The Uses of Computational Argumentation" AAAI Report Number: FS-09-06.

[73] Collin F. Lynch, Kevin D. Ashley, Niels Pinkwart, and Vincent Aleven. Concepts, structures, and goals: Redefining ill-definedness. *International Journal of Artificial Intelligence in Education*, 19(3):253–266, 2009.

[74] Collin F. Lynch, Kevin D. Ashley, Niels Pinkwart, and Vincent Aleven. Ill-defined domains and adaptive tutoring technologies. In Paula J. Durlach and Alan M. Lesgold, editors, *Adaptive Technologies for Training and Education.*, chapter 9, pages 179–203. Cambridge, UK: Cambridge University Press., 2012.

[75] Collin F. Lynch, Niels Pinkwart, Kevin D. Ashley, and Vincent Aleven. What do argument diagrams tell us about students' aptitude or experience? a statistical analysis in an ill-defined domain. In V. Aleven, K. D. Ashley, Collin F. Lynch, and N. Pinkwart, editors, *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 9th International Conference on Intelligent Tutoring Systems (ITS)*, pages 62 – 73, Montreal, Canada, 2008.

[76] Collin F. Lynch, Niels Pinkwart, Kevin D. Ashley, and Vincent Aleven. What do argument diagrams tell us about students' aptitude or experience? a statistical analysis in an ill-defined domain. In Vincent Aleven, Kevin D. Ashley, Collin F. Lynch, and Niels Pinkwart, editors, *Assessment and Feedback in Ill-Defined Domains: Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains.*, pages 56–67, 2008. Held at the 9th International Conference on Intelligent Tutoring Systems.

[77] Scott M. Lynch. Multicollinearity, 2003. [Online; accessed 11-October-2013].

[78] Sherry E. Marcus, Melanie Moy, and Thayne Coffman. Social network analysis. In Cook and Holder [23], chapter 17, pages 443–468.

[79] B. M. McLaren, O. Scheuer, and J. Mikšátko. Supporting collaborative learning and e-discussions using artificial intelligence techniques. *Submitted to: International Journal of Artificial Intelligence in Education*, January 2009.

[80] Bruce M. McLaren, Rupert Wegerif, Jan Miksatko, Oliver Scheuer, Marian Chamrada, and Nasser Mansour. Are your students working creatively together? automatically recognizing creative turns in student e-discussions. In Dimitrova et al. [31], pages 317–324.

[81] Jan Mikšátko and Bruce M. McLaren. What's in a cluster? automatically detecting interesting interactions in student e-discussions. In Woolf et al. [144], pages 333–342.

[82] Tom Mitchell. *Machine Learning*. McGraw Hill, 1997.

[83] Roger Nkambou, Engelbert Mephu Nguifo, and Philippe Fournier-Viger. Using knowledge discovery techniques to support tutoring in an ill-defined domain. In *ITS '08: Proceedings of the 9th international conference on Intelligent Tutoring Systems*, pages 395–405, Berlin, Heidelberg, 2008. Springer-Verlag.

[84] Takashi Okada. Mining from chemical graphs. In Cook and Holder [23], chapter 14, pages 347–379.

[85] Niels Pinkwart, Vincent Aleven, Kevin D. Ashley, and Collin F. Lynch. Toward legal argument instruction with graph grammars and collaborative filtering techniques. In et al. Ikeda, editor, *Proceedings ITS 2006*, pages 227–236. Berlin: Springer., 2006.

[86] Niels Pinkwart, Vincent Aleven, Kevin D. Ashley, and Collin F. Lynch. Evaluating legal argument instruction with graphical representations using largo. In Luckin et al. [66], pages 101–108.

[87] Niels Pinkwart, Vincent Aleven, Kevin D. Ashley, and Collin F. Lynch. Evaluating legal argument instruction with graphical representations using largo. In Luckin et al. [66], pages 101–108. Invited for submission to the International Journal of AI in Education.

[88] Niels Pinkwart, Vincent Aleven, Kevin D. Ashley, and Collin F. Lynch. Adaptive rückmeldungen im intelligenten tutorensystem largo. *E-Learning & Education*, 5, 2009.

[89] Niels Pinkwart, Kevin D. Ashley, Vincent Aleven, and Collin F. Lynch. Graph grammars: An its technology for diagram representations. In Wilson and Lane [139], pages 433–438.

[90] Niels Pinkwart, Kevin D. Ashley, Vincent Aleven, and Collin F. Lynch. Graph grammars: An its technology for diagram representations. In Wilson and Lane [139], pages 433–438.

[91] Niels Pinkwart, Kevin D. Ashley, Collin F. Lynch, and Vincent Aleven. Evaluating an intelligent tutoring system for making legal arguments with hypotheticals. *International Journal of Artificial Intelligence in Education*, 19(4):401–424, 2009.

[92] Niels Pinkwart, Collin F. Lynch, Kevin D. Ashley, and Vincent Aleven. Re-evaluating largo in the classroom: Are diagrams better than text for teaching argumentation skills? In Woolf et al. [144], pages 90–100.

[93] Niels Pinkwart, Collin F. Lynch, Kevin D. Ashley, and Vincent Aleven. Re-evaluating largo in the classroom: Are diagrams better than text for teaching argumentation skills? In Woolf et al. [144], pages 90–100.

[94] The Jess Project. JESS the rule engine for the java platform, 2013. [Online; accessed 11-11-2013].

[95] J. Ross Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers; San Francisco, 1993.

[96] Chris Reed, Douglas Walton, and Fabrizio Macagno. Argument diagramming in logic, law and artificial intelligence. *Knowledge Eng. Review*, 22(1):87–109, 2007.

[97] J. Rekers and Andy Schürr. Defining and parsing visual languages with layered graph grammars. *J. Vis. Lang. Comput.*, 8(1):27–55, 1997.

[98] Nicholas Rescher. *Hypothetical Reasoning.* Studies in Logic and The Foundations of Mathematics. North-Holland Publishing Company, Amsterdam, 1964.

[99] Rod D. Roscoe, Erica L. Snow, and Danielle S. McNamara. Feedback and revising in an intelligent tutoring system for writing strategies. In Lane et al. [62], pages 259–268.

[100] Patrick Royston. An extension of shapiro and wilk's w test for normality to large samples. *Applied Statistics*, 31:115–124, 1982.

[101] D. E. Rumelhart. Schemata: The building blocks of cognition. In R. J. Shapiro, B. C. Bruce, and W. F. Brewer, editors, *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education*, pages 33–58. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1980.

[102] Roger C. Schank and David B. Leake. Creativity and learning in a case-based explainer. *Artif. Intell.*, 40(1-3):353–385, 1989.

[103] O. Scheuer, S. Niebuhr, T. Dragon, B. M. McLaren, and N. Pinkwart. Adaptive support for graphical argumentation - the lasad approach. IEEE Learning Technology Newsletter 14(1), p. 8 - 11, 2012.

[104] Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce McLaren. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning*, 5:43–102, 2010. 10.1007/s11412-009-9080-x.

[105] Oliver Scheuer, Bruce McLaren, Frank Loll, and Niels Pinkwart. Automated analysis and feedback techniques to support argumentation: A survey. In Niels Pinkwart and Bruce M. McLaren, editors, *Educational Technologies for Teaching Argumentation Skills.* Bentham Science Publishers, 2012. (in press).

[106] Oliver Scheuer, Bruce M. McLaren, Maralee Harrell, and Armin Weinberger. Will structuring the collaboration of students improve their argumentation? In Gautam Biswas, Susan Bull, Judy Kay, and Antonija Mitrovic, editors, *AIED*, volume 6738 of *Lecture Notes in Computer Science*, pages 544–546. Springer, 2011.

[107] Oliver Scheuer, Bruce M. McLaren, Frank Loll, and Niels Pinkwart. An analysis and feedback infrastructure for argumentation learning systems. In Dimitrova et al. [31], pages 629–631.

[108] Mary Ann Schroeder, Janice Lander, and Stacey Levine-Silverman. Diagnosing and dealing with multicollinearity. *Western Journal of Nursing Research*, 12(2):175–187, 1990.

[109] Baruch B. Schwarz and Amnon Glassner. The role of floor control and of ontology in argumentative activities with discussion-based tools. *I. J. Computer-Supported Collaborative Learning*, 2(4):449–478, 2007.

[110] S. J. Buckingham Shum, A. MacLean, V. M. E. Bellotti, and N. V. Hammond. Graphical argumentation and design cognition. *Human-Computer Interaction*, 12(3):267300, 1997.

[111] Michael Sipser. *Introduction to the Theory of Computation.* PWS Publishing Company, San Francisco, 1997.

[112] Rand J. Spiro, Walter P. Vispoel, John G. Schmitz, Ala Samarapungavan, and A. E. Boerger. Knowledge acquisition for application: Cognitive flexibility and transfer in complex content domains. In Bruce K. Britton and Shawn M. Glynn, editors, *Executive Control Processes in Reading*, pages 177–199. Lawrence Earlbaum Associates, 1987.

[113] John C. Stamper, Michael Eagle, Tiffany Barnes, and Marvin J. Croy. Experimental evaluation of automatic hint generation for a logic tutor. *I. J. Artificial Intelligence in Education*, 22(1-2):3–17, 2013.

[114] Common Core Standards. Common core standards for english language arts & literacy in history/social studies, science, and technical subjects., 2013. [Online; accessed 01-2011].

[115] Dan Suthers. Representations for scaffolding collaborative inquiry on ill-structured problems. Technical report, University of Hawaii, 1998. Presented at the 1998 conference of the American Educational Research Association, April 1998, San Diego.

[116] Daniel D. Suthers. Representational guidance for collaborative inquiry. In *Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments*, page 2746. 2003.

[117] Daniel D. Suthers. Empirical studies of the value of conceptually explicit notations in collaborative learning. In Alexandra Okada, Simon Buckingham Shum, and Tony Sherborne, editors, *Knowledge Cartography*, pages 1–23. Springer Verlag, 2008.

[118] Olaf Tans. The fluidity of warrants: Using the toulmin model to analyse practical discourse. In David Hitchcock and Bart Verheij, editors, *Arguing on the Toulmin Model: New Essays in Argument Analysis and Evaluation. Series: Argumentation Library , Vol. 10*. Springer-Verlag, 2006.

[119] Owen Thomas, editor. *Walden and Civil Disobedience: Authoritative Texts, Background, Reviews and Essays in Criticism*. W.W. Norton & Company Inc. New York, 1966.

[120] S. E. Toulmin. *The uses of Argument*. Cambridge University Press, 1958.

[121] S. E. Toulmin, R. D. Rieke, and A. Janik. *An Introduction to Reasoning*. New York, London: MacMillan Publishers, 2nd edition, 1984.

[122] J. Gregory Trafton and Susan B. Trickett. Note-taking for self-explanation and problem solving. *Hum.-Comput. Interact.*, 16(1):1–38, 2001.

[123] Tim J. van Gelder. A reason!able approach to critical thinking. *Principal Matters: The Journal for Australasian, Secondary School Leaders*, pages 34–36, 2002.

[124] Mark Vorobej. *A Theory of Argument*. Cambridge, 2006.

[125] James F. Voss. Toulmin's model and the solving of ill-structured problems. In David Hitchcock and Bart Verheij, editors, *Arguing on the Toulmin Model: New Essays in Argument Analysis and Evaluation*, pages 303–311. Springer, Berlin, 2006.

[126] James F. Voss, Terry R. Greene, Timothy A. Post, and Barbara C. Penner. Problem solving skill in the social sciences. *The Psychology of Learning and Motivation*, 17:165 – 215, 1983.

[127] Erin Walker, Nikol Rummel, and Kenneth R. Koedinger. To tutor the tutor: Adaptive domain support for peer tutoring. In Woolf et al. [144], pages 626–635.

[128] Douglas N. Walton. *Informal Logic: A Handbook for Critical Argumentation*. Cambridge University Press, 1989.

[129] Wikipedia. Breuschpagan test — wikipedia, the free encyclopedia, 2013. [Online; accessed 10-November-2013].

[130] Wikipedia. Coefficient of determination — wikipedia, the free encyclopedia, 2013. [Online; accessed 27-February-2013].

[131] Wikipedia. Greedy algorithm — wikipedia, the free encyclopedia, 2013. [Online; accessed 22-December-2013].

[132] Wikipedia. Holmbonferroni method — wikipedia, the free encyclopedia, 2013. [Online; accessed 16-December-2013].

[133] Wikipedia. Linear regression — wikipedia, the free encyclopedia, 2013. [Online; accessed 19-October-2013].

[134] Wikipedia. Multiple correlation — wikipedia, the free encyclopedia, 2013. [Online; accessed 8-November-2013].

[135] Wikipedia. Root-mean-square deviation, 2013. [Online; accessed 24-September-2013].

[136] Wikipedia. Shapirowilk test — wikipedia, the free encyclopedia, 2013. [Online; accessed 6-November-2013].

[137] Wikipedia. Spearman's rank correlation coefficient — wikipedia, the free encyclopedia, 2013. [Online; accessed 27-February-2013].

[138] Wikipedia. Wilcoxon signed-rank test — wikipedia, the free encyclopedia, 2013. [Online; accessed 16-December-2013].

[139] David Wilson and H. Chad Lane, editors. *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference, May 15-17, 2008, Coconut Grove, Florida, USA*. AAAI Press, 2008.

[140] Christopher R. Wolfe. Argumentation across the curriculum. *Written Communication*, 28(2):193–219, 2011.

[141] Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall, 2006.

[142] Simon N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.

[143] Simon N. Wood. mgcv: Mixed gam computation vehicle with gcv/aic/reml smoothness estimation, 10 2013. [Online; accessed 10-26-2013].

[144] Beverly Park Woolf, Esma Aïmeur, Roger Nkambou, and Susanne P. Lajoie, editors. *Intelligent Tutoring Systems, 9th International Conference, ITS 2008, Montreal, Canada, June 23-27, 2008, Proceedings*, volume 5091 of *Lecture Notes in Computer Science*. Springer, 2008.

[145] Youjae Yi. On the evaluation of main effects in multiplicative regression models. *Journal of the Market Research Society*, 31(1):133–138, January 1989.

[146] Chang Hun You, Lawrence B. Holder, and Diane J. Cook. Graph-based data mining in dynamic networks: Empirical comparison of compression-based and frequency-based subgraph mining. In *ICDM Workshops*, pages 929–938. IEEE Computer Society, 2008.