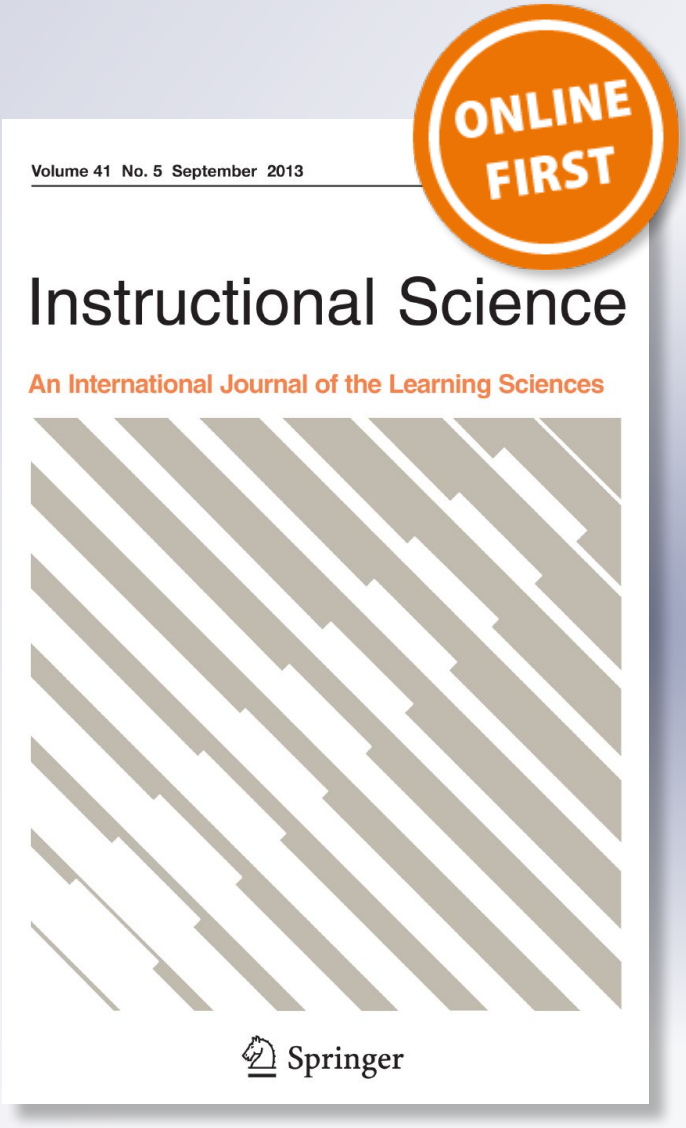
Experimental evidence for diagramming benefits in science writing

Brendan Barstow, Lisa Fazio, Christian Schunn & Kevin Ashley

Instructional Science
An International Journal of the Learning Sciences

ISSN 0020-4277

Instr Sci
DOI 10.1007/s11251-017-9415-3



Springer

Instr Sci
DOI 10.1007/s11251-017-9415-3

Experimental evidence for diagramming benefits in science writing

Brendan Barstow[1] · **Lisa Fazio**[2] · **Christian Schunn**[1] · **Kevin Ashley**[1]

Received: 20 June 2016 / Accepted: 2 May 2017

Abstract 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 replications (i.e. fail to find and include opposition to their hypothesis) or in contrast include no strong support for their hypothesis (i.e. relevant, valid evidence). How can we support novice scientists in generating and defending high quality hypotheses? A long history of research supports the affordances provided by structured representations of complex information. More recently, argument diagramming has gained traction in instruction for philosophy, social studies, and law. However, its effectiveness for supporting students in science is relatively untested. The purpose of the current study was to test the effectiveness of a simple argument diagram optimized for supporting students' research writing in psychology. Two groups of undergraduate students in research methods lab courses were randomly assigned to diagramming support or no support. In the research papers, 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 signs of transfer to a second paper written later in the course that did not require use of the diagramming tool.

Keywords Argument diagram · Science writing · Scaffolds · Argumentation

✉ Christian Schunn
schunn@pitt.edu

[1] LRDC, University of Pittsburgh, 3939 O'Hara St, Pittsburgh, PA, USA

[2] Psychology and Human Development, Vanderbilt University, Hobbs 315A, Nashville, TN, USA

Published online: 06 May 2017

Introduction

Argumentation is a central aspect of science, and is thought to be particularly important both as a means for learning and a skill to be learned (Osborne et al. 2013). Argumentation in science happens in multiple forms. One type of argumentation is organizing the results of a study towards a conclusion (Andrews 1995; Andrews and Mitchell 2001; Oostdam et al. 1994; Oostdam and Emmelot 1991), that is, writing the results and discussion sections of a paper. While complex, this aspect of writing tends to be manageable for secondary students (Hand et al. 2004), perhaps partially because the range of evidence to be integrated within student projects is relatively small.

Another aspect of argumentative writing that is less well-studied is arguing for the need for a study as in writing the introduction to a research report. This form of argumentative writing poses unique challenges for students (Osborne et al. 2013). They need to grasp a large body of conceptual, procedural, and epistemic knowledge and 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. Thus, the student must decide which research papers are relevant to his hypothesis, which are irrelevant and how to combine the relevant papers into a cohesive argument. This process is made more difficult by science instruction that obscures the argumentative, and frequently ambiguous nature of interpreting scientific evidence (Gray and Kang 2012).

For journal articles written by scientists, there are many complexities to writing a strong introduction, and it often takes years of practice to master all aspects. Indeed, it is an open area of scholarship to analyze research papers and uncover the many critical rhetorical aspects they contain, and how these aspects might be handled through different strategies (Berkenkotter and Huckin 2016). To name just a few aspects, there is arguing for the practical importance of the topic to society, arguing for the theoretical importance of the question, and reviewing differing prior claims about the question. When students begin to write introduction arguments, they are unlikely to be able to take on all the important aspects, and some aspects may involve domain knowledge that is well beyond them. For example, understanding the theoretical importance of the topic likely requires a deep understanding of a field.

And yet, current science education scholarship calls for students to engage in authentic practices, rather than overly-scripted and decontextualized tasks (NRC 2012). One approach to remain authentic, but stay within student capabilities, is to focus on a doable aspect of introductions that still brings meaning to the overall study being conducted: providing a rational for the study.

Types of argumentation in research paper introductions

Introductions to research papers have a central feature that makes their argument structure unique. In typical dialogic argumentation (e.g., in a results section or general discussion section), multiple competing perspectives may be explored, but the end goal is resolution in favor of one perspective. In contrast, research paper introductions seek to clarify an open question for which there is supporting evidence, but the prior evidence is insufficient. That is, writers present convincing arguments in favor of their hypotheses, but often also leave enough ambiguity that the issues still appear worthwhile to test, often described as a research gap or working hypothesis.

In other words, the hypothesis needs to be risky enough to be interesting, avoiding retesting settled science (Chinn and Brewer 1993) or attacking purely straw-men hypotheses (Klayman and Ha 1987). Although not usually described this way in research methodology textbooks, a recent systematic analysis of journal articles in psychology revealed that explicit writing about hypothesis 'risk' is a very common feature of published articles within top social, cognitive, and developmental psychology journals (Barstow et al. 2015). Further, these analyses indicated that hypothesis risk is typically 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. In expert writing, this kind of argumentation about risk can demand extensive knowledge of a field. But simpler forms of the argument can be made from more limited knowledge of the literature (e.g., provide some support and yet also identify a gap based on a small set of prior work). Indeed, even expert writers sometimes include the hedge phrase "to the best of our knowledge" in describing gaps. Given the common focus in secondary science on replication as hands-on science (Chinn and Malholtra 2002; NRC 2012), it is not surprising that explicit writing about hypothesis risk in research paper introductions is commonly missing in student work, and instead students simply list in a sequential form descriptions of related work (Barstow et al. 2017).

Focusing on risk while teaching how to write introductions also allows for students to practice applications of two core concepts in science, validity and relevance. For example, prior work in an area might suffer from validity problems such as only having correlational evidence or containing confounds, thereby creating the need for a study that addresses the gap in validity evidence. Alternatively, prior work may have been on measures, populations, or contexts quite different from the student's proposed study, and thus not relevant, providing a rational for hypothesis risk (i.e., a test of generalization). These two ways of establishing risk—validity problems in prior work or testing a previously supported hypothesis in a new setting—are the most common forms of identifying explicit risk in psychology research writing (Barstow et al. 2015). Learning to reason about validity (correlation vs. causation, confounds in an experimental design, poor measures) is commonly a core focus on research methods courses. In sum, this way of conceptualizing and structuring introductions may provide meaning to the science being taught in introductory courses, while providing an opportunity to practice core concepts and skills. However, this conceptualization of introductions is not typically taught in research methods classes and does not naturally appear in student writing (Barstow et al., in press). Given its inherent complexities (discussed below), students will likely need some support to address it properly.

Argument diagramming

Argument diagramming is one way of providing that support and explicitly structuring science writing as an argument. At the basic level, students may fail to include strong support for their hypothesis (Schwarz et al. 2003), while at more intermediate levels, students may fail to include any reason to doubt their tested hypothesis (i.e., fail to note weakness in evidence or possible counter-evidence) (Nussbaum and Schraw 2007). Failure to consider alternatives has sometimes been considered a skill deficit (Crowell and Kuhn 2014; Kuhn et al. 2016), 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 integrate these conflicts into a coherent argument for their hypothesis. An argument diagram makes visually salient the absence of support, alternative explanations, or counterarguments.

Researchers have been studying the affordances of different representation formats for problem solving and learning for nearly half a century (Mandler and Ritchey 1977; Paivio 1986; Shepard 1967; Standing 1973). On the one hand, spatial representations have been studied as important external tools that afford benefits to reasoning and problem-solving (Cheng 1992; Cheng and Simon 1992; Larkin and Simon 1987; Novick 2000; Trafton et al. 2005), such as the use of diagrams of kinematic equations or hierarchical diagrams of evolutionary relationships. On the other hand, textual representations can also be structured, as with a structured abstract, to help users and learners. Regardless of format, structured representations help the viewer to attend to task-relevant features while reducing distractions from task-irrelevant aspects of the problem-solving situation.

Argument diagrams are a particular form of structured representation that have been successfully employed as instructional tools in education (e.g., Dwyer et al. 2012; Harrell 2013; Suthers and Hundhausen 2003; Van Amelsvoort et al. 2007). Argument diagrams 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, these might be a hypothesis, various study findings, and counterarguments. Argument diagrams have been shown to facilitate student retention. For example, in a social studies context, students who diagrammed novel learning material retained the information better than their classmates who did not diagram (Griffin et al. 1995). Argument diagramming also improves students' ability to critically analyze arguments (Dwyer et al. 2012; Harrell 2008, 2011, 2012, 2013; Suthers and Hundhausen 2003) and to generate them (Harrell 2013). Such diagramming also shows potential for helping students write argumentative essays across various disciplines such as science and social studies (Chryssafidou and Sharples 2002; Chryssafidou 2014), an important task that is a source of struggle for many students (Andrews 1995; Andrews and Mitchell 2001; Hahn and Oaksford 2012; Kuhn 2013).

What are the mechanisms by which diagramming could support reasoning about risk in terms of the relevance and validity of existing prior evidence? 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 notice the amount of evidence that is supporting or opposing.

The textual element of diagrams allows students to 'zoom in' on 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 appropriateness of the methods of the cited prior work. These aspects of the study are easily represented by text, but are not easily represented spatially.

Although there is a good conceptual 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 that the quality of college students'

argument diagrams was correlated with the quality of the research paper introductions that students later produced (Lynch et al. 2014; Lynch 2014). But it is unclear from this study whether the diagrams improved writing, or whether misconceptions revealed by students' diagrams were also manifested in the students' writing.

Another complication is that argument diagramming tools and frameworks likely differ in their effectiveness for supporting students. Prior studies in this area have typically employed domain-general Toulmin (1958)-style models (e.g., Stegmann et al. 2007, 2012; Harrell 2013) that lend themselves to cross-domain transfer. For this study, we will employ a psychology-specific argument ontology to target and support more nuanced concepts in the domain. Domain-general Toulmin diagrams are very inefficient at representing realistic cases because they force a detailed representation of each inference rule. In addition, there is nothing in the Toulmin diagram that forces student attention towards issues of relevance and validity, and thus they may fail to practice these target reasoning skills. A domain-specific representation can be made more efficient, so that it can be feasibly applied to the amount of information required to consider in an introduction, and it can scaffold student attention towards targeted skills.

The most similar prior research comes from a quasi-experimental study by Barstow et al. (2017) who found that argument diagrams plus peer review increased students' discussion of hypothesis risk in psychology introductions relative to a prior cohort of students who were not provided with such supports. Further, they tested different ontologies, one more general and one more psychology-specific, and found that the psychology-specific ontology produced the best overall performance in writing introductions. However, it is unclear whether these benefits were due to the argument diagrams or the peer review process. 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. In addition, the current study improves on the experimental design by including randomization to condition within a single cohort population.

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.

Methods

Participants

Participants were drawn from 182 undergraduate students enrolled in 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 course consisted of 2.5 lecture hours and 3 lab hours per week over a 14-week semester; the intervention took

place during three of the weekly lab sessions. The lectures were taught to the entire class and focused on theoretical issues in psychology research, such as different threats of validity, while the lab sections were smaller (18 to 24 students) and focused on practicing basic skills related to observational and experimental research.

The intervention was implemented in the lab sections. Participants were recruited into the experiment by their teaching fellows (TF) who taught the lab sections. Seven TFs participated in the experiment: six taught only one lab section and the seventh taught two lab sections. The six TFS who taught only one section were matched into pairs based on teaching experience and class characteristics (e.g., time of day). For example, there were two TFs who had never taught before and had lab sections early in the day. TFs within each of these pairs were then randomized into either the experimental (diagramming) or control condition. For the remaining TF, who taught two lab sections, one section was assigned to the experimental condition and one section was assigned to the control condition. Data from this TF, allows for comparison of effects of diagramming on students that controls for TF effects.

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 enabled 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.[1] 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.

Students were instructed to include the following elements in their diagrams: two hypotheses, citations of relevance to each hypothesis organized as either supporting or opposing study citations, and counterarguments for any evidence opposing a proposed hypothesis. These guidelines were presented to students along with a diagram template (see Fig. 1), 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.

In the example diagram shown in Fig. 2, the author proposes a study on college students learning Swahili words. The author hypothesized 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

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

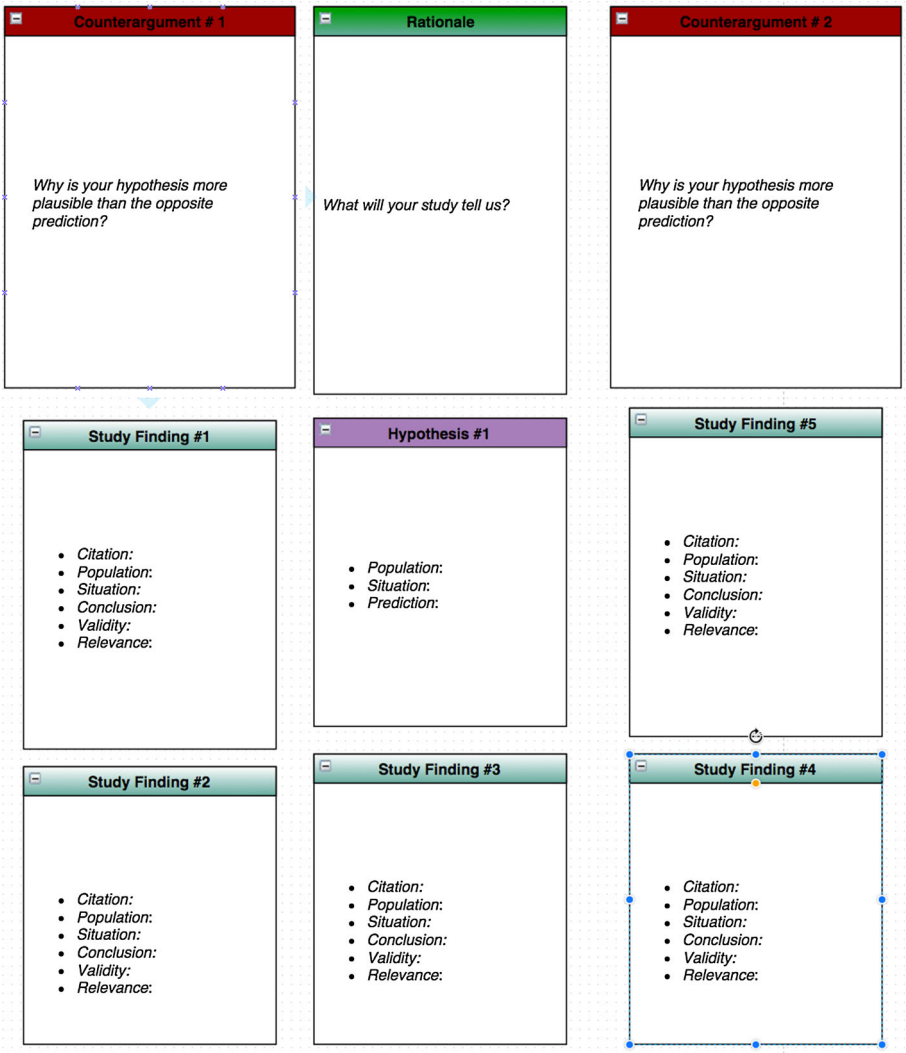


Fig. 1 Argument diagram template. Study findings are duplicated in the template because students are expected to need many of these

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.

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 Fig. 3, 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 Fig. 3).

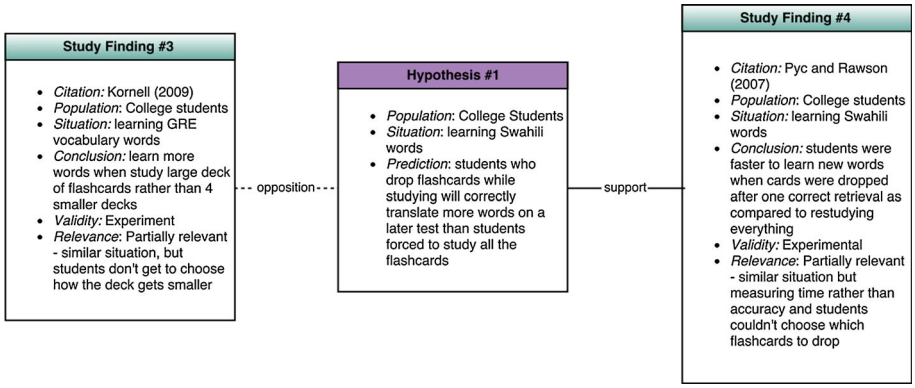


Fig. 2 Filled diagram from introductory demo activity

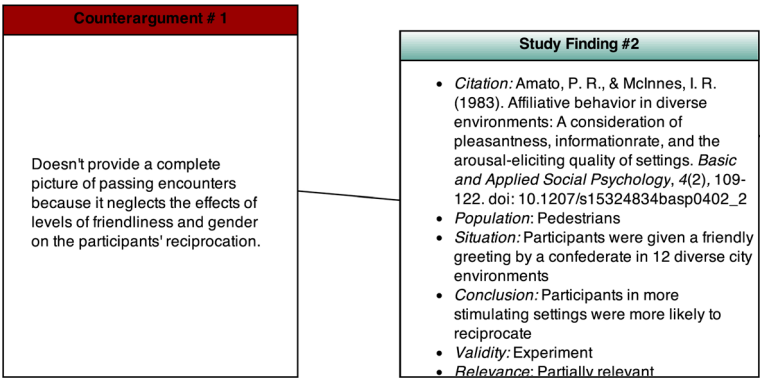


Fig. 3 Example (*subset*) of an actual student diagram created in Draw.IO

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 Fig. 3 rated Study Finding #2 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.

Procedures

Because the argument diagrams involve new and complex components for the students, it was necessary to include training on the purposes and procedures of the diagramming before students attempted to diagram their own studies. Therefore, 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 be successful. The lecture also conveyed the importance of relying on valid and relevant research in locating strong support for the presented

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 the control lab sections, with TF feedback, the students discussed finding research papers in more general terms, including how to identify the study designs, and the independent and dependent variables. Both groups considered issues of relevance and validity of prior work, but only the argument diagramming condition considered explicitly the role of relevance and validity relative to their own hypothesis risk. All sections then moved to brainstorming possible study questions.

In a later class, students in the experimental condition 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. Then 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. Finally, students spent time planning the design of their study. Students in the control condition spent more time planning the design of their study and receiving feedback from the TFs because they did not have to spend time learning how to diagram.

The paper assignment for both conditions included writing an abstract, introduction, methods, results, and discussion. Students were asked to include at least five research journal article 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. That is, all students were at least tacitly encouraged to take up hypothesis risk in their introductions, providing evidence for and against their hypothesis, although only the experimental group was given the hypothesis risk term explicitly. Further, all students received standard lectures in psychology research methods that discuss the purposes of research as forwarding understanding of psychological effects and processes using various study designs that vary in terms of internal and external validity.

Students also completed a second paper assignment one month later 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 unlikely that any cross-contamination occurred between paper 1 and paper 2. We did not initially plan to collect data from this second assignment, 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 discussion of risk using the protocol outlined below.

Measures

Risk coding scheme

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

Although Barstow et al. also found a third form, Risk through Difficulty (i.e., a complex set of predictions tested in an elaborate design), this form was rare and is not applicable to student experiments.

Risk through uncertainty (RU) occurs when the author claims that there is only insufficient (low relevance) or problematic evidence (low validity) for his/her hypothesis(es) in the literature, for example, "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 et al. 2015).

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, for instance, "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 et al. 2015)." Opposition does not itself involve relevance or validity issues in noting a conflicting prior study. However, the resolution of the opposition (i.e., why the student's hypothesis might stand despite the prior counter-evidence) does involve invoking either relevance (e.g., the conflicting result occurred in a different population or situation) or validity (e.g., the conflicting result had a confound).

Using these definitions, we coded the introductions of each student paper by annotating individual sentences that addressed risk, and coded each paper based on the types of risk addressed, regardless of the number of instances. For example, if a paper had four sentences tagged as RU and one as RO, that paper 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 the first author and a subset of these ($n = 25$) were then double-coded by a second coder ($\kappa = .91$).

Relevance and validity coding schemes

A subsample of student papers (102 evenly sampled across sections) were then coded at greater depth for relevance and validity of each citation using an iteratively developed coding scheme.

Relevance coding noted instances where authors discussed the *population* of a study (e.g., "College students"), the *context* of a study (e.g., "a busy street"), and *comparisons* of 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.

The validity codes included 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 blank (no code), $k = .40$. Disagreements between coders were resolved on a weekly basis by the first author 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 was removed from the reliability analysis, the reliability was quite high, $k = .79$.

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 within matched pairs of sections to insure the same pattern generally held across the data, rather than being driven by just one section.

Manipulation check

The intervention consisted of multiple components related to argument diagraming, including a lecture on risk and diagrams, example diagrams, and then completing their own diagrams. To provide further support to the notion that diagramming itself produces benefits (vs. just receiving a lecture and seeing examples), it is important to know that students did complete the diagrams, and did so in a substantive manner. 69 of the students submitted diagrams to their TFs in a form that could be automatically analyzed (i.e., as an xml instead of a pdf). On average, these diagrams included 4.6 citations, with 4.4 supporting and 1.0 opposing their 1.6 hypotheses (note that some citations could involve both supporting and opposing evidence). Table 1 presents the amount of text found in each of the nodes within each of the diagrams. On average, students completed the fields in substantive ways, writing several sentences or a paragraph of text for each field. Only one student wrote fewer than 10 words for any of the fields (summed across nodes of that type).

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 4 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

Table 1 Mean and standard deviation for words included in each student's diagram in each of the fields (summed across nodes of that type)

Node type	Field	Mean number of words (SD)
Hypothesis	Population	11.0 (7.3)
	Situation	34.2 (26.0)
	Prediction	45.7 (27.2)
Study rationale		70.1 (32.0)
Study finding	Population	81.7 (57.3)
	Situation	270.3 (180.8)
	Conclusion	279.5 (124.1)
	Validity	35.4 (21.9)
	Relevance	220.5 (197.3)
Counterarguments		65.5 (78.0)

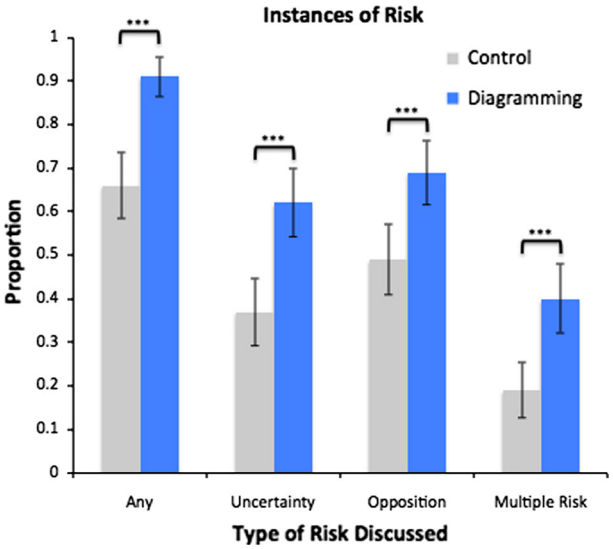


Fig. 4 Proportion of student papers addressing any risk (*any*), risk through uncertainty, risk through opposition, and a combination of risk types with SE bars. *$p < .05$, **$p < .01$, ***$p < .001$

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, $p < .001$, (100% vs 65%; $d = 1.02$) and more writing about RU, χ^2 (1, $n = 42$) = 4.8, $p = .03$, (82% vs. 50%; $d = 0.71$) in the diagramming condition, but no significant difference across conditions in writing about RO, $\chi^2(1, n = 42) = 0.8$, $p = .37$ (59% vs. 50%; $d = 0.28$), 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 Fig. 5). 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.

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 2). Note that there was no significant difference in the mean number of citations per paper across the two conditions (4.6 vs. 4.5 citations, $t(101) < 1$). The mean difference in attention to relevance and validity 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 (0.3 instances; See Fig. 6).

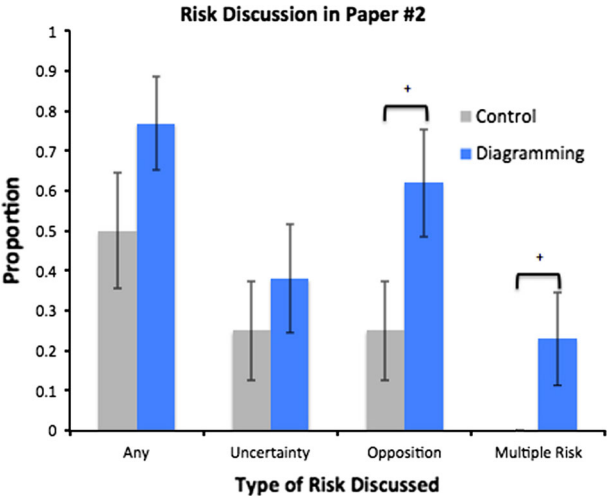


Fig. 5 Proportion of student papers addressing any risk, risk through uncertainty, risk through opposition, and a combination of risk types on paper 2 with SE bars. $*p < .05$, $+p < .10$

Table 2 *t* tests comparing writing about relevance and validity across conditions

	t	df	Sig.	Mean difference	*d*
Relevance					
Population	−2.6	136.5	.01	−0.54	0.44
Context	−5.1	134.9	.00	−1.12	0.88
Comparison	−3.0	134.5	.00	−0.57	0.52
Validity					
Sample size	−2.4	130.1	.02	−0.39	0.42
Exp. design	−2.3	151.1	.03	−0.53	0.37
Confounds	−3.5	92.4	.00	−0.51	0.73
Evaluation	−2.7	107.8	.01	−0.30	0.52
Citations	0.7	152.0	.47	0.20	0.11

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, $t(40) = 3.1$, $p = .003$, (1.7 vs. 0.6; $d = 0.98$), context, $t(40) = 5.6$, $p < .001$, (2.9 vs. 0.5; $d = 1.77$), comparisons, $t(40) = 2.4$, $p = .02$, (1.5 vs. 0.5; $d = 0.75$), and validity evaluations, $t(40) = 2.1$, $p = .048$, (0.5 vs 0.1; $d = 0.66$) than those in the control condition. However, there were no significant differences in this sample for writing about sample size (1.2 vs. 0.5), experimental design (2.1 vs. 1.8), or confounds (0.7 vs 0.1), even though the rates were always directionally higher in the diagramming condition.

Interrelationship of risk, relevance, and validity

One-way analyses of covariance (ANCOVA) were conducted to test the relationship between these dependent variables, spurred on by the moderate correlation observed between writing about relevance and writing about validity ($r = .51$, $n = 102$, $p < .001$).

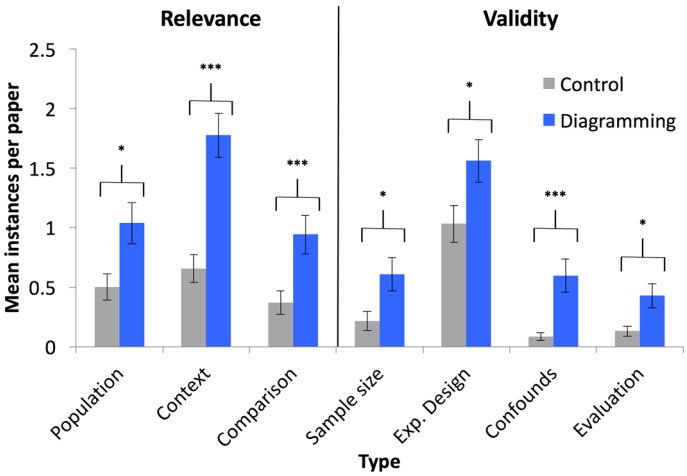


Fig. 6 Writing about relevance and validity separated by subcomponent on paper 1 with SE bars. $*p < .05$, $**p < .01$, $***p < .001$

See Fig. 7 for a visual representation of these relationships. While controlling for writing about relevance ($F(1,99) < 1$, $p = .37$) and validity ($F(1,99) = 3.71$, $p = .06$), condition had a significant effect on writing about risk ($F(1,99) = 17.59$, $p < .001$). While controlling for writing about risk and validity, condition had a significant effect on writing about relevance ($F(1,99) = 11.73$, $p = .001$). While controlling for writing about risk ($F(1,99) = 3.71$, $p = .06$), and relevance ($F(1,99) = 42.62$, $p < .001$), condition did not have a significant effect on writing about validity.

Discussion

Introductions to research papers are deceptively difficult to write both because of the complexity in establishing hypothesis risk and because of the complexity of understanding and using general research concepts such as relevance and validity. The currently studied intervention was designed to support novice scientists in applying the concepts of relevance and validity, and thus hypothesis risk, to their research paper introductions. 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,

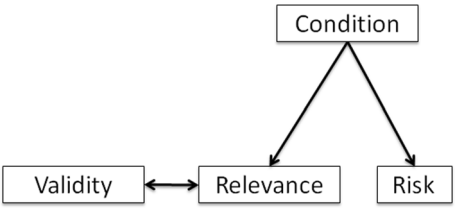


Fig. 7 Structure of relationships between writing characteristics and condition

but some of the other components had smaller effects sizes. For writing about validity issues, there were effects of diagramming on all issues, but of different sizes. The largest effect was for experimental design, in which a student describes the basic structure of a study (e.g., experimental, correlational).

This 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. One would expect every published study to both describe the context in which it was conducted and involve some kind of context difference. Experimental design is likely similar; it requires little or no reasoning to extract and should be described clearly in any published study. At the other end of the spectrum (i.e., the smallest effect) was the evaluation dimension, which had the smallest effect across all dimensions of 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, arguing that there has not been any or enough research 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 authors generally make quite clear the area in which a particular study is grounded. Opposition, in contrast, is both rarer (topics with highly conflicted findings are limited), and more difficult. 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).

Based on the ANCOVA results, these variables in students' writing appear to be moderately interrelated. The intervention's effects on writing about risk were independent of writing about relevance and validity, while writing about relevance was partially mediated by writing about validity and writing about validity was completely mediated by writing about relevance. First, regarding the independence of risk, the overall spatial structure of the diagram draws attention to risk, whereas the contents of the textboxes draws attention to relevance and validity; thus, the differential ways in which students attended to spatial structure vs. node contents may have led to independent effects. Second, the mediation of validity differences via relevance differences may have been caused by the low attention paid to validity in the diagram compared to relevance; it may be that only those students who understood the relevance of citations enough to address it in their papers were also able to address their validity.

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 (Barstow et al. 2017; Chryssafidou 2014; Harrell 2013; Nussbaum and Schraw 2007), this study provides further evidence that the utility of argument diagramming for improving some aspects of writing is robust across argument diagramming ontologies, especially because this ontology differs greatly from conventional frameworks previously studied like those based on Toulmin's (1958) framework of argument.

Argument diagramming frameworks have varied greatly in complexity and content across the literature (Chryssafidou 2014; Griffin et al. 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 is unique in its breaking down elements of writing in psychology about risk into key, discipline-specific components and its demonstration 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 and 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). It 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 less frequently encountered 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. Although this additional measurement noise reduces study power, statistically significant effects were still found, suggesting little concern about false negatives in this case. 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.

The current intervention necessarily had multiple components to introduce students to the diagramming tool and framework prior to completing the diagrams. This multi-part nature of the intervention raises the possibility that one of the components (e.g., just the lecture on risk or just the diagramming of example papers) produced some or all of the observed condition differences. However, it is important to note that students in the control condition had many lectures on experimental validity, and the additional lecture exposure of the diagramming group to validity in the context of diagrams was unlikely to produce substantial effects on their writing. In general, the largest time difference between the conditions was the time spent diagramming (vs. otherwise preparing their study). But future studies would explore the (likely smaller) benefits of just lecture-exposure to the concepts of hypothesis risk.

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 was actually correlational). 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. 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 operationalization 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. Indeed, findings from a lab study by Suthers and Hundhausen (2003) suggest that representations per se can have diverse effects on problem solving and learning. As noted in the introduction, the contents of nodes and diagram structure in the current study were designed to support particular reasoning and writing: node contents to support documentation of relevance and validity, and diagram structure to support overall evaluation of hypothesis risk. Therefore, it is likely that both node contents and diagram structure were important. However, it is possible that any approach which provides similar structure to students' thinking about the literature and the relationship of their hypothesis to the literature will likely have similar effects (e.g., a highly structured outline tool).

It is encouraging that some differences between the experimental groups in their use of risk 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.

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, are not the only 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's introduction.

In conclusion, building on prior quasi-experimental evidence of similar interventions (Barstow et al. 2017), the current study provides experimental evidence for the beneficial effects of argument diagramming intervention on undergraduate students' writing in psychology. It also adds evidence to a growing literature on the benefits of diagramming and use of structured 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. This operationalization 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 some limited evidence for the transfer of diagramming benefits over time.

Acknowledgements Work on this project was funded by Grant IIS-1122504 from the National Science Foundation to the 3rd and 4th authors.

References

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

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

Barstow, B., Fazio, L., Lippman, J., Falakmasir, M., Schunn, C., & Ashley, K. (2017). The impacts of domain-general vs. domain-specific diagramming tools on writing. *International Journal of artificial intelligence in education*.

Barstow, B. J., Schunn, C. D., Fazio, L. K., Falakmasir, M. H., & Ashley, K. (2015). Improving science writing in research methods classes through computerized argument diagramming. In *Proceedings of the 37th annual conference of the cognitive science society*.

Berkenkotter, C., & Huckin, T. N. (2016). *Genre knowledge in disciplinary communication: Cognition/culture/power*. Routledge.

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.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*(1), 1–49.

Chinn, C. A., & Malhotra, B. A. (2002). Children's responses to anomalous scientific data: How is conceptual change impeded? *Journal of Educational Psychology, 94*(2), 327.

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

Chryssafidou, E., & Sharples, M. (2002). Computer-supported planning of essay argument structure. *International society for the study of argumentation proceedings 2002*.

Crowell, A., & Kuhn, D. (2014). Developing dialogic argumentation skills: A three-year intervention study. *Journal of Cognition and Development, 15*(2), 363–381. doi: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 on fifth-grade students. *The Journal of Educational Research, 89*(2), 98–107. doi: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., 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. doi: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. doi:10.5840/inquirtyct201227210.

Harrell, M. (2013). Improving first-year writing using argument diagramming. In *Proceedings of the 35th annual conference of the cognitive science society.*

Holden, M. P., Newcombe, N. S., & Shipley, T. F. (2015). Categorical biases in spatial memory: The role of certainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(2), 473–481.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*(2), 211.

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

Kuhn, D., Hemberger, L., & Khait, V. (2016). Tracing the development of argumentive writing in a discourse-rich context. *Written Communication, 33*(1), 92–121.

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

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. doi: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. doi:10.3758/BF03196949.

Novick, L. R. (2000). Spatial diagrams: Key instruments in the toolbox for thought. *Psychology of Learning and Motivation, 40*(2000), 279–325. doi: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. doi:10.3200/JEXE.76.1.59-92.

Oostdam, R. J., & Emmelot, Y. W. (1991). Education in argumentation skills at Dutch secondary schools. In *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. doi:10.1002/tea.21073.

Ozmen, R. G. (2011). Comparison of two different presentations of graphic organizers in 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.

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. doi: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. doi:10.1016/S0022-5371(67)80067-7.

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

Stegmann, K., Weinberger, A., & Fischer, F. (2007). Facilitating argumentative knowledge construction with computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning, 2*(4), 421–447.

Stegmann, K., Wecker, C., Weinberger, A., & Fischer, F. (2012). Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. *Instructional Science, 40*(2), 297–323.

Suthers, D. D., & Hundhausen, C. D. (2003). An experimental study of the effects of representational guidance on collaborative learning processes. *The Journal of the Learning Sciences, 12*(2), 183–218.

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

Toulmin, S. E. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

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. doi:10.1007/s10699-005-3007-4.

Van Amelsvoort, M., Andriessen, J., & Kanselaar, G. (2007). Representational 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.

Westerman, D. L., Lanska, M., & Olds, J. M. (2015). The effect of processing fluency on impressions of familiarity and liking. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(2), 426–438.

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