Supporting Students' Research Writing in Psychology through Argument Diagramming

by

Brendan Jeffrey Barstow

B.A. Applied Psychology, Bryant University, 2014

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

Master of Science

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This thesis was presented

by

Brendan Jeffrey Barstow

It was defended on

April 8th, 2016

and approved by

Christian Schunn, Professor, Department of Psychology

Timothy Nokes-Malach, Professor, Department of Psychology

Benjamin Rottman, Associate Professor, Department of Psychology

Thesis Director: Christian Schunn, Professor, Department of Psychology

Supporting Students' Research Writing in Psychology Through Argument Diagramming

Brendan Jeffrey Barstow, M.S.

University of Pittsburgh, 2016

Arguing for the need for a scientific research study (i.e. writing an introduction to a research paper) poses significant challenges for students. When faced with these challenges, students often generate overly 'safe' studies, or replications, or in contrast include no strong support for their hypothesis. Additionally, instruction on argumentation has been slow to integrate into scientific education and discourse. This raises the question—how can we support novice scientists in generating and defending high quality hypotheses? A long history of research supports the affordances provided by spatial representations of complex information, particularly in the sciences. More recently, argument diagramming— the process of spatially representing an argument by its component parts and their relationships— has gained traction in instruction for philosophy, social studies, and law. However, its effectiveness for supporting students in science is relatively untested. Additionally, many of these studies have focused on basic contrasts between diagramming and no diagramming. The purpose of these studies was to test the effectiveness of argument diagrams for supporting students' research writing in psychology, and to learn how different diagram ontologies affect the benefits afforded by the activity. In the first study, three groups of undergraduate students (n=120) in research methods lab courses were given either no diagramming support, support with a domain-general ontology, or support with a domain-specific ontology to help them write a research paper introduction. Students given any diagramming support included more relevant citations and included more opposing citations in their papers. Students using the domain-general ontology included more supporting citations than those in the control and those using the domain-specific wrote more about the scientific validity of cited studies than the control, but these latter two effects only approached statistical significance. In the second study, two groups of undergraduate students in research methods lab courses (total n=182) were randomly assigned to psychology-specific diagramming support or no support. Those given diagramming support were more likely to argue for an appropriately 'risky' hypothesis and wrote more about the relevance and validity of cited studies. Some of these gains show evidence of transfer to a second paper written later in the course.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

This research was supported by National Science Foundation Grant IIS-1122504. Any opinions expressed in this work reflect those of the authors and do not necessarily reflect those of the National Science Foundation or the University of Pittsburgh. We would like to thank Diane Litman and Huy Nyugen for their help and feedback on this project.

1.0 INTRODUCTION

This work explores argument diagramming as a tool to support students in psychology writing across a series of two studies. Argument diagramming is the process of visually representing an argument by breaking it down into its component parts (e.g. hypothesis, study findings, counterarguments). Diagramming has been shown to support students in understanding and generating arguments in Philosophy and Social Studies (Griffin, Malone, & Kameenui, 1995; Harrell, 2011, 2012; 2013;), but its effectiveness in Science education is under-researched.

The study reported in Chapter 2 tests an argument diagramming intervention on three cohorts of students in undergraduate research methods courses. Hypothesis risk is measured by the inclusion of opposing evidence in student papers. Relevance was coded for by experts who rated the actual relevance of cited studies to a student's hypothesis. Validity was coded for by experts who rated the validity of cited studies. An additional goal of this work was to compare the affordances of a domain-general diagramming ontology versus a psychology-specific ontology. One cohort was given diagramming support with a domain-general ontology, one was supported by a psychology-specific ontology, and one was unsupported (control).

Reflections on this first study raised questions about the complicated nature of the ontologies and our conceptualizations of risk, relevance, and validity. The design of the study described in Chapter 2 also leaves multiple opportunities for improvement. These reflections led to a similar but evolved study, described in Chapter 3. In this study, our definition of risk was expanded to include not only opposition, but also uncertainty—an expansion supported by an analysis of risk writing in published psychology papers. Study one indicated that diagramming can help students bring in more relevant studies, but from an educational perspective it is more important that students are extracting the right bits of information from studies in order to make these judgments. In other words, our focus shifted from the outcome of writing in study one to the process for study two. Additionally, we developed a far simpler ontology for the study in Chapter 2. This decision was supported primarily by student feedback from the prior study. Study 2 used a randomized, controlled experimental design within one cohort of students to more rigorously test the effect of argument diagramming on students' writing.

The inclusion of both studies in this investigation supports a significantly stronger argument for the value of argument diagramming in the domain of science writing. The exploration of different ontologies in study one informed the single ontology chosen for study two, and this variety in ontology indicates more robust support for the core practice of diagramming. Additionally, the experimental design limitations of study one are complemented by the tighter controls and design of study two. Lastly, the expansion of our conception of risk, relevance, and validity between studies provides a more holistic understanding of the impacts of these interventions.

Chapter 4 bridges these two studies and discusses the conclusions that may be drawn from the evidence presented in Chapters 2 and 3. It will address connections of this work to theories of representation and other empirical work on argument diagramming. It will also address the practical implications for this work for psychology instruction and science instruction more generally. Finally, Chapter 4 will address new questions that this research raises and likely directions for the continuation of this line of work.

2.0 DO ONTOLOGICAL CHARACTERISTICS OF ARGUMENT DIAGRAMMING TOOLS IMPACT WRITING GAINS IN SCIENCE?

2.1 INTRODUCTION

Argumentation and argumentative writing are difficult skills for students to learn (Andrews, 1995; Andrews & Mitchell, 2001; Hahn & Oaksford, 2012, Kuhn, 2013), yet these are important skills in a wide variety of fields. Learning to argue means not only acquiring cognitive skills, but internalizing the social, epistemological, and metacognitive dimensions also involved in the production and evaluation of argument (Kuhn, Zillmer, Crowell, & Zavala, 2013).

Unfortunately, argumentative writing suffers from a dearth of practice opportunities in formal education—In high school, students may have only one or two opportunities per semester to write evidence-based essays (Kiuhara, Graham, & Hawken, 2009). This may be at least partially explained by the intensive demands of conventional writing instruction on teachers, which may involve multiple cycles of drafting and detailed feedback. Further, in lower and mid-ranked American colleges, students' writing skill shows little to no improvement over four years—a problem apparent to employers as well as researchers (Arum & Roksa, 2011). Existing instruction for argumentative writing tends to have misplaced emphasis on the *presentation* of arguments instead of their *generation* (Andrews, 1995; Andrews & Mitchell, 2001; Oostdam, de Glopper, & Eiting, 1994; Oostdam & Emelot, 1991).

Teaching and learning argumentation in science can pose unique difficulties to both instructors and students (Osborne, Simon, Christodoulou, Howell-Richardson, & Richardson, 2013). The breadth and depth of conceptual, procedural, and epistemic knowledge that many scientific arguments require can make their development and analysis both time-consuming and challenging. Scientific theories and scientific evidence are frequently complex on their own, and their integration into a coherent argument is especially complex. For example, a given research paper can have a range of findings—some findings may contradict a theory, and other findings may be just irrelevant. Scientific theories are frequently multi-faceted, with each facet requiring its own support. The integration of argumentation into scientific instruction does not appear to come naturally, likely requiring a significant investment into teachers' professional development to achieve (Osborne et. al, 2013).

One kind of scientific argument structure that is especially challenging to both develop and defend is the main argument for the research question(s) found in an introduction to a research paper. In contrast to typical dialogic argumentation where multiple sides of an argument must be explored but the goal is for one side to be definitively stronger, research seeks to clarify open questions, issues for which prior knowledge is not definitive. Thus, the writer must strike a balance slightly in favor of the arguments for, but maintain a certain (and even desirable) ambiguity. Novice writers may not know that science uses methods to resolve open questions, and that a literature review serves as an argument for a hypothesis rather than just a historical summary. In addition, novices may fail to include strong support for their hypothesis (Schwarz, Neuman, Gil, & Ilya, 2003) or include obvious or unsupported arguments. Intermediate writers may fail to include any reason to doubt their tested hypothesis (i.e., fail to note possible counter-evidence) (Nussbaum & Schraw, 2007).

Because of these issues in science writing, instructional tools can help students improve their argumentative writing while minimizing instructional burdens. One tool, the Science Writing Heuristic, for example, seeks to provide students with more opportunities to practice informal writing in science by developing a framework for students to reflect on and discuss course concepts (Keys, Hand, Prain, & Collins, 1999). These informal writing experiences appear to help students create richer representations of scientific concepts and enable them to respond more deeply to related test questions (Hand, Prain, & Wallace, 2002; Hand, Wallace, & Yang, 2004; Keys, Hand, Prain, & Collins, 1999). Although a useful instructional tool, the Science Writing Heuristic emphasizes writing to learn science rather than learning to write effective scientific arguments. Thus, the demand for a method to improve students' formal writing in science still remains.

In developing a solution to this problem, one question to ask is: what medium of representation is ideal for the problem domain? At the highest level of design, this means choosing between a text or spatial representation for the instructional tool. A prior relevant debate about effective representations covers this exact issue, determining the relative benefits of text versus spatial representations of instructional content. Early research examined differences in memory between the two, indicating first that differences in storage and/or retrieval leads to better recognition for pictures rather than sentences or words (Shepard, 1967). This finding was later refined to show that recognition *accuracy* was better for pictures, but that recognition *speed* was better for words (Standing, 1973), perhaps due to the difference in content complexity. Building on this work, Mandler and Ritchey (1977) tested individuals' memory for pictures after altering either their meaning or their non-meaningful details, finding that people tended to have better memory for the meaning of pictures rather than their details. Paivio (1986) elaborated on this and other findings with the proposal of his dual-coding theory; positing that information is represented in memory through both verbal *and* visual modes rather than abstract propositional representations. Thus, if we attend to visual information at storage and retrieval we may have better memory for meaning but sacrifice detail, whereas attending to verbal information provides the opposite tradeoff.

Larkin and Simon (1987) presented seminal work on differences in *reasoning* afforded by text versus spatial representations. First, diagrammatic representations can directly represent complex topological relationships among problem components, while sentential representations directly represent temporal relationships or hierarchical relationships (e.g., in an outline). Second, diagrams *can* be superior to sentential representations in problem solving by improving computational efficiency for some tasks. For example, diagrams can reduce the effort spent on searching for problem elements by grouping information by similar use. In the context of science, diagramming has long been understood to play an important role, from the role they played in Galileo's and Huygen's discoveries (Cheng, 1992; Cheng & Simon, 1992) to the central role they play in modern science (Novick, 2000; Trafton, Trickett, & Farilee, 2005).

There is also evidence indicating that diagramming can facilitate student learning in science and other domains. Students who diagrammed new material in social studies performed better on a follow-up retention task than those who did not (Griffin, Malone, & Kameenui, 1995), although this appears to be an effect of the diagram itself rather than the student's creation of it (Stull & Mayer, 2007). The process of diagramming is cognitively demanding and may temper benefits if not applied mindfully (Chang, Sung, & Chen, 2002), but this may only be an issue for younger students as college students in psychology show robust, long-term benefits of diagram creation (McCagg & Dansereau, 1991). In spite of the volume of research establishing these and other affordances of diagrams as a class of representation, much less research has focused on cognitive aspects of argument diagramming. In particular, how does the environment used to construct diagrams and the specific content contained within diagrams influence the benefits gained from their employment for argumentation? The focus of the present work is to explore these questions in the context of science writing.

A growing body of research supports the effectiveness of argument diagramming in the classroom, the specific form of diagramming under investigation here. Argument diagramming is the process of visually representing an argument by its component elements. In philosophy education, multiple studies indicate the power of argument diagramming for improving students' argument analysis skills (Harrell, 2008, 2011, 2012) as well as their ability to generate arguments that are more elaborate and cohesive (Harrell, 2013). Nussbaum and Schraw (2007) found that the practice of diagramming arguments enabled students to refute more counterarguments in their opinion writing, although there were tradeoffs in essay quality between argument diagramming and more traditional criteria instruction—possibly indicating a cost for this improvement.

There is also some indirect evidence supporting the use of diagramming for argumentative writing in science education. Recent modeling work has established a direct link between the quality of college students' diagrams and the resulting science writing, indicating that the coherence and complexity of a student's diagram can be used to predict the grade earned by the resulting essay (Lynch, Ashley, & Chi, 2014; Lynch, 2014). But it is not known whether the diagrams improve writing, or whether conceptual challenges revealed in students' diagrams are also found in students' writing.

A further open question relates to the choice of ontology. An ontology specifies the fundamental types of things or concepts that exist for purposes of constructing a particular kind of argument, and sets out the relations among them. The ontology used to represent an argument may differ significantly by discipline or assignment purposes. For example, a diagram of a research study could use *hypotheses, findings, studies*, and other science-specific node types, but one could also utilize a more generic ontology like Toulmin's (1958) which involves claims, warrants, and rebuttals. More general ontologies might be more useful for a wider range of writing and lend themselves more easily to knowledge transfer.

More specific ontologies, however, might better support student reasoning about the concepts found within a discipline or writing genre. For example, in psychology the concepts of a cited study's relevance and validity are particularly important. To properly judge a piece of evidence in relation to a hypothesis, one needs to know the similarity of their goals and methods (i.e., the study's relevance), and also the rigor of the cited study's methods (i.e., its validity). Including these domain-specific elements in a diagram ontology may be helpful for writing in psychology, but perhaps add complexity to how much must be learned at once. We are also curious how diagramming support may generalize to situations of more or less complexity. It is possible that diagramming is only helpful when students are being heavily challenged.

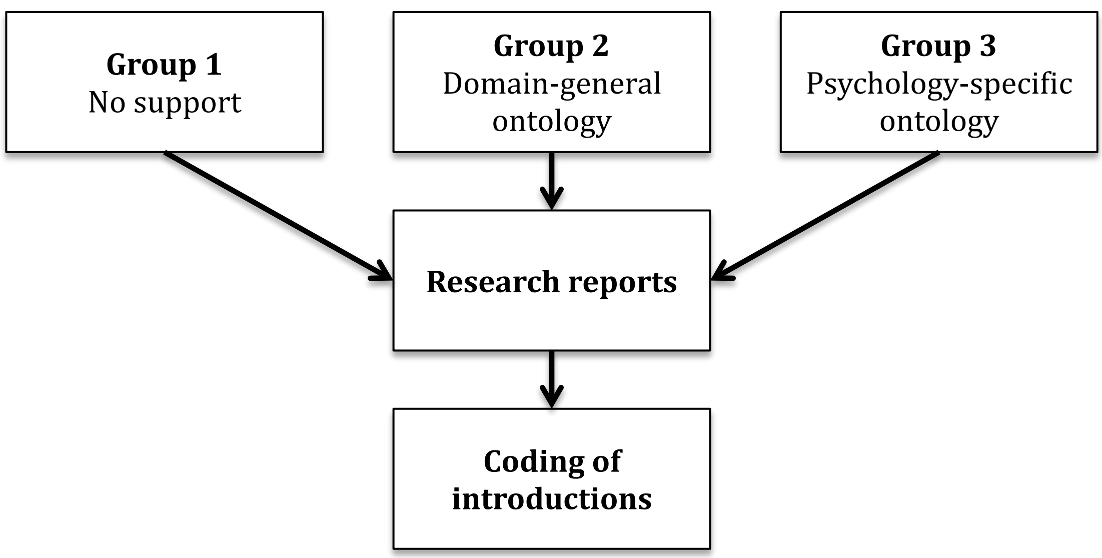
In sum, diagrams in general can have both memory and reasoning benefits, and argument diagrams in particular may help students think about the complex, multi-faceted relationships among hypotheses and prior findings needed to produce a strong argument for a hypothesis in a scientific paper introduction. The present study utilizes the online diagramming software LASAD (Loll & Pinkwart, 2013) to contribute to this growing research area first by determining the effect of a diagramming activity versus no diagramming on university students' writing quality of research paper introductions, and secondly to determine how the domain-specificity of ontological components in diagrams impact this effect.

In students' research paper introductions, we will examine the inclusion of opposing evidence (a common problem in college level writing (Knudson, 1992; Leitão, 2003; Stapleton, 2001; Perkins, Farady, & Bushey, 1991)) and the relevance and validity of citations (a specific challenge in research writing) as measures of writing quality.

We hypothesize that students who do any diagramming activity before writing will be more likely to include supporting and opposing evidence in their introductions. Additionally, we expect that students who construct diagrams, which explicitly prompt them to include information about the relevance and validity of citations, will include more of this information in their introductions.

We will test these hypotheses by analyzing introductions produced by students enrolled in research methods classes across three different semesters. The first group had an unaltered experience in the course to serve as a baseline for comparisons. The second group was given diagramming support for their papers in the form of a generic argument ontology. The third group was also given diagramming support, but in the form of a psychology-specific argument ontology. To address concerns about comparability across cohorts, the students across semesters are matched on general academic performance, and a similar pool of instructors enacted an otherwise shared and fixed curriculum. It was not logistically feasible to implement different interventions within a lab section in a given semester. Further, implementing the intervention across lab sections within a semester would have raised the risk of confounding teaching fellow (TF) quality and intervention effect given the smaller pool of TFs. Finally, the design of the third group's intervention arose from analyses of the strengths and weaknesses of the initial diagramming intervention, as is commonly done in design-based research.

Figure 1: Research design overview



2.2 METHODS

2.2.1 Instructional Context

The current study was conducted within a psychology department at a large, relatively selective public university in the United States. All undergraduate students at this university complete a composition course in their first year, which provides some training in argumentative writing. But due in part to the size of the university, many of the other early general education courses have large-enrollments and require relatively little writing. The entry-level courses for psychology majors are typically large lectures (150 to 300 students) with little-to-no writing and a focus on textbook readings, and thus there is little early exposure to disciplinary argumentation in written form. Students' first major introduction to disciplinary argumentation is in a psychology research methods course, the successful completion of which is required to officially declare a psychology major and enroll in advanced psychology courses. This course was the focus of our interventions and research.

The diagramming intervention was implemented in the laboratory (lab) sections of the Introduction to Psychology Research Methods course. The lecture was a large class that met once a week and was focused on theoretical research issues (e.g., validity, reliability, different research designs). The lab activities were worth 40% of the overall grade in the course and were designed to complement the lecture providing students with hands-on experience in conducting and writing about research. The lab sessions occurred in small sections of approximately 25 students that met twice a week with a standardized curriculum of weekly topics, in-class activities, and homework assignments.

There are typically 10 lab sections each semester, each run by a teaching fellow (TF), who most commonly was a graduate student in a psychology Ph.D. program. TFs met as a group on a weekly basis with a TF coordinator, who encouraged uniform implementation of the curriculum and grading across sections. Lab activities and homework centered on designing research projects, conducting literature searches, collecting and analyzing data, writing up the results of studies, and revising the written report.

Thus, students in this context are simultaneously learning about the nature of research in general, forms of research in psychology, written argumentation in research reports, psychology conventions for research writing, details of particular experimental paradigms, and statistical analyses. Such multi-leveled learning is typical in the behavioral sciences, and presents significant learning challenges for students.

Lab sections customized the hypotheses and designs of two studies, collected data, and then individual students wrote lab reports. A number of homework assignments were dedicated to helping students prepare the first lab report. Students wrote a first draft, received both rating and text-based feedback based on the rubric for the paper, and then revised their paper into a final draft.

The particular focus of the present research is the first draft written for their first study, the integrative moment at which students may experience the greatest struggles. To support students at this difficult moment, we created short activities involving argument diagramming and peer review of argument diagrams.

The paper assignment was a complete APA-style research report that students prepared based on a study that was designed as a class and conducted in small groups. Papers were approximately 10-12 double-spaced pages total with the introduction typically 1 to 2 pages long. As described in the grading rubric given to the students, the introduction of the lab report was to:

(a) Describe your research problem or question and say why it is important

(b) Contextualize your study and distinguish it from prior research

(c) Preview your study design

(d) Describe your hypotheses

(e) Provide a convincing justification for each hypothesis

All students read one common instructor-selected journal article on the topic, but then students had to find their own articles to include in their research report as supporting a hypothesis. Students in this class were encouraged to investigate simple hypotheses of the following form: Independent variables (IVs) X and Y cause changes in a dependent variable (DV) Z (possibly among population W). For instance, the hypothesis may concern the effects of gender and time of day on gratitude among coffee drinkers or the role of seat location and class size on student participation in class. In the first two cohorts, students were instructed to include two hypotheses in their paper (X and Y), but in the third cohort, students were given the option of including one or two hypotheses to study. All students were instructed to include both opposing and supporting studies as part of the justification for their hypotheses.

2.2.2 Participants

Control Group. Thirty essays were randomly selected from eight different lab sections from one fall semester of research methods classes that did not receive diagramming support. These essays were then coded and analyzed.

Domain-General Group. All students across nine different lab sections of the same course, also during the fall semester, but in the following year, were given diagramming support using a generic argument ontology. From this group, a stratified random sample of 30 essays was coded and analyzed.

Psychology-Specific Group. All students across nine different lab sections of the same research methods course taught during the fall semester of a subsequent year were given diagramming support using a psychology-specific argument ontology. Out of nine original lab sections, data from six sections (n=134) were retained. One TF did not attend training sessions and another TF, teaching two sections, fundamentally altered the writing assignment. From this set, a stratified random sample of 60 essays was coded and analyzed.

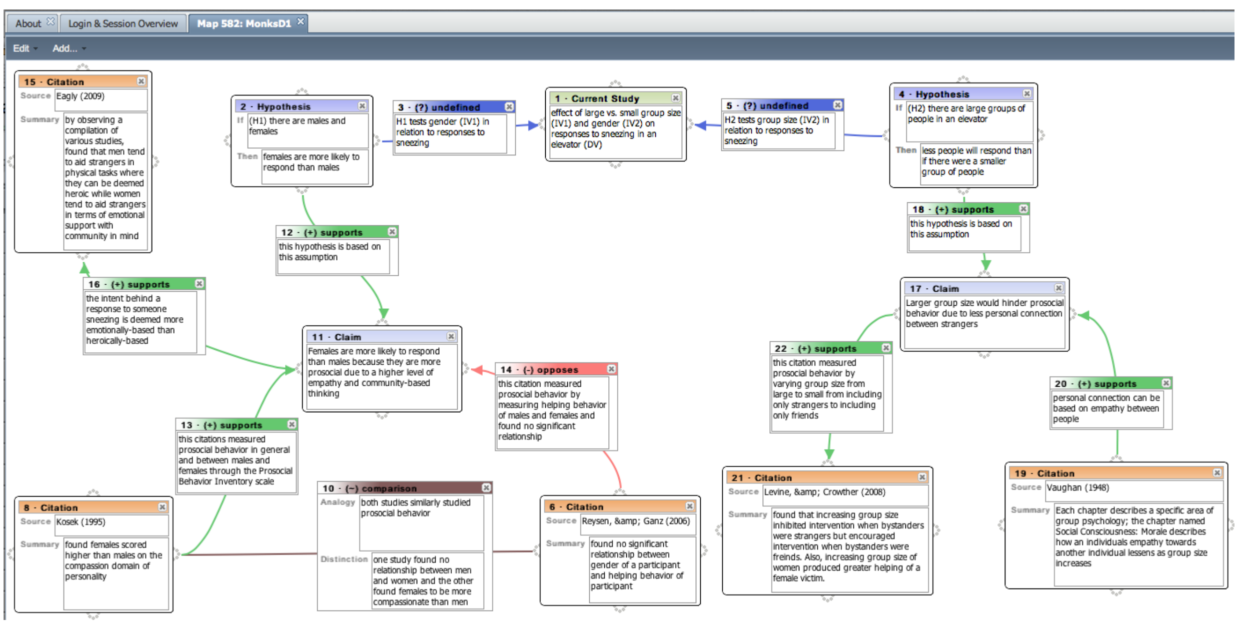
2.2.3 Argument Diagrams

***Domain-General Ontology*.** Our study in both iterations utilized LASAD, an online diagramming tool that allows users to create visual representations of arguments, including both the elements of an argument and their relationships. In LASAD, arguments are represented using a structured argument ontology of specific object and relationship types. Ontologies can be customized for each learning context. We customized the ontologies to represent the core elements of scientific argumentation that students were expected to include in the introductions to their laboratory reports. Specifically, our ontologies supported students in mapping out an argument for their hypotheses based on a review of studies and theories.

The first ontology used a more domain-general structure, with objects that were specific to science but relationships more generically cast in terms of supporting and opposing claims. Figure 2 presents an example student diagram. Note that LASAD, unlike many simpler diagramming tools, allows for detailed descriptions of relationships among nodes (i.e., links with multiple text fields), and thus may be particularly useful for reasoning about support and opposition relationships.

The node types of the first ontology are illustrated in Figure 2. *Hypothesis* nodes state the student's prediction of a data pattern in the form of a conditional (if/then) statement e.g., "If it is a busy time of day and the area in question has low traffic, then drivers will not obey the law and will not stop at the stop sign." *Current Study* nodes provide a general description of the study. *Supports* and *Opposes* nodes indicate the relationship of a study to a hypothesis node or a *Claim* node and explain why either relationship is indicated. *Claim* nodes provide reasoning for the hypothesis (analogous to Toulmin's *Warrant*) and are supported by *Citation* nodes (analogous to Toulmin's *Grounds/Evidence*). *Comparison* nodes compare two *Citation* nodes or a *Citation* and the *Current Study* node on the basis of study design and findings. The *Comparison* node is separated into 'analogies' and 'distinctions' (similarities and differences).

Figure 2: Example diagram from the domain-general ontology



***Psychology-Specific Ontology*.** For the second diagramming iteration, we sought to develop and test a LASAD ontology that was more domain-specific, especially including features particularly relevant and important for argumentation in psychology. *Finding* nodes replaced *Claim* nodes to represent the empirical findings supported by one or multiple studies because some students were confused about what exactly should go in a claim node, since the term 'claim' is borrowed from a language more common to argumentation in philosophy rather than psychology or science more generally. Note that multiple studies could relate to one research finding (e.g., two studies both support the finding that people are more likely to help when there are fewer bystanders), and one study could produce multiple findings (e.g., a single study finds that within larger groups people are both less likely to help and slower to help).

As before, students identified studies that supported or opposed more general findings (which in turn supported or opposed their hypothesis). Specific content was also added to these *Supports* and *Opposes* nodes in which students were now explicitly prompted to write about the relevance and validity of a citation used to support a finding or a finding used to support a hypothesis. In addition, for each study and finding, students rated how relevant it was to their hypothesis (close, medium, far, unsure), how valid it was (strong, medium, weak, unsure), and provided justification for both ratings. For the link between a study and a finding, relevance was defined as how strongly the study supported the finding (e.g. how large was the effect) and validity was determined by the methodological soundness of the study. For the link between a finding and a hypothesis, relevance was the amount of conceptual overlap between the finding and hypothesis (e.g., did they use similar independent and dependent variables) and validity was the overall validity of all the studies related to the finding.[1]

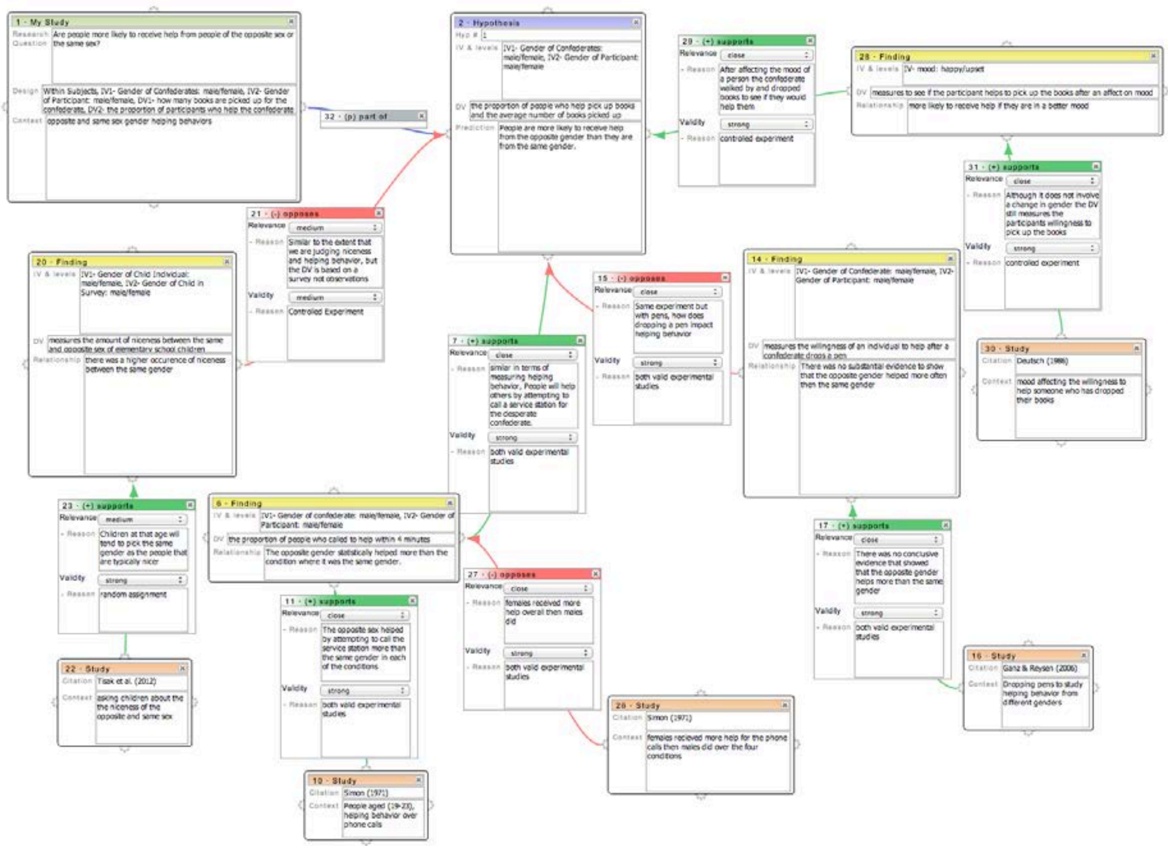
By following one thread of a student's argument diagram from *Current Study* to *Citation*, the nature of the ontology differences can be better understood. In the domain-general ontology (Figure 2), the student's *Current Study* is the effect of group size on

[1] In addition to the ontology changes for the second iteration, basic artificial intelligence features were added into the LASAD environment. These provided instant feedback during the diagram construction process using logical rules to analyze students' developing argument diagrams and to provide suggestions to make their diagrams more complete.

responses to sneezing (e.g. "Bless you"), and they *Hypothesize* that with a larger group less people will respond. This *Hypothesis* is *Supported* by the *Claim* that larger group size inhibits prosocial behavior through reduced personal connection; which is *Supported* by a *Citation* of Levine & Crowther (2008) who found that larger group size inhibited helping behavior when bystanders were strangers to a victim.

In the psychology-specific ontology this would look slightly different. The *Claim* node of larger group size inhibiting prosocial behavior would be labeled a *Finding* node instead, since students are citing empirical psychological studies. The *Supporting* node connecting the *Citation* of Levine & Crowther (2008) to the *Finding* would have a rating of relevance (close) and a subsequent justification (helping behavior and responses to sneezing are both forms of prosocial behavior), as well as a rating of validity (strong) and a subsequent justification (controlled experiment). Figure 3 presents a different student's argument in the psychology-specific ontology.

Figure 3: Example diagram from the Psychology-specific ontology



2.2.4 Procedures

For the latter two course iterations, as part of the argument diagramming intervention, we made minor changes to existing assignments in the research methods course and added two new assignments. These modifications were the same for the two diagramming cohorts. These changes to the class assignments are summarized in Table 1. The modifications to existing assignments included (1) adding an in-class lecture and activity to assignment 1 that introduced the components of the LASAD argument diagramming ontology, and (2) adding to assignment 4 the task of creating an argument diagram that justified their hypotheses using the sources collected for assignments 3 and 4. Additional assignments included conducting blind peer reviews of three other student’s argument diagrams using the SWoRD online peer-review system (Cho & Schunn, 2007) and a revision of their initial argument diagram based on peer feedback. The revised argument diagram was submitted to their TF for grading. Students then used feedback from their TFs on the argument diagram to generate a rough draft of their introductions for their lab reports.

For training, students first made an argument diagram in pairs based on a short text describing a hypothetical student's study, hypotheses and supporting and opposing studies (see Appendix). When most pairs had completed at least half of the diagram, the teacher handed out a completed diagram to serve as a model for their own study diagram, and the class discussed whether each hypothesis shown in the diagram was appropriately risky. The students then separated from their partner and began diagramming their own study.

Table 1: Laboratory Section Homework Assignments for baseline and diagramming cohorts

Assignment	Baseline	Changes to Baseline for Diagramming Conditions
1	Worksheet on Hypotheses, independent & dependent variables, & operational definitions	**Added**: Argument diagram practice activity
2	Reading research articles and APA style	*No change*
3	Statistics Exercise	*No change*
4	Reading and understanding research articles	**Added**: Create an argument map for hypotheses using sources
5	No assignment	**New**: Peer review of argument map
6	No assignment	**New**: Revision of argument map
7	Paper draft	*No change*

Peer reviews**.** For both iterations, to further deepen their understanding of the argument diagrams and repair the diagrams before use in writing, students submitted their completed argument diagrams to an online peer review system called SWoRD (Cho & Schunn, 2007). The system assigned four student reviewers to each diagram; the reviewers provided written comments and ratings for six dimensions of writing quality. Reviews were completed out of class. Each student received both a diagram grade and a reviewing grade. The diagram grade was based on the ratings of the four reviewers (proportionally weighting ratings by how generally consistent each reviewer's ratings were with the mean ratings of the other reviewers of the same diagrams). The reviewing grade was based on how similar a reviewer's ratings were to the other three reviewers, along with how helpful the diagram author found their written comments. Both the reviewers and authors remained anonymous.

Student survey**.** Near the end of the semester, in return for participation points, students completed an online survey about their experiences creating the diagram, using the peer review system, and writing their paper.

2.2.5 Measures.

Coding Scheme*.* To assess the quality of students' writing, we developed a set of coding schemes for the variables of interest. Relevance was coded on a per-citation basis, where each citation in a student's paper was rated on a 1-5 scale. A rating of one was defined as "not at all relevant", a rating of three as "somewhat relevant", and a rating of 5 as "very relevant", and coders could use ratings of two and four to denote intermediate degrees of relevance. If a student did not include enough information to determine the relevance of a citation, it was not included in analyses. For each citation, the two coders' ratings were averaged (α=.62), and these values were then averaged across all citations in a student's paper to produce three values of mean, minimum, and maximum citation relevance per student paper. Thirty essays each were coded for relevance from the control and domain-general cohorts, and 40 from the psychology-specific cohort.

Thirty essays each from the control and domain-general cohorts and 60 essays from the domain-specific cohort were coded for a second set of dimensions. This set of dimensions included: Clear hypotheses (k=.80), supporting citations (k=.68), opposing citations (k=.70), and writing about validity (k=.52), coded as present (1) or absent (0) for each dimension by two coders. For instance, if a student had at least one opposing citation, that dimension would be marked as present (1); if they had at least one instance of writing about citation validity that dimension would be marked as present.

Students in the domain-specific cohort were more likely to have two hypotheses than those in the domain-general or control cohorts. Despite this, there was no significant difference between the average number of study citations between the control and domain-specific cohorts (t=2.67, p=.08) or the domain-general and domain-specific groups (see Table 2; t = 1.39, p = .14). Number of hypotheses was not related to minimum (F = 0.76, p = .39), maximum (F = 1.14, p = .29), or average relevance of citations (F = 2.09, p = .16). Table 2 shows descriptive statistics for these dimensions. Number of hypotheses was also not related to inclusion of support, χ^2(1, n=60)= 0.48, p = .49, writing about support validity, χ^2(1, n = 60)= 0.09, p = .77, inclusion of opposition, χ^2(1, n = 60)= 1.270, p = .260, or writing about opposition validity, χ^2(1, n = 60)= 0.58, p = .44.

Student Survey. At the conclusion of the study activities for the psychology-specific cohort, a survey was administered to students asking them about their experience with diagramming and writing their research report. This survey was primarily intended to be a diagnostic tool in interpreting our findings and a resource for future iterations of the diagram ontologies.

2.3 RESULTS

2.3.1 Relevance of Citations

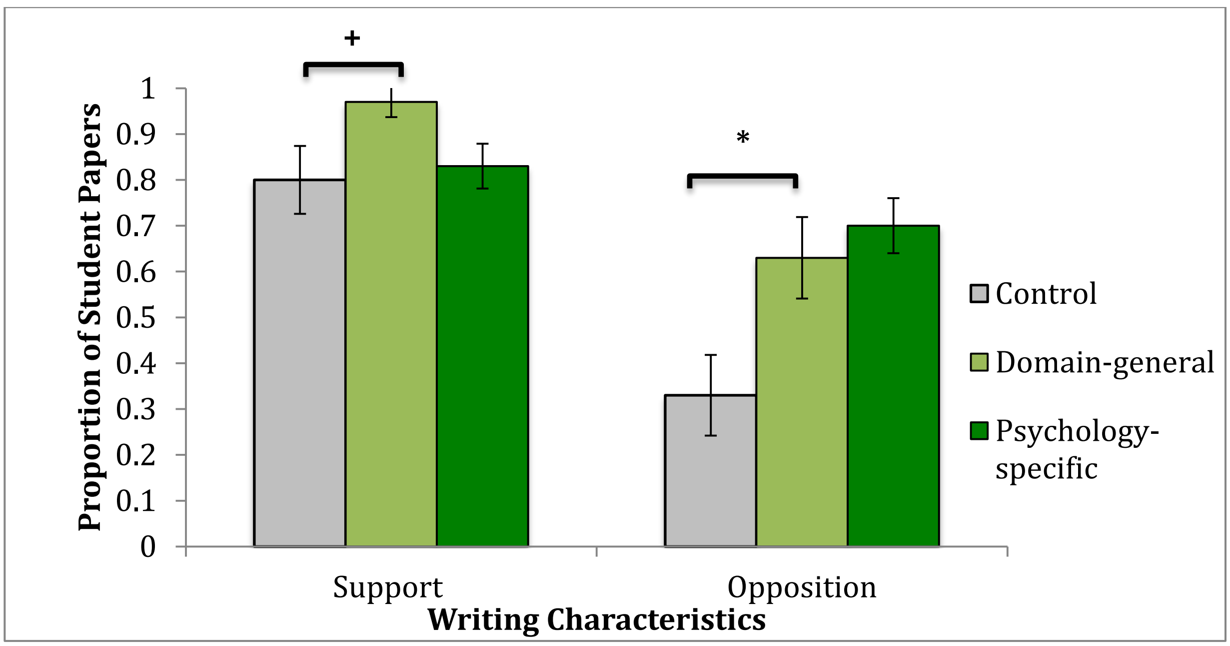
The relevance of study citations generally increased over the three intervention iterations (see Table 2). Relevance was examined in three different ways to ascertain a clearer idea of how any changes manifested in students' writing. The average minimum relevance of citations in a given essay was significantly higher in the domain-general group than in the control group, $t(57) = 4.10$, $p < .001$, $d = 1.09$, and higher in the psychology-specific group than in the control group, $t(65) = 3.21$, $p = .002$, $d = .80$, but not different between the domain-general and domain-specific groups, $t(66) = .82$, $p = .41$, $d = .20$. The average maximum relevance of citations in a given essay was significantly higher for the domain-general group than the control group $t(57) = .27$, $p < .01$, $d = .73$, and higher in the domain-specific group than the control group, $t(65) = 4.65$, $p < .001$, $d = 1.15$, but not different between the domain-general and domain-specific groups. The average relevance of study citations was higher in the domain-general, $t(57) = 3.86$, $p < .001$, $d = 1.24$, and domain-specific groups, $t(65) = 5.04$, $p < .001$, $d = 1.25$, than the control group, but was not different between the two diagramming ontologies, $t(66) = .37$, $p = .70$, $d = .09$. In sum, both types of diagrams improved citation relevance and they did so to an equivalent extent.

2.3.2 Inclusion of Supporting and Opposing Evidence

The inclusion of supporting evidence was not significantly different across iterations, although there was a trend-level difference between the domain-specific and domain-general groups, $\chi^2(1, n = 90) = 3.31$, $p = .092$, $d = .39$. In general, most students included evidence in support of their hypotheses.

Turning to opposing evidence, the rates were much lower across the board. Students using either diagramming ontology $\chi^2(1, n = 60) = 5.41$, $p = .02$, $d = .63$, $\chi^2(1, n = 90) = 11.02$, $p = .001$, $d = .74$, were significantly more likely to include opposing evidence in their essays than those in the control group, although there was no difference in the inclusion of opposing evidence between the two diagramming ontologies $\chi^2(1, n = 90) < 1$, $p = .52$, $d = 0.13$. See Figure 4 for a visual comparison of these results.[2].

Figure 4: Proportion of student papers including supporting and opposing evidence with SE bars



2.3.3 Validity of Provided Evidence

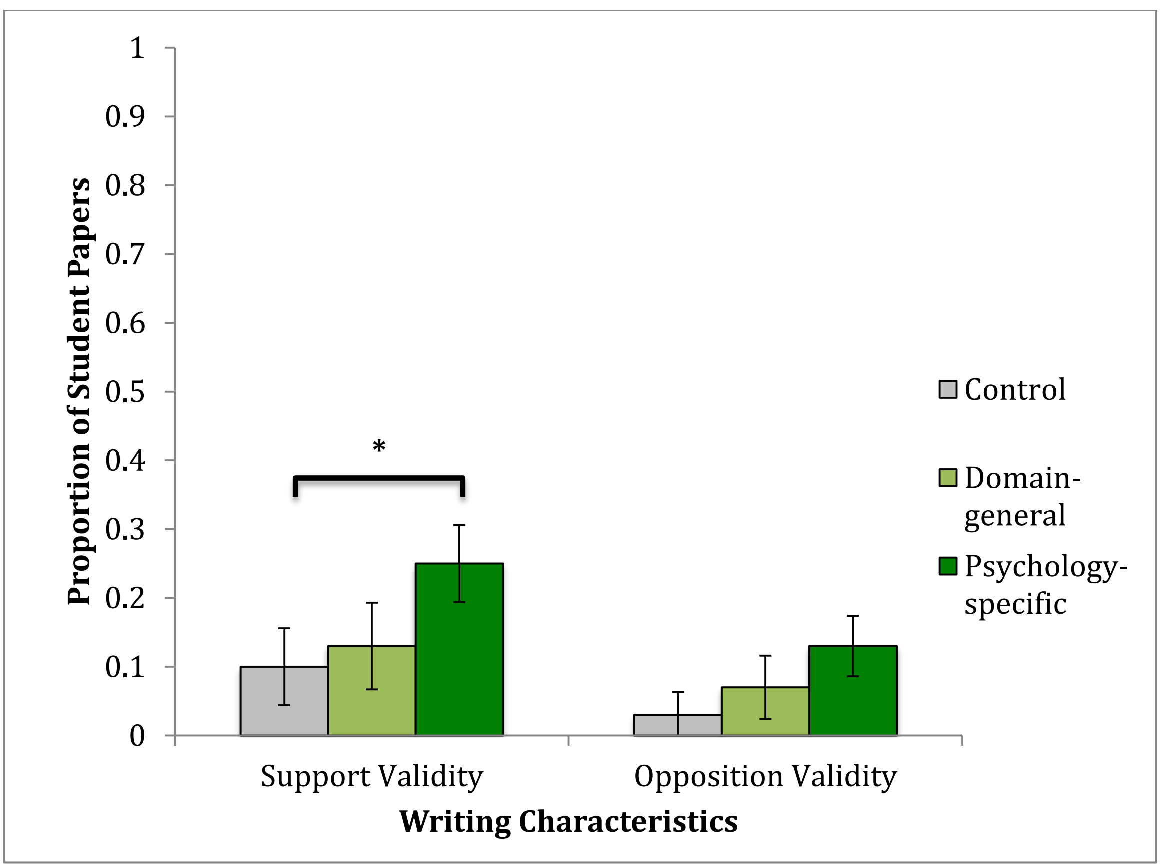
There were no differences between groups in writing about the validity of supporting citations except a trend-level difference between the control and domain-specific groups, $\chi^2(1, n = 90) = 2.81$, $p = .094$, $d = .36$. There were no differences in writing about the validity of opposing citations. See Figure 5 for a visual comparison of these results.

[2] Significance at the $p < .05$ level is denoted by *, $p < .01$ by **, $p < .001$ by ***, and trend-level effects ($p < .10$) are denoted with + in all figures.

Table 2: Number of hypotheses, citations, and relevance of citations in student papers

Year	2 Hyp.	# Cites	Min. Relevance	Max Relevance	Avg. Relevance
Control	100%	4.87	2.38	4.29	3.35
Domain-General	100%	4.53	3.4	4.77	4.08
Psychology-Specific	51%	4.18	3.18	4.91	4.14

Figure 5: Proportion of valid support and opposition in student papers with SE bars



2.4 DISCUSSION

The results indicate that some benefits of argument diagramming are robust to changes in the underlying argument diagram ontology. For example, doing either form of tested argument diagramming helped students to use more relevant citations in their papers. These effects were seen in terms of reducing the frequency of low relevance citations (i.e., changes in minimum relevance), increasing the frequency of high relevance citations (i.e., changes in maximum relevance), and general increases in citation relevance (i.e., changes in average relevance). Additionally, doing either form of argument diagramming appeared to also help students include opposing evidence for their hypotheses.

There results also indicate some, albeit very weak, evidence that there may be differences between the benefits afforded by these ontologies. However, given that these effects only approached statistical significance further testing would be needed to support any distinctions.

2.4.1 Theoretical Implications

Generally, these data point to the two-part value contained in well-designed argument diagramming activities: 1) the spatial structure of argument diagrams makes some kinds of argument aspects particularly salient, and 2) the detailed textual structure of argument diagram components make other aspects of an argument salient. Thus, argument diagrams are importantly a hybrid spatial-symbolic tool for supporting thinking and reasoning. Previous studies on argument diagramming lack theoretical explanations for its effects and are generally focused on classroom applications and implications rather than theoretical understanding. Our explanations may form a starting point for future research to build a deeper theoretical understanding of these representations, which should include investigation of the cognitive mechanisms involved in creating and using argument diagrams.

2.4.2 Practical Implications

At a more practical level, the results of this study indicate that diagramming is a useful practice to employ in college-level psychology courses to improve students' writing, and should be integrated into curricula. Our findings support previous research in this area showing that diagramming can be beneficial for students in many educational domains (Griffin et al., 1995; Harrell, 2013; Nussbaum et al., 2007), including science writing. Previous work has looked at the inclusion of supporting and opposing evidence in argumentation, but we are the first to show that argument diagramming benefits citation relevance and writing about citation validity, important components of scientific argumentation. Our research also explores ontological variations, which suggests that the benefits of diagramming may be relatively robust to ontological variation and ontology specialization may not be necessary or even beneficial.

Regarding whether domain-general or domain-specific ontologies should be used depends upon the relative importance of various learning objectives. Given that the unique effects of either diagramming ontology were few and non-significant, a good argument could be made for using more domain-general ontologies. Such an emphasis would allow for students to use similar diagramming techniques across courses in various disciplines (Philosophy, Psychology, Physics, etc.). This would facilitate corroboration of scientific evidence concerning diagramming and narrow the diversity of diagramming ontologies for comparison. Validity, however, is a central, deep structural concept in research, and perhaps the most important aspect of the research activity. Thus, from the perspective of writing-to-learn about science, potential improvements in treatment of the validity concept could be deemed sufficiently important to warrant use of the domain-specific ontology, if further research indicates the trend-level effects are representative of a stronger relationship. More research is needed before this position can be advocated for with confidence.

2.4.3 Caveats

This study did not utilize a strict experimental design (three iterations with three different cohorts), meaning that cohort effects are possible alternative explanations for the condition differences. However, we attempted to control for this by ensuring that all three cohorts were similar in GPA and other academic characteristics, and we used a variety of teaching fellows, making it unlikely that differences stemmed from a particularly effective teaching fellow. Further, the use of multiple teaching fellows shows some robustness of the effects across a range of qualities/styles of instructional support that are commonly found in these contexts.

Another important consideration is the intervention's combination of techniques. In particular, since the effects of peer review of diagrams or the specific implementation details of LASAD were inseparable from the effects of a pure diagramming task in this study, we do not know how much these elements of the intervention are responsible for the overall effects. Based on their survey responses, however, students did not believe the peer review process of diagrams to be very helpful to their writing. Only 50% of students in the domain-specific group found peer feedback comments helpful to the task, and only 20% of those students found peer feedback ratings helpful. Further, LASAD is similar to many other tools for diagramming at a basic structural level. Thus it is unlikely that other elements or factors played a large role in the writing gains seen here beyond the core diagram structures themselves.

2.5 CONCLUSIONS

The use of argument diagramming in education has been supported by previous research in this area (Griffin et al., 1995; Harrell, 2013; Nussbaum et al., 2007; Lynch et al., 2014; Lynch, 2014), but this study presents the first attempt to rigorously study differences in diagramming ontology, in this case, the difference between a domain-general versus a domain (psychology)-specific ontology. Our results support prior findings (Griffin et al., 1995; Harrell, 2013; Nussbaum et al., 2007) that any kind of diagramming activity can be helpful for writing – science writing in particular. Both of the studied ontologies helped students to include more relevant citations in their papers and include evidence opposing their hypotheses. The data also indicate that these effects are relatively robust across ontology changes, but that some benefits (writing about validity, inclusion of supporting citations) may be sensitive to ontology.

Potential differences in effects between the two ontology types may be explained by the level of writing issues, where high-level issues (relevance, support, opposition) can be identified using any spatial representation, but that lower-level issues (e.g. writing about the validity of citations) may be more easily identified with a domain-specific diagramming ontology. Alternatively or additionally, the difference may be explained by the relative difficulty of writing issues. Citation relevance and the inclusion of support and opposition may be easier for students to grasp, so any ontology facilitates their improvement; while writing about citation validity is harder to deal with so students may benefit from the extra scaffolding of a domain-specific ontology in order to improve them. Additional research in this area will help determine which explanation is stronger, and what other benefits argument diagramming may elicit for students.

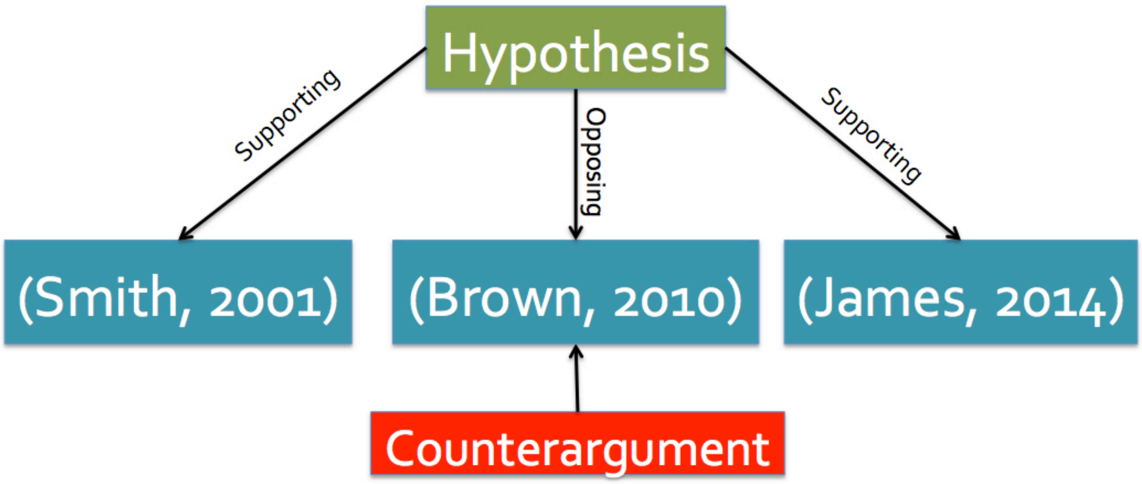
3.0 EXPERIMENTAL EVIDENCE FOR DIAGRAMMING BENEFITS IN SCIENCE WRITING

3.1 INTRODUCTION

Researchers have been studying the affordances of different representation formats for problem solving and learning for nearly half a century (Mandler & Ritchey, 1977; Paivio 1986; Shepard, 1967; Standing, 1973). Spatial representations have specifically been studied as important external tools that afford benefits to reasoning and problem-solving (Cheng, 1992; Cheng & Simon, 1992; Larkin & Simon, 1987; Novick, 2000; Trafton, Trickett, & Farilee, 2005). Argument Diagrams, as one form of spatial representation, have been employed as instructional tools in education with substantial empirical support from a growing body of research.

Argument diagrams visually represent arguments by breaking them down into component parts and their relationships, based on an 'ontology', or system of organization. In the case of science writing, for example, these might be a hypothesis, various study findings, and counterarguments (see Figure 6). Argument diagrams have been shown to facilitate student learning and retention across a variety of domains. In social studies, students who diagrammed novel learning material retained the information better than their classmates who did not diagram (Griffin, Malone, & Kameenui, 1995). Multiple studies have indicated the robustness of argument diagramming for improving students' ability to critically analyze arguments (Dwyer, Hogan, & Stewart, 2012; Harrell, 2008; 2011; 2012; 2013) and to generate them (Harrell, 2013). Such diagramming also shows potential for helping students write argumentative essays across various disciplines (Chryssafidou, 2002; 2014, an important task that is a source of struggle for many students (Andrews, 1995; Andrews & Mitchell, 2001; Hahn & Oaksford, 2012; Kuhn, 2013).

Figure 6: Simplified representation of argument diagram ontology



Learning to argue requires not only the acquisition of cognitive skills (e.g., inference rules in a domain), but also the internalization of social, epistemological, and metacognitive dimensions necessary for effective evaluation and production of argument (Kuhn, Zillmer, Crowell, & Zavala, 2013). Traditional instruction for argumentative writing tends to focus students on how to explain their argument persuasively to others, but does not necessarily teach them how to create an initial hypothesis or thesis (Andrews, 1995; Andrews & Mitchell, 2001; Oostdam, de Glopper, & Eiting, 1994; Oostdam & Emelot, 1991).

Argumentative writing in science takes a number of form. One form relates to organizing the results of a study towards a conclusion (i.e., writing the results and discussion sections of a paper). While complex, this aspect of writing tends to be more manageable in that the discovery of a thesis is not usually problematic, and in student projects the range of evidence to be integrated is relatively small.

Another aspect of argumentative writing that is less well-studied is arguing for the need for a study (e.g. writing the introduction to a research report). This form of argumentative writing poses unique challenges for students (Osborne, Simon, Christodoulou, Howell-Richardson, & Richardson, 2013). They need to grasp a large body of conceptual, procedural, and epistemic knowledge and be able to integrate complex scientific evidence into a coherent argument. For example, an individual research paper can present a range of findings—some may support a theory while others contradict it, and others may be irrelevant to the student's argument. This is made more difficult by science instruction that obscures the argumentative, ambiguous nature of interpreting scientific evidence (Gray & Kang, 2012). Further, different papers may use different methodologies that each have varying kinds of threats to validity. Recent efforts to integrate argumentation into science instruction may make the task easier, but this integration does not happen naturally (Osborne et al., 2013).

Interestingly, introductions to research papers have a central feature that makes their argument structure unique. In contrast to typical dialogic argumentation, where multiple competing perspectives may be explored but the end goal is resolution in favor of one perspective, research paper introductions seek to clarify an open question for which there is supporting evidence but the prior evidence is insufficient. That is, the writer must present a convincing argument in favor of their hypothesis, but also leave enough ambiguity that the issue still appears worthwhile to test. A prior investigation in the case of research psychology found that explicit writing about hypothesis 'risk' is a very common feature of published articles. This risk can be established by noting a gap in situations that had been studied previously, flaws in the evidence that had been previously collected, or contradictions in prior findings (Barstow et al., 2015).

Importantly, explicit writing about hypothesis risk in research paper introductions is also commonly missing in student work (Barstow et al., 2015). In general, this way of conceptualizing and structuring introductions is not typically taught in research methods classes. Further, given its inherent complexities (discussed below), students will likely need support to address it properly. Argument diagramming is one way of providing that support and explicitly structuring science writing as argument. At the basic level, students may fail to include strong support for their hypothesis (Schwarz, Neuman, Gil, & Ilya, 2003), while at more intermediate levels, students may fail to include any reason to doubt their tested hypothesis (i.e., fail to note possible counter-evidence) (Nussbaum & Schraw, 2007). Failure to consider alternatives has sometimes been considered a skill deficit (Crowell & Kuhn, 2014; Kuhn, Hemberger, & Khait, 2015), but failures to write about alternatives might also stem from being overwhelmed by the tasks of managing all the arguments that are for and against (Sweller, 1994). Such an overload seems likely when each piece of evidence is itself complex, as is typically the case in science. Finally, students may include evidence for and against, but fail to find a resolution to these conflicts to provide a satisfactory proposed hypothesis.

In other cases where researchers are interested in writing in science, they have developed and applied various frameworks to define and structure the nature of science, as well as the process of writing it. One example, the Science Writing Heuristic (Keys, Hand, Prain, & Collins, 1999; Hand, Prain, & Wallace, 2002; Hand, Wallace, & Yang, 2004; van Amelsvoort, Andriessen, & Kanselaar, 2007), provides scaffolds (templates) for both students and instructors that structure students' scientific thinking as a series of questions (e.g. "How do I know? Why am I making these claims? How do my ideas compare with other ideas?"). Although this and other conceptualizations may address hypothesis risk and related concepts indirectly, we have chosen to frame the process of science in this study to directly address these important issues.

We have conceptualized the primary challenges specific to writing strong introductions in psychology research (and likely behavioral science more generally) as framed around three components: Hypothesis risk, or the demonstration of uncertainty; citation relevance, the applicability of cited studies to an author's hypothesis; and citation validity, the strength of evidence presented in cited studies. Achieving appropriate risk in student introductions is our goal, and relevance and validity are two powerful components for managing and addressing hypothesis risk. One way that authors in published psychology research tend to argue in support of assertions of risk is through uncertainty—when limited prior research leaves open questions in the field (Barstow et al., 2015). Uncertainty can be understood as a problem of relevance, where no studies are relevant *enough* to the given hypothesis, or as a problem of validity, where there are relevant studies but none executed rigorously enough to be definitive. The other common way that authors in psychology address hypothesis risk is through opposition, where multiple studies in a given area have produced conflicting evidence (Barstow et al., 2015). This is a validity problem, where no one study has been performed carefully enough to promote consensus among researchers. Although there is evidence for student difficulties in successfully navigating these facets of argumentation in research writing, no research has yet directly studied how we may help novice and intermediate writers skillfully address these components in their arguments. The current study tests the extent to which argument diagramming can help students manage these difficult concepts.

What are the mechanisms by which diagramming could support these skills? There are two core elements in argument diagrams: spatial, in which information is embedded in the structure of the diagram, and textual, in which information is presented in the content of the diagram nodes or links. Diagrams are thus a hybrid representation in which each aspect may be involved in improving argumentative writing.

The spatial layout will likely enable students to gain a better understanding of hypothesis risk by indicating the existence of evidentiary relationships between studies and hypotheses, and whether the links are supporting or opposing. It will also help them balance evidence (in the case of mixed evidence); balance, as an inherently a spatial metaphor, could be well supported by diagrams.

The textual element of diagrams allows students to 'zoom in' to their argument and access critical summary information to judge the strengths and weaknesses of each piece of evidence. We expect this aspect to be particularly helpful for understanding the relevance and validity of their cited studies. Relevance involves thinking about the semantic overlap between the hypothesis and the studies being cited. Validity involves thinking about the semantic content of the studies being cited with respect to the claims being hypothesized. Such semantic judgments inherently involve use of text; it is not clear how such content could be represented spatially.

It is possible to support students with separate spatial and text representations (e.g. an outline and a diagram) for these two different types of content, but having them combined in a hybrid tool allows students to easily integrate information from both representations into a cohesive understanding and argument. Indeed, the relative strengths of support and critique being judged in spatial structure depend upon being able to first judge the relevance and validity of each component.

Although conceptually there is a good match between the needs of students in writing research introductions and the affordances of argument diagrams, the research support for such benefits is still preliminary. One study found the quality of college students' argument diagrams was correlated with the quality of the research paper introductions that students later produced (Lynch, Ashley, & Chi, 2014; Lynch, 2014). But it is unclear from this study whether the diagrams improved writing, or whether misconceptions revealed by students' diagrams were also manifest in the students' writing.

Another complication is that all argument diagramming tools and frameworks are not likely equal in their effectiveness for supporting students. Prior studies in this area have typically employed domain-general Toulmin (1958)-style models (e.g. Stegmann, Wecker, Weinberger, & Fischer, 2012; Stegmann, Weinberger, & Fischer, 2007; Harrell, 2013) that likely lend themselves to cross-domain transfer. For this study we will employ a psychology-specific ontology to target and support more nuanced concepts in the domain.

The similar prior work on argument diagramming for psychology introductions by Barstow et al. (2015) had other limitations. First, students in the diagramming condition also did peer reviews of the diagrams, and it is possible it was the peer review rather than the diagrams that was important. Peer review allows students to consider alternative perspectives in evaluating the relative strength of arguments presented by others. The current study removes peer review of diagrams from the intervention. Second, the current study improves on experimental design from the aforementioned study, with a single cohort population randomized to condition by classroom.

In sum, for this study, we examine introductions to APA-style research papers in psychology created by students in a research methods course. Students were randomly assigned to have either diagramming support or no support (beyond what the course offers to all students). Our hypotheses are as follows:

1) Students given diagramming support will be more likely to explicitly address hypothesis risk in their introductions than students given no support
2) Students given diagramming support will write more about the relevance of cited studies than those given no support
3) Students given diagramming support will write more about the validity of cited studies than those given no support

3.2 METHODS

3.2.1 Participants

Participants consisted of 182 students enrolled in a multiple sections of a Research Methods in Psychology course at a large public university. Seventeen students did not complete the paper assignment, which reduced the final sample to 165 students. The formal course components consisted of 2.5 lecture hours (focused on theoretical issues in psychology research) and 3 lab hours (focused on practicing basic skills related to observational and experimental research) per week. The intervention was implemented in the lab. Participants were recruited into the experiment by their teaching fellow (TF). Seven TFs signed up for the experiment and were then matched into pairs based on teaching experience and class characteristics (time of day, day of week). TFs within each pair were then randomized into either experimental (diagramming) or control conditions. One TF taught two lab sections and was treated as a within-instructor pair of experimental and control sections.

3.2.2 Materials

Diagramming is an activity embedded in tools, and there is a reasonable concern that if the tools we construct are too optimized for one task then they are inaccessible or not useful for other tasks. For this study, we utilized a readily accessible (i.e. free, easy to learn) tool so that students could choose to use it again, although this was not monitored. The accessibility of the tool also enables simpler research on scaling and easier application in classrooms.

We constructed the diagramming ontology used in this study through extensive pilot testing and iterative development, beginning with a generic, technologically complex ontology that evolved into a simpler, psychology-specific ontology. This ontology has been refined to draw particular attention to issues of relevance, validity, and thus risk in psychology research.

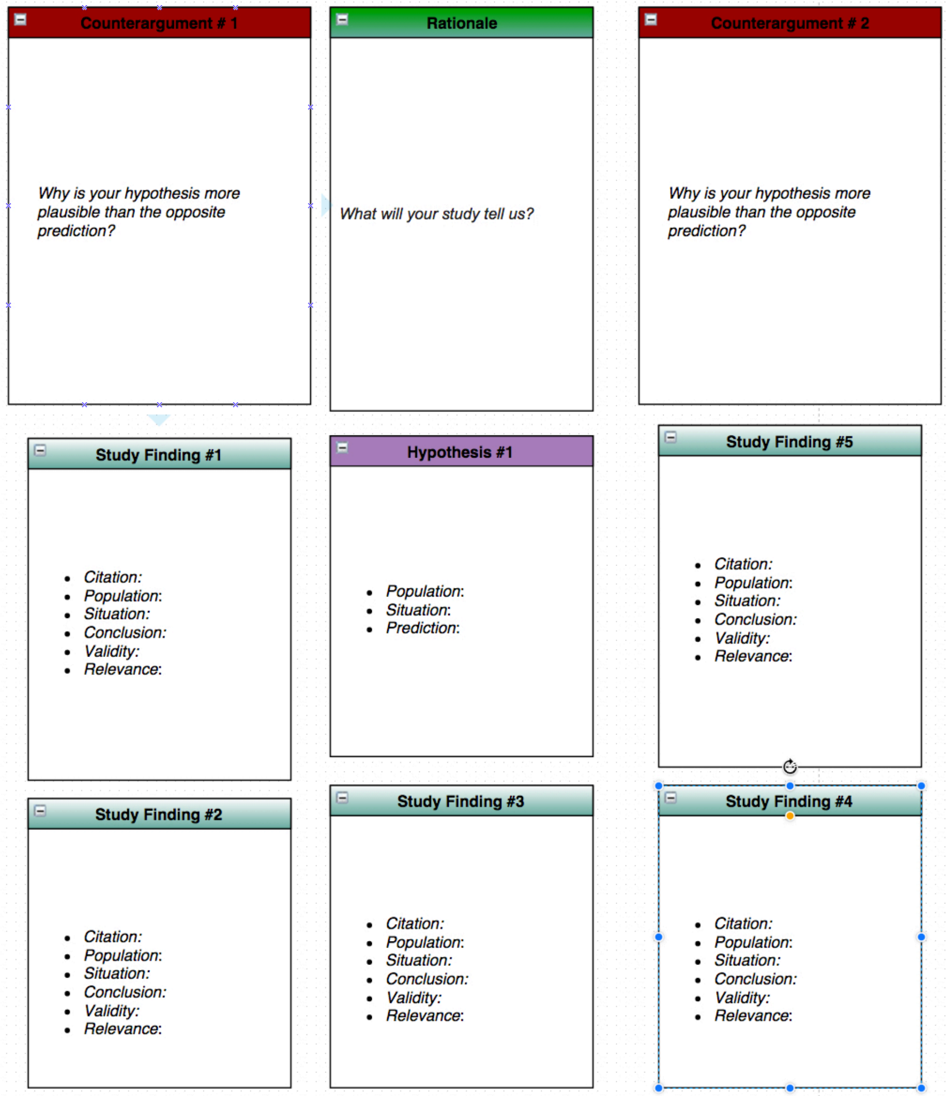
Students constructed argument diagrams in a free, web-based, open-ended application called Draw.IO[3]. This diagramming tool was chosen for its relative simplicity and accessibility, which make it an ideal choice for classroom application, and possible transfer to use in later courses.

Since students were encouraged to have two hypotheses that are being tested in their experiment, students were instructed to include the following elements in their

[3] https://www.draw.io/

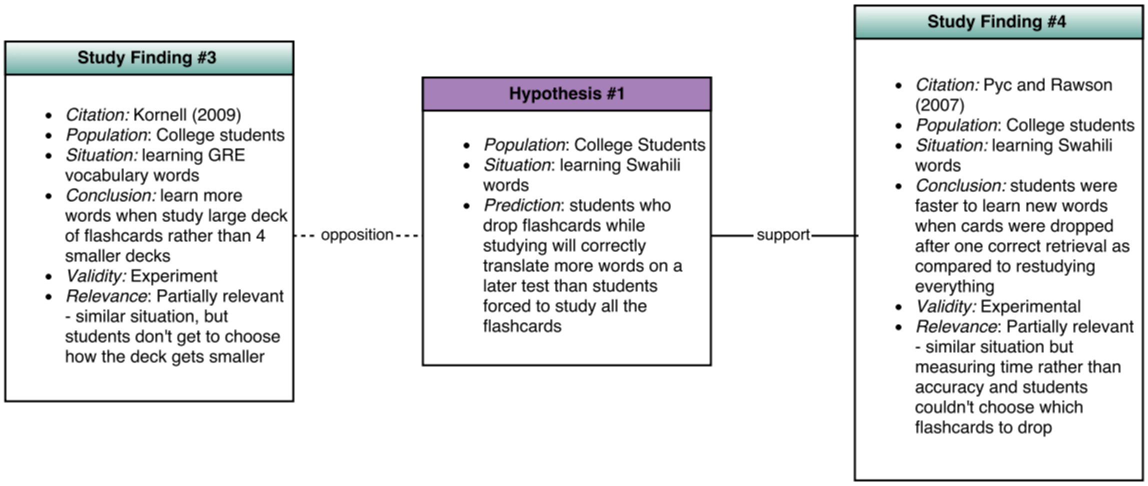
diagrams: two hypotheses, citations of relevance to each hypothesis organized as either supporting and opposing study citations, and counterarguments for any opposing evidence to a proposed hypothesis. These guidelines were presented to students along with a diagram template (see Figure 7), which contained nodes of each type and basic descriptions of the core information to include in each node type. Students were instructed to duplicate these template node types as often as needed, fill out the contents, and connect the completed nodes to one another. Multiple finding nodes were included in the template to emphasize that each hypothesis should be connected to multiple findings in the literature.

Figure 7: Argument diagram template



In the example diagram shown in Figure 8, the author proposes a study on college students learning Swahili words. They hypothesize that students who 'drop' individual flashcards once well-learned will correctly translate more words on a later test than students forced to always study all of the flashcards. This hypothesis is supported by Study Finding #4, where dropping flashcards resulted in improved speed rather than accuracy as a dependent variable. However, it is opposed by Study Finding #3 where separating a larger deck of flashcards into four smaller decks resulted in poorer memory for word definitions.

Figure 8: Filled diagram from introductory demo activity

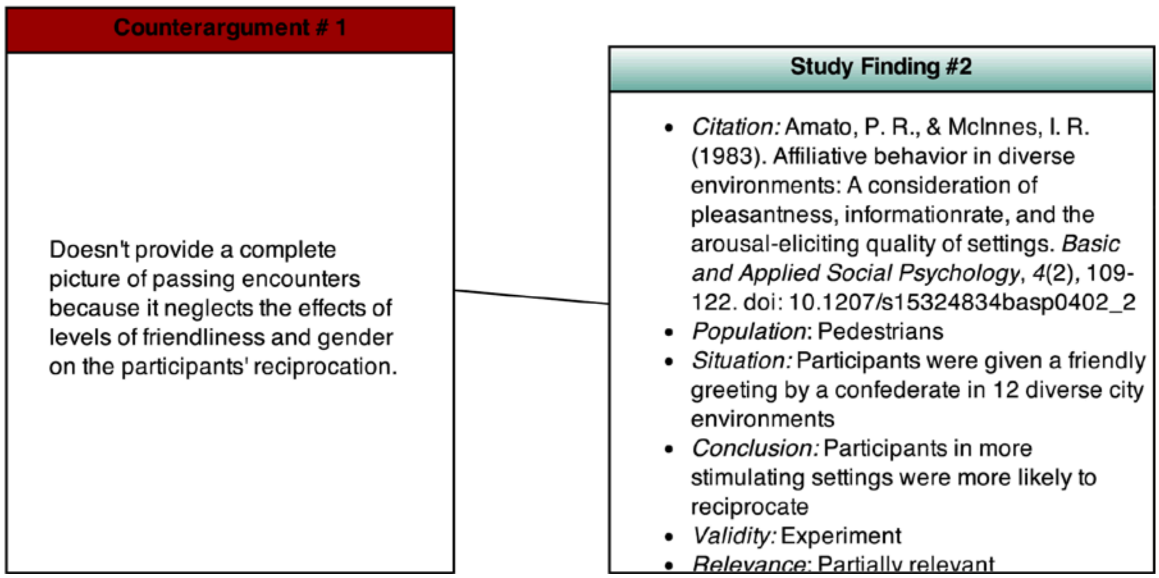


If students could not locate opposing evidence to their hypothesis, they were instructed to demonstrate hypothesis risk in other ways (i.e. through insufficient data, validity issues with prior studies). For example in Figure 9, the author demonstrates risk by noting a gap in existing knowledge regarding the bystander effect in low-risk situations.

For each study cited in their diagram, students were instructed to record the APA-style citation, population tested, situation (tested variables and context), conclusion (findings), validity (e.g. experimental), and the relevance of the evidence to a student's hypothesis(es) (See Figure 9).

On the basis of the information included in the hypothesis and study nodes, students were asked to categorize each study as slightly, partially, or highly relevant to the linked hypothesis, and were encouraged to justify their choice. For example, the author of the diagram in Figure 8 rated Study Finding #4 as partially relevant because the cited study dealt with a high-risk situation versus the author's proposed low-risk context. Students also labeled the connection between a study and a hypothesis as either opposing or supporting evidence.

Figure 9: Actual subset of student diagram created in Draw.IO



3.2.3 Measures

Risk coding scheme. All research paper introductions were coded for risk using a coding scheme validated through coding the treatment of risk in the introduction sections of journal articles in psychology (Barstow et al., 2015). We focused on the two categories that commonly occur in psychology: risk through uncertainty and risk through opposition:

Risk through uncertainty (RU) occurs when the author claims that there is only insufficient or problematic evidence for his/her hypothesis(es) in the literature. E.g., "First, although the effect of fluency on a variety of judgments has been well documented, it is unknown whether fluency can influence two different attributes at once" (Westerman, Lanska, & Olds, 2014).

Risk through opposition (RO) occurs when the author claims that there is both supporting and opposing evidence for his/her hypothesis(es) in the literature. E.g., "While there is strong evidence for such a process of combination, there has been some debate as to when metric and categorical cues are combined…" (Holden, Newcombe, & Shipley, 2014).

Using these definitions, we coded the introductions of each student paper by annotating individual sentences that addressed risk, and coded each article based on the types of risk addressed, regardless of the number of instances. For example, if an article had four sentences tagged as RU and one as RO, that article would be tagged as [RU, RO]. In other words, one instance of a risk type was sufficient to be considered. The introductions were first coded by an expert and a subset of these (n=25) were then double-coded by a second coder (κ=.91).

Relevance & validity coding scheme*.* A subsample of student papers (102 sampled in a stratified way across sections) were then coded at greater depth for relevance and validity of each citation using an iteratively developed coding scheme.

Relevance coding was separated into *categorical* coding dimensions, in which raw characteristics of a study are discussed, and *comparison* coding dimensions, in which a student directly compared studies on the basis of these characteristics (although explicit reference to them was not necessary). Validity coding also involved *categorical* coding dimensions, but as a second type of code involved *evaluative* coding, in which a student makes a judgment about the validity of a study.

Categorical coding of relevance included population (e.g., "College students") and context (e.g., "a busy street"). Comparison coding of relevance involved noting instances in which the author discusses similarities and/or differences between the cited study and their own proposed study. These comparisons needed to be based on study characteristics rather than study findings.

Categorical coding of validity the common factors influencing validity in psychology research that were also discussed explicitly in the lecture portions and associated textbook: sample size (e.g., "145 participants"), experimental design (e.g., "a meta-analytic review of social psychological literature…"), and confounds (e.g., "The authors looked at sign color and compliance but used two different locations for the signs."). Evaluative coding of validity was defined as an evaluation of the scientific rigor of a study (e.g., "But, this was just a correlational study").

Two coders were trained on 20 student papers to establish agreement before coding the full set. Kappa was calculated based on 8 possible codes for each coder: population, context, comparison, sample size, experimental design, confound, evaluation, and null (no code), k=.40. Disagreements between coders were resolved on a weekly basis by an expert and most commonly took the form of one coder marking text and the other not marking it at all (code/blank). When this type of disagreement is removed from the reliability analysis, the reliability was quite high, k=.79.

3.2.4 Procedures

Students in the experimental lab sections were first given a 5-minute lecture explaining hypothesis risk. This lecture communicated 'appropriate risk' as a balance between insufficient risk, in which case conducting a study would be redundant, and excessive risk, in which case conducting a study would be unlikely to work. The lecture also conveyed the importance of relying on valid and relevant research in locating strong support for one's hypothesis. Afterwards, these students completed a brief training activity in which they were given printouts for three argument diagrams and were tasked to work in pairs to choose which diagram described a hypothesis that was too risky (insufficient supporting evidence), which was too 'safe' (only supporting evidence), and which demonstrated an appropriate level of risk (some supporting and some opposing evidence).

In a later class, students were introduced to the diagramming software through a practice activity in which students worked in pairs to diagram a short scientific paper given to them by their instructor. The paper was selected to be short and involve a mixture of supporting and opposing findings. The practice task was not graded, and the TF showed an accurate diagram at the end of the activity.

Finally, students were instructed to construct an argument diagram for an observational experiment they would later conduct; this diagram was turned into the TF for grading and feedback. Students were told in advance that the diagram would help them write the introduction to the paper.

The paper assignment for both conditions included an abstract, introduction, methods, results, and discussion. Students were asked to include at least five peer-reviewed references in their introduction (two of which were provided by the instructor) and two hypotheses, and to discuss and explain at least one study or theoretical position that conflicted with their hypotheses.

Students also completed a second paper assignment that was nearly identical to the first, except that: 1) the study involved a factorial design experiment, rather than observational study, and 2) students worked on the paper in dyads. These dyads were within-lab section so it is highly unlikely that any cross-contamination occurred between paper 1 and paper 2. Collecting data from this second assignment was not included in the original experimental protocol, but early results from paper 1 spurred our interest in possible temporal transfer effects. After the end of the semester, we were able to collect papers from the two lab sections taught by the same TF (n=25 papers) and these papers were coded for risk using the protocol outlined above.

3.3 RESULTS

For all of the analyses, we used $\alpha=.05$, and Cohen's *d* is used to indicate effect sizes. Given the relatively small number of students per section and the relatively small number of sections, formal nested regressions that directly account for nesting of students within sections would have been underpowered. However, data patterns were examined by matched pairs of sections to insure the same pattern generally held across the data, rather than being driven by just one section.

Hypothesis Risk. In the control sections, 37% of students addressed risk in at least one form compared to 62% of students in the diagramming condition. Figure 10 shows the proportions of papers in each condition including each form of risk, any form of risk, and multiple forms of risk. A χ^2 test of independence revealed that students in the diagramming condition were more likely to write about risk through uncertainty, $\chi^2(1, n=165)=10.2$, $p=.001$ ($d=0.51$), risk through opposition $\chi^2(1, n=165) = 6.2$, $p=.01$ ($d=0.52$), any form of risk, $\chi^2(1, n=165)=14.6$, $p<.001$ ($d=0.62$), and multiple forms of risk $\chi^2(1, n=165)=8.6$, $p=.003$ ($d=0.47$) than students in the control condition. Only some of these results held when examined for only the two within-instructor lab sections, showing more writing about any risk, χ^2 (1, n=42)=8.8, $p<.001$, ($d=1.02$) and more writing about RU, χ^2 (1, n=42)=4.8, $p=.03$, ($d=0.71$) in the diagramming condition, but no significant difference across conditions in writing about RO, $\chi^2(1, n=42)=0.8$, $p=.37$ ($d=0.28$), or combinations of risk types $\chi^2(1, n=42)=0.5$, $p=.49$, ($d=0.21$).

On paper 2, 50% of students in the control sections addressed risk in at least one form compared to 77% of students in the diagramming condition (See Figure 11). A χ^2 test of independence applied to the two lab sections revealed a trend-level difference in which students in the diagramming condition were more likely to write about risk through opposition, $\chi^2(1, n=25)=3.4$, $p=.07$ ($d=0.79$) and multiple forms of risk, $\chi^2(1, n=25)=3.1$, $p=.08$ ($d=0.75$); but not more likely to write about any risk, $\chi^2(1, n=25)=2.0$, $p=.16$ ($d=0.59$), or risk through uncertainty, $\chi^2(1, n=25)=.5$, $p=.47$ ($d=.29$). Given the relatively low power of this analysis, the trend-level effects are encouraging, although not conclusive.

Figure 10: Proportion of student papers addressing any risk (Any), risk through uncertainty, risk through opposition, and a combination of risk types with SE bars

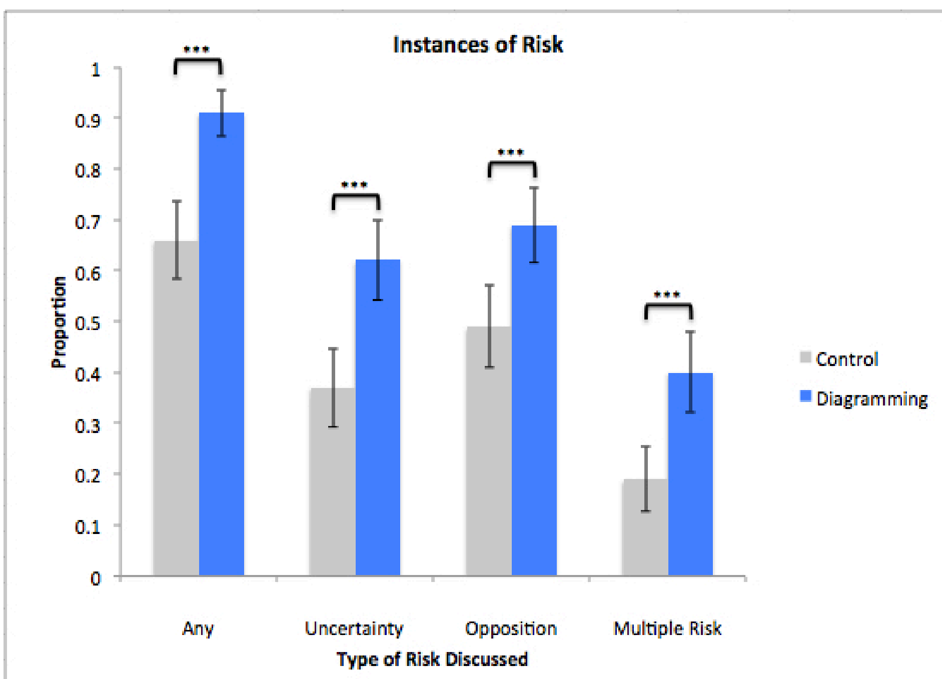
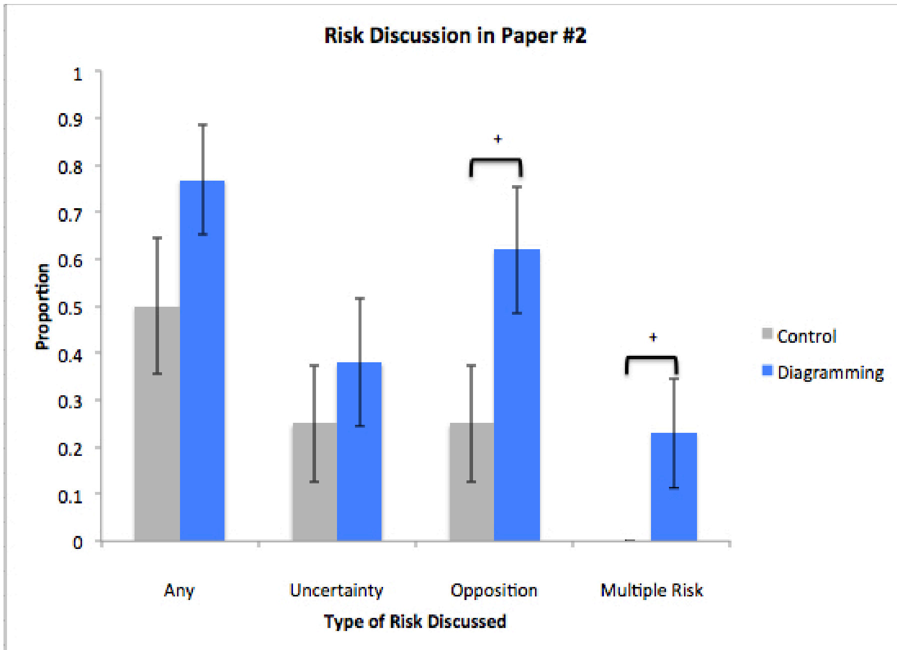


Figure 11: Proportion of student papers addressing any risk, risk through uncertainty, risk through opposition, and a combination of risk types on paper 2 with SE bars



Relevance and Validity. Students in the diagramming condition wrote significantly more about the relevance and validity of citations than those in the control condition on all seven dimensions. T-tests revealed that all of these differences are significant (see Table 4). There was no significant difference in the mean number of citations per paper across the two conditions. The mean difference across conditions is largest for writing about the context of cited studies (1.1 instances) and smallest for writing about evaluations of the validity of cited studies (.3 instances) (See Figure 12).

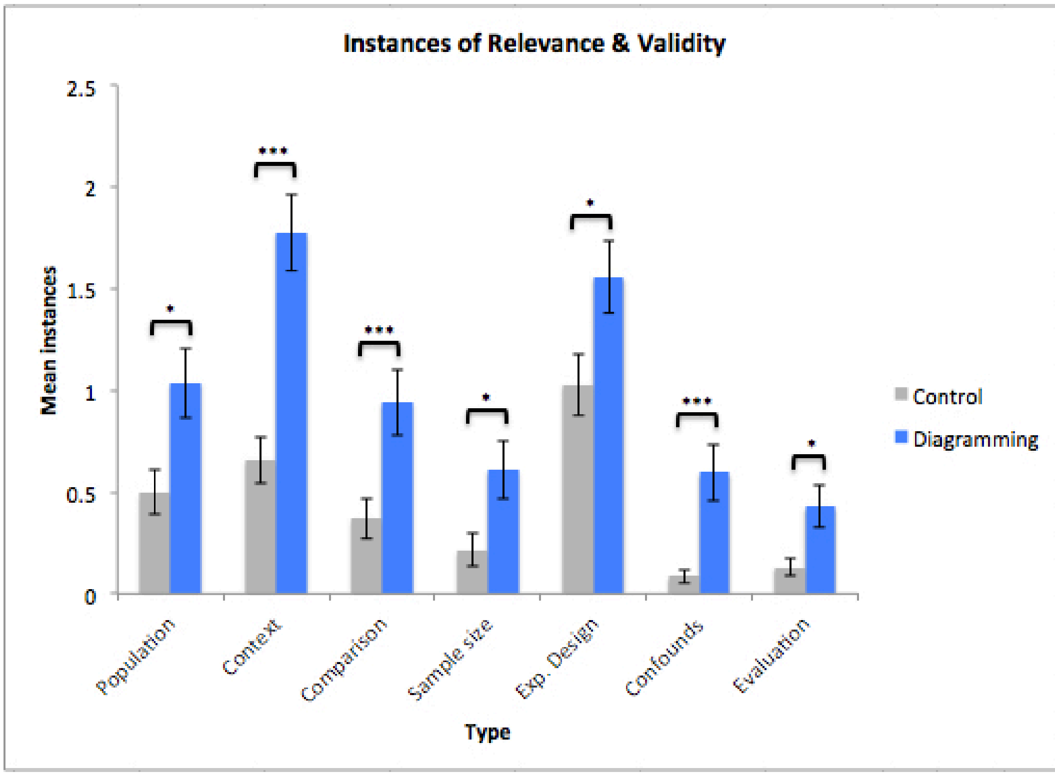
As with hypothesis risk, some but not all of these results held when examined for the within-instructor sample. Students in the diagramming condition wrote more about population, $t(40)=3.1$, $p=.003$, $d=.98$, context, $t(40)=5.6$, $p<.001$, $d=1.77$, comparisons, $t(40)=2.4$, $p=.02$, $d=.75$, and validity evaluations, $t(40)=2.1$, $p=.048$, $d=.66$ than those in the control condition. However, there was no difference in this sample for writing about sample size, experimental design, or confounds.

Table 3: T-tests comparing writing about relevance and validity across conditions

T tests for all by condition

	t	df	sig.	Mean Difference	*d*	Confidence Interval Lower	Confidence Interval Upper
Population	-2.6	136.5	.01	-0.54	0.44	-0.94	-0.13
Context	-5.1	134.9	.00	-1.12	0.88	-1.55	-0.68
Comparison	-3.0	134.5	.00	-0.57	0.52	-0.95	-0.19
Sample size	-2.4	130.1	.02	-0.39	0.42	-0.71	-0.07
Exp. Design	-2.3	151.1	.03	-0.53	0.37	-1.00	-0.06
Confounds	-3.5	92.4	.00	-0.51	0.73	-0.80	-0.22
Evaluation	-2.7	107.8	.01	-0.30	0.52	-0.52	-0.08
Citations	0.7	152.0	.47	0.20	0.11	-0.34	0.74

Figure 12: Writing about relevance and validity separated by subcomponent on paper 1 with SE bars



3.4 DISCUSSION

Introductions to research papers are deceptively difficult to write because of the complexity of the structure that must be created in organizing the prior literature and because of the complexity of embedded research concepts that must be used in creating a sensible organization. One key element that connects the challenges of organization to complexity of the underlying concepts is hypothesis risk. Hypothesis risk is essential to the conduct of research: replications of settled science are not considered useful and yet there should be some prior theoretical or empirical support for a hypothesis as well. In psychological research, authors use it to explicitly demonstrate the need for a study or even a larger project to be conducted in relation to existing research and gaps in scientific understanding (Barstow et al., 2015). Discussing the relevance and validity of cited studies in a research paper introduction are common, powerful ways of addressing hypothesis risk. An author can note through relevance that there is a shortage of experiments in a given research area, which suggests the need for the author's study, or note through validity that there are relevant but underpowered studies that have produced inconclusive results, which also demonstrates the scientific utility of the author's proposed study.

Yet relevance and validity are themselves abstract and difficult concepts to grasp (Chalmers, Hedges, & Cooper, 2002; Cooper, 1982; 1998;). This study was designed to support novice scientists in applying the concepts of relevance and validity, and thus hypothesis risk, to their research paper introductions. Applying the underlying concepts of relevance and validity to the research literature was scaffolded through the node content in the diagrams. Appropriately evaluating risk was scaffolded through diagram structure. Therefore, effects of diagramming were expected on the amount of writing about relevance, validity, and risk. Statistically significant effects were found for all three.

For writing about different kinds of relevance, the largest effect was for context comparisons, in which a student may note, for example, that a cited study looked at the bystander effect in person, where instead their study is examining the effect in an online environment. Writing about all aspects of relevance was higher in the diagram condition, but some of the other components had smaller effects sizes. For writing about kinds of validity issues, there were effects of diagramming on all kinds, but of difference sizes. The largest effect was for experimental design, in which a student describes the basic structure of a study (e.g. experimental, correlational). The variation across the different components of relevance and validity can most likely be explained through differences in frequency, or applicability, and difficulty. For example, context comparisons are an easily accessible feature of studies, require little reasoning, and are frequently appropriate, as we would expect every published study to both describe the context it was conducted in and involve some kind of context difference. Experimental design is likely similar, where it requires little to no reasoning to extract and should be described clearly in any published study. On the other end of the spectrum (i.e. the smallest effect) was the evaluation dimension, which had the smallest effect across all dimensions across relevance and validity. Opportunities for evaluations are both rare (i.e. not every study will have notable confounds or power issues), and difficult (discovering validity issues requires a higher amount of reasoning and scientific understanding).

We may understand our findings related to hypothesis risk in a similar way. The intervention had its largest effect for risk addressed through uncertainty, in which, for example, a student makes an appeal to scientific ignorance in a particular area, that there has not been any or enough research conducted in their topic of interest to be conclusive. This is both a relatively accessible appeal (one can imagine a near-infinite number of unexplored research areas) and an easy one because the area a particular study is grounded in is generally made quite clear. Opposition, in contrast, is both more rare (topics with highly conflicted findings are limited), and more difficult (i.e. it requires a deep and nuanced understanding of multiple related articles). Indeed, even published papers in psychology address risk more frequently through uncertainty than through opposition (Barstow et al. 2015).

Overall, this study provides novel evidence from a randomized and controlled experiment that spatial representations (i.e. diagrams) can help novices in science better apply the concept of hypothesis risk to their research paper introductions. Building on prior work examining other effects of argument diagramming (Chryssafidou, 2014; Harrell, 2013; Nussbaum & Schraw, 2007;), this study provides further evidence that the utility of argument diagramming for improving some aspects of writing is robust across ontologies, especially because this ontology highly differs from conventional frameworks previously studied like those based on Toulmin's (1958) discipline-specific framework of argument.

Argument diagramming frameworks have greatly varied in complexity and content across the literature (Chryssafidou, 2014Griffin, Malone, & Kameenui, 1995; Harrell; 2013; Ozmen, 2011). Here we tested the benefits of a disciplinary-specific framework that prompts students for highly specific contents, and does not directly represent the rules of inference. We moved to this approach through iterative testing and refinement in our tested context: less disciplinary specific diagrams left too much unprompted for students to complete and produced diagrams that were too large to be useful in writing. But this may reflect the nature of introductions in psychology research, in which many papers, often not very closely related to the experimental situation at hand, must be reviewed. In other areas in which introductions cover more theory and less experimental work, a different kind of diagramming structure may be more relevant. Alternatively, if the goal is to improve writing experimental introductions in science more generally, rather than in psychology specifically (e.g., in a K-12 context), then a more general diagramming approach could be better.

The current study also uniquely broke down elements of writing in psychology about risk into key, discipline-specific components and shows that these individual components all benefit from the support of structured argument diagramming. Past research on computer-based tools for supporting problem-based learning (e.g., Quintana et al., 2004; White & Frederiksen, 1998) has suggested that students need additional supports and scaffolds to manage the complexity embedded into complex inquiry tasks. Here we explored a simple tool that could be added into a wide variety of instructional situations (e.g., in-class activity or homework assignment worksheets for highly structured or very open-ended inquiry projects), but focuses on particular issues in a relatively psychology-specific framework that attends to particular writing challenges in psychology research.

At a more nuanced level, our findings suggest that some elements of relevance, validity, and hypothesis risk may be more difficult than others for students to address, and that perhaps additional scaffolding could be built into a diagram ontology or an intervention to further support students in understanding and applying these concepts. For example, assignments could be created that involve reasoning about given studies and hypotheses that necessarily involve rarer and more difficult issues in relevance, validity, and risk.

There are limitations to the conclusions one can draw from this study, related to both our methods and analysis. For example, the inter-coder reliability of the detailed content of student papers was not ideal. However, this should not be of great concern to the conclusions drawn here because additional measurement noise reduces study power and yet statistically significant effects were still found. The primary issue in coding these elements is that student writing is often brief and informal, and thus it is not always clear what students intended to say.

Another potential issue relates to the operationalization of relevance and validity as the presence of reasoning about those topics rather than the accuracy of what was said (e.g., coding the presence of comments about correlational designs rather than checking whether the described study actually had validity problems). Thus, the current findings point more directly to an increased tendency to write explicitly about those topics rather than an increased ability to correctly assess validity in other research or correctly categorize the relevance of prior research to the current student. The point of these argument diagrams was to prompt thinking and inclusion of these concepts in writing, and did not directly provide instruction on these concepts. Thus, the operalization of the outcome measures appropriately matched the nature of the intervention. However, it remains an open question as to whether students also improved the quality of their reasoning about validity and relevance or were generally less likely to cite studies of low relevance to their hypotheses.

The complex nature of the tested diagram ontology as a hybrid representation (combining textual and spatial elements) makes it difficult at this point to separate the effects of diagram structure vs. node contents or understand the possible synergy between them. Given the contents of each and the posited mechanisms by which they influence the different studied aspects of writing, it is likely that both node contents and diagram structure were important.

It is encouraging that some differences between the experimental groups were present on the second paper in the course after a significant time delay — transfer in this domain has been under-discussed and rarely found. Although the differences only trended towards statistical significance, the effect sizes were moderate and point to power issues as the likely cause. Using a free, highly accessible diagramming tool for this study enables students to use the tool at their will—which, while practically beneficial—means that we are unable to determine whether differences on the second assignment represent a glimpse at true temporal transfer or if students actually revisited the tool. Student use of the tool was not measured outside the context of the first paper, where it was required. It is also unclear what the cognitive mechanism of transfer would be in this domain. Students may be improving their conceptual knowledge of different facets of argumentation (e.g. citation relevance), internalizing better structural and organizational knowledge of argumentation, or perhaps a combination of these and other improvements. Further research on transfer from diagramming activities could elucidate the relevant mechanisms at play.

3.5 CONCLUSIONS AND FUTURE DIRECTIONS

In future work, it would be interesting to determine what other elements of scientific writing and argumentation might be supported by this type of tool and ontology. Hypothesis risk, relevance, and validity, although important, do not constitute all of the components useful in constructing strong scientific arguments. To better understand the unique contributions of different elements of our ontology, content in students' diagrams could be compared directly to their writing. For example, one might compare the amount of text written about validity in a student's diagram to the amount present in their paper introduction.

In conclusion, the current study provides novel experimental evidence for the beneficial effects of argument diagramming activities on undergraduate students' writing in psychology. It also adds evidence to a growing literature on the robust benefits of diagramming and use of spatial representations in general in a variety of domains and applications, educational and professional. This study also introduces an operationalization of key components for writing in psychology and potentially science in general. These conceptualizations may lay the groundwork for additional fine-level analyses of science writing and provide a foundation for improving our theoretical understanding of the implicit and explicit components of scientific discourse. Finally, this work presents limited but exciting evidence for the transfer of diagramming benefits over time.

4.0 CONCLUSIONS

Together these two studies present a strong argument for the benefits of argument diagramming for improving research writing in psychology. In the first study, argument diagramming is shown to improve the relevance and validity of citations in student paper introductions, but does not address whether students are able to more effectively communicate the nature of these cited studies. The second study improves upon this by examining relevance and validity through student writing about these concepts, providing insight into their understanding and thought process in the literature review task. Additionally, the first study studies diagramming across three separate cohorts not randomized to condition, inviting the possibility of cohort effects. The second study boasts a tighter design—one cohort with a large sample of matched classrooms, randomly assigned to condition within pairs.

In both studies, argument diagramming is shown to help students bring in and write about evidence more relevant to their hypothesis, and also helps them to address hypothesis risk in different forms. In study one diagramming did not help students bring in more valid evidence, but in study one diagramming helps students to *write* more about the validity of their citations. This may be because writing about the nuances of scientific evidence represents an easier, foundational step towards selecting stronger evidence. Relevance, as we have conceptualized it here, should require less reasoning skills to access (i.e. explicitly addressed) when compared with validity (implicitly addressed), which may explain the difference in findings across these two studies. This is further evidenced by the trend-level effect for validity present in the first study, in that diagramming likely has an effect on the actual validity of cited studies as well, albeit a weaker one than that of related writing.

A notable history of research supports the utility of spatial representations for supporting reasoning in science for experts (Cheng & Simon, 1992; Larkin & Simon, 1987; Novick, 2000; Trafton, Trickett, & Farilee, 2005). The present research extends this body of evidence to show that diagramming can also help support novices in science wrestle with difficult meta-scientific concepts and their application in science writing. The present work also indicates that some types of spatial representations may afford unique benefits, and raises the possibility that the benefits of spatial representations may be maximized if the ontology used is adapted to the particular task at hand. It remains to be understood how this specialization impacts cross-domain transfer of learning gains afforded by the representation.

This research produces a number of new questions to be addressed in future work. In the second study, some temporal transfer of diagramming benefits was observed. Although only trend-level effects were found, effect sizes were relatively large and suggest the need for a study with tighter monitoring protocols and a larger sample size looking specifically at diagramming and knowledge transfer. The scarcity of transfer discussion in the existing spatial representation literature only further emphasizes the need for closer study. Additional analysis of the present data could shed light on the unique contributions of textual and spatial elements of hybrid representations, which may provide unique synergistic affordances above either alone. Lastly, neither of the two studies presented here studied student beliefs, motivations, or learning which may underlie the gains seen in writing, encouraging future work to develop a more holistic picture of how this type of intervention impacts students.

5.0 BIBLIOGRAPHY

Andrews, R. (1995). Teaching and learning argument. London: Cassell.

Andrews, R., & Mitchell, S. (2001). *Essays in argument*. Middlesex University Press.

Arum, R. & Roska, J. (2011). *Academically Adrift: Limited Learning on College Campuses.* Chicago: University of Chicago Press.

Barstow, B. J., Schunn, C. D., Fazio, L.K., Falakmasir, M.H., & Ashley, K. (2015). *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Chang, K.-E., Sung, Y. T., & Chen, I.-D. (2002). The effect of concept mapping to enhance text comprehension and summarization. *The Journal of Experimental Education, 71*1), 5-23. http://dx.doi.org/10.1080/00220970209602054

Chalmers, I., Hedges, L.V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation and the Health Professions, 25*(1), 12-37.

Cheng, P. C.-H. (1992). Diagrammatic reasoning in scientific discovery: modeling Galileo's kinematic diagrams. In H. Narayanan (Ed.), *AAAI technical report on reasoning with diagrammatic representations* (Report No. SS-92-02, pp. 33-38). Menlo Park: CA: American Association for Artificial Intelligence.

Cheng, P. C.-H. & Simon, H. A. (1992). The right representation for discovery: Finding the conservation of momentum. In D. Sleeman & P. Edwards (Eds.), *Machine learning: Proceedings of the Ninth International Conference* (pp. 62-71). San Mateo, CA: Kaufmann.

Chryssafidou, E. & Sharples, M. (2002). Computer-supported planning of essay argument structure. *International Society for the Study of Argumentation Proceedings 2002*.

Chryssafidou, E. (2014). *Argument diagramming and planning cognition in argumentative writing* (unpublished doctoral dissertation). University of Birmingham, Edgbaston, Birmingham, United Kingdom.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, *52*, 291-302.

Cooper, H. M. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Crowell, A., & Kuhn, D. (2014). Developing dialogic argumentation skills: A three-year intervention study. *Journal of Cognition and Development*, *15*(2), 363-381. http://dx.doi.org/10.1080/15248372.2012.725187

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2012). An evaluation of argument mapping as a method of enhancing critical thinking performance in e-learning environments. *Metacognition & Learning*, *7*, 219-244.

Gray, R. & Kang, N-H. (2012). The structure of scientific arguments by secondary science teachers: Comparison of experimental and historical science topics. *International Journal of Science Education*, *36* (1), 46-65.

Griffin, C. C., Malone, L. D., Kameenui, E. J. (1995). Effects of graphic organizer instruction non fifth-grade students. *The Journal of Educational Research,* 89(2), 98-107. http://dx.doi.org/10.1080/00220671.1995.9941200

Hahn, U. & Oaksford, M. (2012). Rational Argument. In K. J. Holyoak & R. G. Morrison (Eds.) *The Oxford handbook of thinking and reasoning* (pp. 277-300). New York: Oxford University Press.

Hand, B., Prain, V., & Wallace, C. (2002). Influences of Writing Tasks on Students' Answers to Recall and Higher-Level Test Questions. *Research in Science Education, 32*, 19-34. http://dx.doi.org/10.1023/A:1015098605498

Hand, B., Wallace, C. W., & Yang, E. (2004). Using a Science Writing Heuristic to enhance learning outcomes from laboratory activities in seventh-grade science: quantitative and qualitative aspects. *International Journal of Science Education, 26*(2), 131-149. http://dx.doi.org/10.1080/0950069032000070252

Harrell, M. (2008). No computer program required: Even pencil-and-paper argument mapping improves critical thinking skills. *Teaching Philosophy, 31*, 351-374.

Harrell, M. (2011). Argument Diagramming and Critical Thinking in Introductory Philosophy. *Higher Education Research & Development, 30*(3), 371-385.

Harrell, M. (2012). Assessing the Efficacy of Argument Diagramming to Teach Critical Thinking Skills in Introduction to Philosophy. *Inquiry*, *27*(2), 31-38. http://dx.doi.org/10.5840/inquirtyct201227210

Harrell, M. (2013). Improving First-Year Writing Using Argument Diagramming. *Proceedings of the 35th Annual Conference of the Cognitive Science Society.*

Keys, C. W., Hand, B., Prain, V., & Collins, S. (1999). Using the Science Writing Heuristic as a Tool for Learning from Laboratory Investigations in Secondary Science. *Journal of Research in Science Teaching, 36*(10), 1065-1084. http://dx.doi.org/10.1002/(SICI)1098-2736(199912)36:10<1065::AID-TEA2>3.0.CO;2-I

Kiuhara, S. A., Graham, S., & Hawken, L. S. (2009). Teaching writing to high school students: A national survey. *Journal of Educational Psychology*, *101*, 136-160. http://dx.doi.org/10.1037/a0013097

Knudson, R. E. (1992). Analysis of argumentative writing at two grade levels. *The Journal of Educational Research, 85*(3), 169-179. http://dx.doi.org/10.1080/00220671.1992.9944434

Kuhn, D. (2013). Reasoning. In P. Zelazo (Ed.), Oxford handbook of developmental psychology (pp. 744–764). New York, NY: Oxford University Press.

Kuhn, D., Zillmer, N., Crowell, A., Zavala, J. (2013). Developing norms of argumentation: Metacognitive, epistemological, and social dimensions of developing argumentative competence. *Cognition and Instruction, 31*(4), 456-496. http://dx.doi.org/10.1080/07370008.2013.830618

Kuhn, D., Hemberger, L, & Khait, V. (2016). *Argue with me: Argument as a path to developing students' thinking and writing* (2nd ed.). New York: Routledge.

Larkin, J. H. & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words.
Cognitive Science, 11(1), 65-100. http://dx.doi.org/10.1111/j.1551-6708.1987.tb00863.x

Leitão, S. (2003). Evaluating and Selecting Counterarguments: Studies of children's rhetorical awareness. *Written Communication, 20*(3), 269-306. http://dx.doi.org/10.1177/0741088303257507

Loll, F. & Pinkwart, N. (2013). LASAD: Flexible Representations for Computer-Based Collaborative Argumentation. *International Journal of Human-Computer Studies*, *71*(1), 91-109. http://dx.doi.org/10.1016/j.ijhcs.2012.04.002

Lynch, C. (2014). *The Diagnosticity of Argument Diagrams*. Doctoral Dissertation, University of Pittsburgh.

Lynch, C., Ashley, K. D., & Chi, M. (2014). Can Diagrams Predict Essay Grades? *Lecture Notes in Computer Science*, *8474*, 260-265. http://dx.doi.org/10.1007/978-3-319-07221-0_32

Mandler, J. M. & Ritchey, G. H. (1977). Long-term memory for pictures. *Journal of Experimental Psychology, 3*(4), 386-396. http://dx.doi.org/10.3758/BF03196949

McCagg, E. C., & Dansereau, D. F. (1991). A convergent paradigm for examining knowledge mapping as a learning strategy. *Journal of Educational Research, 84*, 317-324. http://dx.doi.org/10.1080/00220671.1991.9941812

Novick, L. R. (2000). Spatial diagrams: Key instruments in the toolbox for thought. *Psychology of Learning and Motivation, 40(2000)*, 279-325. http://dx.doi.org/10.1016/s0079-7421(00)80023-7

Nussbaum, E. M., Schraw, G. (2007). Promoting Argument-Counterargument Integration in Students' Writing. *The Journal of Experimental Education, 76*(1), 59-92. http://dx.doi.org/10.3200/JEXE.76.1.59-92

Oostdam, R. J., & Emmelot, Y. W. (1991). Education in argumentation skills at Dutch secondary schools. *Proceedings of the Second International Conference on Argumentation.* Amsterdam: Sic Sat.

Oostdam, R., Glopper, K. D., & Eiting, M. H. (1994). Argumentation in written discourse: secondary school students' writing problems. *Studies in Pragma-dialectcs.* Amsterdam: Sic Sat.

Osborne, J., Simon, S., Christodoulou, A., Howell-Richardson, C., & Richardson, K. (2013). Learning to argue: A study of four schools and their attempt to develop the use of argumentation as a common instructional practice and its impact on students. *Journal of Research in Science Teaching, 50*(3), 315-347. http://dx.doi.org/10.1002/tea.21073

Ozman, R.G. (2011). Comparison of two different presentations of graphic organizers of recalling information in expository texts with intellectually disabled students. *Educational Sciences: Theory and Practice 11*(2), 785-793.

Paivio, A. (1986). Mental representations: A dual coding approach. New York: Oxford University Press.

Perkins, D.N., Faraday, M. and Bushey, B. 1991. Everyday reasoning and the roots of intelligence. In: J.F. Voss; D.N. Perkins and J.W. Segal, eds. Informal Reasoning and Education. Hillsdale, NJ: Lawrence Erlbaum Associates, (Ann Arbor: University of Michigan Press), 83-105.

Quintana, C., Reiser, B. J., David, E. A., Krajcik, J., Fretz, E., Duncan, R. G., et al. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences, 13*(3), 337-386.

Schwarz, B. B., Neuman, Y., Gil, J., & Ilya, M. (2003). Construction of Collective and Individual Knowledge in Argumentative Activity. *Journal of the Learning Sciences, 12*(2), 219-256. http://dx.doi.org/10.1207/S15327809JLS1202_3

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior, 6*(1), 156-163. http://dx.doi.org/10.1016/S0022-5371(67)80067-7

Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology, 25*(2), 207-222. http://dx.doi.org/10.1080/14640747308400340

Stapleton, P. (2001). Assessing critical thinking in the writing of Japanese university students: Insights about assumptions and content familiarity. *Written Communication, 18*(4), 506-548. http://dx.doi.org/10.1177/0741088301018004004

Stegmann K., Weinberger A., Fischer F. Aktives Lernen durch Argumentieren: Evidenz für das Modell der Argumentativen Wissenskonstruktion in Online-Diskussionen. [Active learning by argumentation: Evidence for the model of argumentative knowledge construction in online discussions.] Unterrichtswissenschaft. 2011;39:231–244.

Stull, A. T. & Mayer, R. E. (2007). Learning by doing versus learning by viewing: Three experimental comparisons of learner-generated versus author-provided graphic organizers. *Journal of Educational Psychology, 99*(4), 808-820. http://dx.doi.org/10.1002/pfi.20028

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295-312. http://dx.doi.org/10.1016/0959-4752(94)90003-5

Toulmin, S. E. (1958), The Uses of Argument, Cambridge University Press, Cambridge.

Trafton, J. G., Trickett, S. B., & Mintz, F. E. (2005). Connecting internal and external representations: Spatial transformations of scientific visualizations. *Foundations of Science, 10,* 89-106. http://dx.doi.org/10.1007/s10699-005-3007-4

Van Amelsvoort, M., Andriessen, J., & Kanselaar, G. (2007). Representaitonal tools in computer-supported collaborative argument-based learning: How dyads work with constructed and inspected argumentative diagrams. *The Journal of the Learning Sciences*, *16*(4), 485-521.

White, B. & Frederiksen, J. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3-118.