Science
Education

Promoting Students' Attention to Argumentative Reasoning Patterns

ANDY R. CAVAGNETTO,[1] KENNETH J. KURTZ[2]
[1]*Department of Teaching & Learning and School of Biological Sciences, Washington State University, Pullman, WA 99164, USA;* [2]*Psychology Department, Binghamton University (SUNY), Binghamton, NY 13902, USA*

Received 1 May 2014; accepted 30 November 2015
DOI 10.1002/sce.21220
Published online 29 April 2016 in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: Argument-based interventions in science education have largely been motivated by the perspective that students lack knowledge of argument. Recent studies, however, suggest that contextual factors influence students' argument quality. The authors hypothesize that a key limiting factor lies in students' abilities to recognize when to employ knowledge related to argumentative reasoning. Students hold knowledge that can remain inert if the context fails to trigger access. The article reports on two experiments exploring the influence of context on recognition of argumentative reasoning. Each experiment used performance on a categorization task as the dependent variable. The first experiment (E1) tested the effect of targeted cues on participants' identification of fallacious reasoning patterns. The second experiment (E2) examined the role of scenario context. E1 revealed a significant advantage in promoting detection of fallacious reasoning patterns when targeted cues were present. E2 suggests that the broad context of scenarios also influences reasoning patterns. These findings reveal that lack of knowledge of argument may not be the principal constraint on students' argument abilities. Rather, contextual details can have a significant influence on activation of knowledge bases that contribute to argumentative reasoning. Instructional implications are discussed. *Sci Ed* **100:**625–644, 2016

Correspondence to: Andy R. Cavagnetto; e-mail: andy.cavagnetto@wsu.edu
Supporting Information is available in the online issue at wileyonlinelibrary.com.

INTRODUCTION

Argument has been viewed as a key aspect of science literacy because it authentically captures the often neglected social aspect of science (Driver, Newton, & Osborne, 2000) and holds considerable potential for enhancing student understanding of domain concepts. As such, efforts to centralize argument within instruction, particularly science, have been emphasized in recent standards in the United States (Common Core State Standards [CCSS] and Next Generation Science Standards [NGSS]) and across a wide variety of educational levels in science education research (Cavagnetto & Hand, 2011; Grooms, Sampson, & Golden, 2014; Kelly & Takao, 2002; Ryu & Sandoval, 2012).

Research around human argumentative reasoning, both outside of science education and within science education, has led the science education research community to view students as deficient in knowledge of argumentative reasoning skills. The early work of D. Kuhn and colleagues has been largely influential in the uptake of a deficiency perspective. Kuhn (1991) studied the argumentative thinking of 160 participants of diverse ages and ethnicity. Participants were prompted to take positions, provide counterarguments, rebuttals, and consider evidence in relation to the cause of three social topics: (a) prisoners returning to crime upon release, (b) children failing in school, and (c) unemployment. In general, participants conflated theory with evidence as demonstrated by failure to adequately support their claims with evidence and an inability to let go of or critique their initial positions. Furthermore, when provided with contradictory evidence, participants tended to assimilate the evidence to their theories or disregard the evidence rather than reconsider their theories in light of the evidence. These findings reflected earlier studies by Kuhn and colleagues (Kuhn, Amsel, & O'Loughlin, 1988). Many studies within K–12 science contexts also could be interpreted in support of a deficiency perspective. Jimenez-Aleixandre, Rodriguez, and Duschl (2000) studied how ninth graders construct and assess arguments during genetics instruction. Students most often provided claims, but did not connect them to evidence, warrants, or backings. Students also struggle to critically evaluate arguments. For example, rebuttals were rarely found in small group argumentative sessions with upper elementary science students (Cavagnetto, Hand, & Norton-Meier, 2010). When rebuttals did occur, they were generally weak—often short and failing to provide a clear idea of the nature of the rebuttal. Similar findings were reported by Sadler (2004) in his review of 13 studies related to informal reasoning in socioscientific contexts. Sadler concluded that (a) students often make unjustified claims and struggle to recognize opposing arguments during argument construction in socioscientific contexts, (b) students do not commonly use scientific evidence to inform their personal decision making, (c) content knowledge influences reasoning ability in contexts associated with the particular content, and (d) people are not very competent at analyzing and evaluating arguments. The deficiency perspective has had a considerable influence on how argument is implemented in science education contexts with a number of interventions designed around the development of student argumentative reasoning via explicit instruction (Cavagnetto, 2010; Nussbaum, Sinatra, & Poliquin, 2008; Osborne, Erduran, & Simon, 2004). For example in the Ideas Evidence and Argument in Science (IDEAS) project, students were instructed in Toulmin's pattern of argumentation and then practiced argumentation in the context of both social and scientific issues (Osborne et al., 2004). Similarly, McNeill, Lizotte, Krajcik, and Marx (2006) supported student written explanations of chemical change through explicit instruction in a *claims, evidence, reasoning* framework. Explicit instruction was then followed by prompts that faded in frequency as students gained experience with the framework.

More recently, studies have suggested that a deficiency perspective does not adequately represent students' abilities to engage in construction and critique of arguments or the

complexities of the processes influencing argumentation in classrooms. For example, a recent series of studies of 65 seventh-grade students illustrated that the goal of the argument task influences quality of the argumentative discourse (Felton, Garcia-Mila, & Gilabert, 2009; Garcia-Mila, Gilabert, Erduran, & Felton, 2013; Gilabert, Garcia-Mila, & Felton, 2012). The authors compared students who engaged in an argumentative task with the goal of reaching consensus to students who engaged in an argumentative task with the goal of persuading others. Students in the consensus group outperformed students in the persuasive group with respect to argument quality. One of the studies (Felton et al., 2009) examining the influence of task goal on student science content learning resulted in a nonsignificant trend toward the consensus group.

Even within a consensus-based task, there are contextual factors that are influencing students' engagement. In her study of middle school students, Berland (2011) found that classroom practices or norms of classroom activities influenced how students engaged in argumentation. For example, how students perceived themselves in relation to other students, i.e., an individual goal orientation may not require that students communicate or listen to one another as much as a shared goal orientation. Additionally, Berland found that more immediate aspects of the classroom, such as task structure, positioning, and access to information, also are influential. Similar findings were reported by Berland and Lee (2012) who found that social pressures can influence the nature of argumentative discourse. In their examination of student groups attempting to reach consensus, one important factor was students' abilities to *save face*. That is, recognition of the various perspectives allowed disparate groups to move toward consensus. In general, environmental characteristics (e.g., social interactions, classroom practices, teacher positioning, task purpose) influence students' and teacher's *framing* of the task which can result in variation in task engagement (Berland & Hammer, 2012). Berland and Hammer characterize framing as a "schema that organizes past experience" (p. 87). In doing so, how one frames the task, based on environmental clues, influences what students see as the purpose of the task, the norms of the task, and subsequently what they see as important for task engagement. Taken together, these studies suggest that students' argumentative reasoning is heavily influenced by how they interpret the task and other contextual clues that influence the knowledge that students bring to bear in response to the task.

We see framing or a contextual perspective of argument as an evolution of the literature related to knowledge transfer. The transfer literature highlights the so-called problem of inert knowledge (Whitehead, 1929) and wide-ranging evidence (see Barnett & Ceci, 2002; Day & Goldstone, 2012; Gick & Holyoak, 1980, 1983; Gentner, Rattermann, & Forbus, 1993; Kurtz & Loewenstein, 2007; Loewenstein, 2010; Ross, 1984; Spencer & Weisberg, 1986) of low levels of transfer of abstract knowledge to new situations—even when that knowledge is intact, available, and directly relevant. In essence, people fail to transfer knowledge to new situations either because they (a) do not have the knowledge or (b) do not recognize the appropriate time to access the knowledge. Importantly, we must consider *what is the unit of knowledge* in this transfer equation? What is being transferred? The *concept* or *skill* has historically been assumed to be the unit of knowledge transferred. However, scholars have called this assumption into question. Humans draw upon multiple cognitive resources or attributes when applying a concept or skill to a particular context (Galbraith, 1999; Hammer, Elby, Scherr, & Redish, 2005; Lamb, Vallett, Akmal, & Baldwin, 2014). From this component or manifold perspective of cognition, transfer is the coordinated recruitment of appropriate resources or attributes—which collectively lead to spontaneous application of the broader concept or skill. In argumentation, students are simultaneously accessing schemas related to reasoning, punctuation, interpersonal communication, and content information about the topic—not to mention their general world view, perceptions

of how knowledge is constructed, and what counts as knowledge. As such, every aspect of the context has the potential to cue or activate resources that lead toward or away from "successful" task completion. Thus, the manner with which a task is framed has considerable implications for outcomes.

Task framing occurs at multiple levels. Framing is considered a social process as students are continuously interpreting signals from the teacher, other students, and the classroom environment to define the broad task—Berland and Hammer (2012) present three such cases in their work. Framing is also a personal process. Individual students are continually reframing the task based on the environmental information that is continually flooding their senses. Therefore, a task may be framed at the classroom level as an argumentative task, but an individual student may still fail to fully draw on their cognitive resources leading to appropriate argumentative reasoning. For example, there is evidence that students struggle to draw upon scientific evidence when engaging in argument around socioscientific issues (Sadler, 2004). We can imagine a topic, such as human evolution, in which students contribute to an argument from a faith-based perspective. While the class may frame the activity as an argumentative discussion, an individual student may fail to fully evidence their position because they rely heavily on a particular religious perspective that does not require evidence in the same way that traditional logic or scientific argument does. So while the class may maintain the frame of an argumentative discussion, individual students' engagement in the argumentative discussion may be influenced by the extent to which they access their knowledge of argumentative reasoning. In the above example, we cannot say that the student is deficient in argumentative reasoning abilities. We only have evidence that argumentative reasoning was absent. In this case, the student drew more on their faith schema as opposed to their cognitive resources related to argumentative reasoning. In essence, the details of the scenario pushed them away from their knowledge of argumentative reasoning by cueing knowledge of their faith. To be clear, we are not suggesting that only their knowledge of faith is activated. Argumentative reasoning is activated, but the student's faith schema changes the relative amount with which they rely on argumentative reasoning for task completion.

Previous studies in science education contexts have illustrated that students often fail to engage in high-quality argumentation (often evidenced by their inability to engage in argumentative rebuttals). This has presumed to be a deficiency in students' knowledge of argumentation. An alternative explanation is that students do not realize that a situation calls for a rebuttal—that is, students may be failing to activate their knowledge-bases related to argumentative reasoning, i.e., contextual details may guide students away from the cognitive resources most relevant for high-quality argumentative reasoning. We hypothesize that people (including school aged students) hold some knowledge to successfully engage in argumentative reasoning but that the context may influence their recognition of when to draw on those knowledge bases. To be clear, such knowledge is unlikely to be formal and systematic like that of a logician—it may not even exist in a form that can be effectively verbalized. Knowledge that may be characterized as weak, poorly structured, inaccurate, or inchoate is still knowledge that matters—perhaps the problem is not that students know so very little about argumentative reasoning, but instead that they fail to leverage what they do know. That is, when an individual experiences actual examples of good and bad reasoning, they are unlikely to activate the cognitive resources related to reasoning itself – and this might go some distance in explaining poor performance by students in their ability to argue.

This study seeks to build on the developing context-dependent perspective on argumentative abilities by testing whether student recognition of argumentative reasoning can be advanced by modifying contextual cues and reducing extraneous information that potentially divert attention away from the reasoning patterns. While previous work has focused on changing the task goals (toward consensus building) or how the broad task is

interpreted, this study targets contextual cues within the broader goal of the task—those that most directly influence personal negotiation of meaning. Modification of contextual cues is intended to provide immediate payoff in terms of improved reasoning performance (as explored in our studies) and to serve as a means for promoting effective teaching and learning.

A major issue that arises is selecting an appropriate measure of argument performance. Toulmin's argument pattern (TAP) has been commonly used in science education research to analyze student abilities. We see failure to engage in appropriate argumentative reasoning resulting from lack of knowledge or the failure to activate that knowledge. The TAP analysis (Erduran, Simon, & Osborne, 2004) does not account for the failure to recognize when to employ or access knowledge of argumentative reasoning. As such, we chose to employ a task that has been well established in the psychological literature related to transfer to examine whether failure to activate knowledge is a limiting factor.

In a classic study, Chi, Feltovich, and Glaser (1981) used a sort task to demonstrate the nature of expertise: Physics experts were able to sort sets of problems based on the underlying solution principles, whereas novices sorted based on their surface properties (objects and settings). Subsequently, sorting has been used to evaluate the nature and quality of domain knowledge in psychological (e.g., Rottman, Gentner, & Goldwater, 2012; Shafto & Coley, 2003) and science education research (e.g., Smith et al., 2013; Stains & Talanquer, 2008). Importantly, in the present work, we extend the technique by comparing sort outcomes for the same population under different contexts—as opposed to comparing populations (novices versus experts). In this sense, an analogy is suggested between the power of expertise and the power of context—this holds considerable promise, because if context can be effective, it is certainly far easier to put in place than the development of expertise. Along these lines, we assume generally equivalent domain knowledge across experimental conditions, but our question is whether we can create differences in the ability to access and invoke existing knowledge.

Our specific goal is to use sort performance to measure the efficacy of cues that allow a student to effectively see examples as being valid or fallacious reasoning. Within the sort task, we seek to leverage contextual cues that bridge the gap between actual examples of reasoning and abstract principles related to reasoning. The cues have to do with how the task of evaluating examples of reasoning is presented, i.e., the use of instructions and the form and content of the materials. More specifically, we manipulate the degree of specificity of task instructions and the presence of extraneous information that could cloud one's view of the quality of the reasoning embedded within the task content. Can we get students to sort more like domain experts—not by turning them into actual experts on argument—but simply by providing cues that enhance the likelihood that students will activate the knowledge they have related to argumentative reasoning? To explore this possibility, the current study is guided by the following research questions:

RQ1: Can detection of fallacious argumentative reasoning patterns be increased by providing targeted cues in the form of more defined directions, precise language, and reduced extraneous information?

RQ2: Does context influence activation of argumentative reasoning knowledge, i.e., are participants better able to detect fallacious reasoning patterns in scientific or everyday contexts?

Based on the aforementioned literature on framing and spontaneous transfer, we expect that variation in context will influence activation of argumentative reasoning knowledge. Participants operating without targeted cues will struggle, whereas participants who receive

examples and task instructions designed to help them make the connection between an example and existing knowledge about argument quality will fare better.

CONTEXT OF STUDIES

We conducted the studies upon noticing poor performance on a categorization task targeting argumentative reasoning in everyday contexts. The categorization task required participants to sort six scenarios (one on each of six note cards) into two equal-sized groups based on common argumentative reasoning patterns within the scenarios. Student abilities to successfully sort and explain their rationales would suggest that they are able to recognize common fallacious reasoning patterns and thus hold some knowledge of appropriate and inappropriate argumentative reasoning patterns. During the spring semester, we engaged undergraduate students in the categorization task with minimal supports and found performance to be low with regard to sort success. The subsequent fall we conducted a study that included targeted cues embedded within the directions and sort card scenarios and simultaneously tested the effect of context (everyday or scientific) on attention to reasoning patterns. In that study, students completed three sort tasks. We accessed the first sort task of each participant (from the fall semester) in the everyday context group to compare the influence of targeted cues with the control group from the previous spring. This comparison will be reported as Experiment 1. The comparison of context utilizing all three sorts per student during the fall semester (everyday vs. scientific) will be reported as Experiment 2. Figure 1 provides an overview of the two experiments.

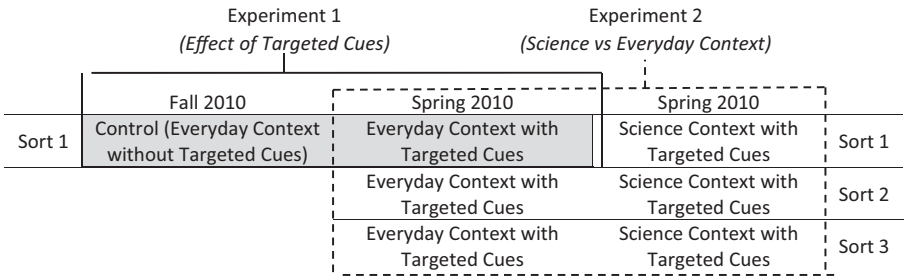

Figure 1. overview of two experiments.

EXPERIMENT 1

Experiment 1 (E1) follows a quasi-experimental design to assess the effect of targeted cues (e.g., defined directions, precise language, and reduced extraneous information in scenarios) on participants' identification of logical fallacies. The dependent variables in the study were (a) sort outcomes and (b) ability to provide an explanation showing that the basis of the sort was related to the reasoning in the examples. The logic guiding the experiment is as follows: If recognizing when to employ knowledge of argumentation is a factor in argumentation ability, then one would expect to see a difference in categorization (sorting) success depending on the targeted cues. If participants are inhibited due to a lack of knowledge about argumentative reasoning, then targeted cues (in the form of more defined directions, precise language, and less extraneous information in the sort scenarios) should not result in improved performance on the sort task. Thus, if our hypothesis for RQ1 is substantiated, then activation of argumentative reasoning knowledge is an important factor

TABLE 1
Study 1 Participant Majors (Reported as Relative Frequencies)

	Control Group	Targeted Cues Group
Natural sciences	27	27
Mathematics/engineering	8	14
Management	14	5
Humanities	22	35
Social sciences	29	19

that needs to be accounted for as we seek to enhance the level of argumentation in science learning environments. Substantiation would also suggest that the study participants do hold a basic understanding of poor and appropriate reasoning.

Method

Participants. Eighty-eight undergraduate students from a large state university in the northeast United States participated in the study. Most participants were enrolled in an upper level undergraduate education course *Teaching, Learning, and Schooling* designed for students to explore teaching as a career. A small percentage were enrolled in an upper level biology course, *Entomology*. Participants (63% female, 37% male) were juniors or seniors and represented a wide range of undergraduate majors at the university. Students were not randomly assigned as the Control and Targeted Cues groups were not run in parallel (as previously mentioned in the Context of Studies section of this article); however, participants were recruited in the same manner from the same population of students (juniors and seniors across a range of undergraduate majors) for both the Control and Targeted Cues groups. One clear confounding variable was time of participation (spring semester or fall semester); we are aware of no likely basis by which to attribute an impact to this difference. The Control trial was conducted during the spring semester, whereas the Targeted Cues trial occurred during the subsequent fall semester (as illustrated previously in Figure 1). Fifty-one students (56% female, 45% male from a diverse array of majors) participated in the Control group, whereas 37 students (73% female, 27% male from a diverse array of majors) participated in the Targeted Cues group. The difference in sample sizes resulted from maintaining consistency in recruiting from the same course during semester two of the study. As will be explained in E2, we ran two treatment groups (everyday context or science context) during the fall semester and this resulted in splitting the sample of fall semester participants ($n = 74$) so as to maintain consistency in context (comparing the everyday context group in the fall with the everyday context-based scenarios of the spring). Attempting to maintain consistency in context of the sort scenarios required that we use only first sort data from the fall semester everyday group ($n = 37$) during E1. Participants in the Control group were primarily senior academic standing (86% with 14% being juniors). The Targeted Cues group also largely consisted of senior-level participants (73%), 11% were juniors, and 16% unidentified. Major areas of study for participants are illustrated in Table 1.

Design and Materials. Participants in both the Targeted Cues and Control groups were asked to sort six scenarios (on note cards) into two equal groups based on the arguments

in the scenarios. The categorization task measured whether students were able to identify fallacious reasoning patterns common to argumentative contexts. Materials across the Control and Targeted Cues groups differed both in the instructions used to guide participants in completing the task and in the structure of the scenarios.

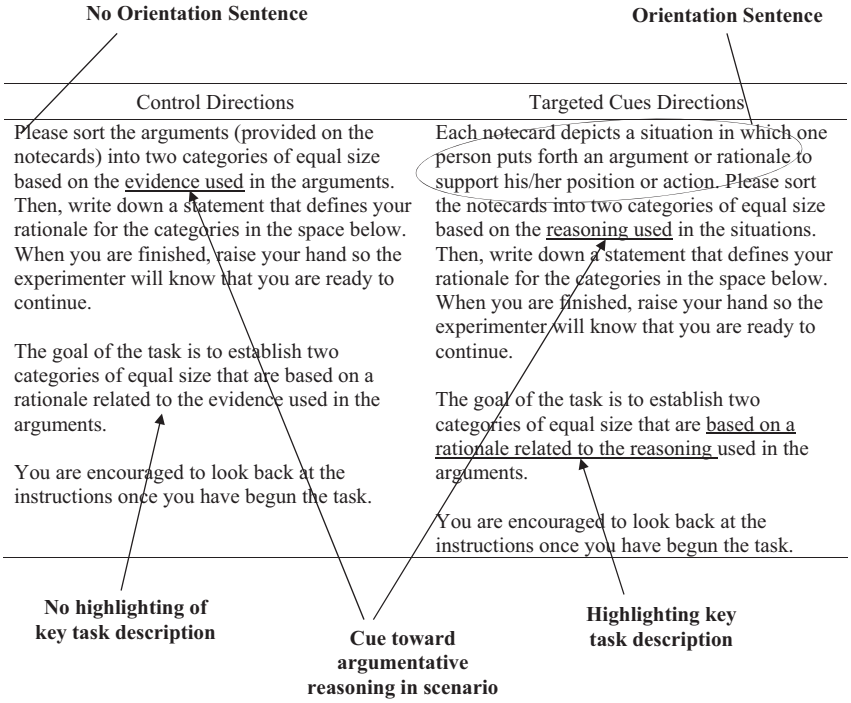

Figure 2. comparison of sort directions.

Differences in Task Instructions. The task instructions for the two conditions differed in three ways. First, the instructions differed in the language used to cue attention toward the argument within the scenario. In the Control, the task directions explicitly highlighted the "evidence used" as opposed to the "reasoning used" in the Targeted Cues group (Treatment) (see Figure 2). The shift toward more explicit language in the Targeted Cues group was motivated by a concern that the term *evidence* is open to a wider range of interpretations than is the term *reasoning*. While we view evidence as the product of reasoning about data or information (Cavagnetto & Hand, 2011), it is possible that participants might interpret evidence to mean empirical data; thus not focus on reasoning. *Reasoning* may connote a greater emphasis on logic. Therefore, there was potential that the directions in the Control group were not cueing participants' attention on the reasoning patterns in the sort card scenarios. The instructions also differed in establishing the context for the task. Specifically, the Targeted Cues instructions included an orientation sentence that more fully explicated and situated the scenarios: "Each notecard depicts a situation in which one person puts forth an argument or rationale to support his/her position or action." This was intended to orient participants to the general structure of the scenarios and make it less likely they would get caught up in surface level distinctions (i.e., person characteristics, setting, etc.) for the sort task. The Control instructions offered no orientation of the scenario to the reader.

TABLE 2
Comparison of Control and Targeted Cues Group Sort Cards

	False Cause	Sound Reasoning
Control	Jackie and Heidi are having lunch. While eating burgers and fries, they discuss the new video editing software that Heidi has recently installed onto her laptop.	Ally and Bill have just received their scores for their performance in their dance class that they attended that morning. Both of them did well on the performance assessment.
	Jackie: How do you like the new software Heidi?	Ally: See Bill, I always do well when I have a cup of coffee in the morning.
	Heidi: It works great. Also, I think the new software has caused my computer to perform much better. I downloaded the software last week and since then my computer navigates the internet much faster.	Bill: Coffee tastes good, but maybe your performance score is because the instructor only assessed the most basic steps in the routine.
Targeted Cue	Gwen is at a hip downtown bar and her friend Bill offers to buy her an apple martini. While considering Bill's offer she reflects, "the last few times I drank an apple martini I did not feel hungry for the rest of the night. Apple Martinis block hunger."	After a long, boring phone call home to his mother, Ryan went to class and did poorly on an exam. He thinks to himself, "Did that phone call ruin my ability to get a good score? I don't know, but I'll keep it in mind."

Differences in Structure of the Scenarios. The scenarios themselves were constructed slightly differently depending on condition to vary the amount of *noise* or narrative similarities available in the scenarios. The Targeted Cues scenarios featured less narrative flourish and extraneous information than the Control scenarios. The scenarios in the Targeted Cues condition depicted one individual's thinking; by contrast, the Control scenarios depicted a conversation between two people in a social setting. Extraneous information has the potential to cue frames that are not critical to examining the argumentative reasoning –for example, the conversations between two people have greater potential to evoke frames related to social interaction or dynamics. We note that these characteristics led the Control scenarios to be slightly longer than the Targeted Cues scenarios. Table 2 provides examples of Control and Targeted Cues cards. As illustrated in Table 2, the Control scenarios are less succinct and have more information to sift through to recognize key information on which to base the sort upon. Conversely, the Targeted Cues scenarios are succinct; thus a greater proportion (compared to the Control scenarios) of the information in the scenario is relevant to correctly sorting.

Two fallacious reasoning patterns were used in the experiment for both the treatment and control groups: false cause (see Tables S1 and S2 for the full card sets in the Supporting information) and hasty generalization (Tables S3 and S4 in the Supporting information). Two reasoning patterns were used to strike a balance between generalizability and clarity of focus; the two patterns were not a factor in the design (in Experiment 2, we extend the

materials to include a third reasoning pattern and also apply greater consideration to their differences). False cause (also known as post hoc) reasoning occurs when one interprets a correlated variable as a causal variable (Engel, 1995). The other reasoning pattern, hasty generalization, is commonly described in two ways. A hasty generalization can occur when a single instance is used to draw a general conclusion about an entire population or when exceptional qualifications are neglected when making a generalization (Walton, 1995). For this study, the former definition, i.e., when a single instance is used to draw a conclusion about an entire population, was employed. These patterns were chosen as both are important for the appropriate evaluation of arguments in science and science education contexts.

Readability was similar between the treatment scenarios as measured by the Flesch Reading Ease statistic: For false cause, Control and Targeted Cues were 79.4 and 75.9, respectively (indicating that the Control scenarios were slightly easier to read); for hasty generalization, Control and Targeted Cues were both 72.4. All four sets of materials scored between grades 5.1 and 6.9 on the Flesch–Kincaid Grade Level statistic suggesting that the materials were highly accessible to the undergraduate participants.

Procedure. Upon entering the lab, students were asked to complete a consent form. After the consent form was signed, the participant completed the categorization task (either Control false cause, Control hasty generalization, Targeted Cues false cause, or Targeted Cues hasty generalization). The Control groups received direction with minimal cues and scenarios with more extraneous information ($n = 51$), and the Targeted Cues groups received directions designed to highlight the reasoning patterns in the scenarios and scenarios with minimal extraneous information ($n = 37$). The categorization task required participants to sort six notecards into two equal groups based on reasoning patterns that were linked to the rationale for a claim. Three of the scenarios illustrated the same fallacious reasoning pattern (either false cause or hasty generalization), and three illustrated logical reasoning. Participants were asked to complete the categorization task and provide their rationale for grouping the scenarios. The scenario cards were numbered in a manner that allowed for easy identification and recording of the constructed categories. Participants' sorts were captured by a research assistant when participants raised their hands upon completing the categorization task. Their written rationales were then collected by the research assistant. Participants were debriefed regarding the study at the end of the experimental session.

Analysis. A 2×2 chi-square analysis was used to detect the effect of support level on sort performance. The dependent measure of sort performance was coded categorically as correct or incorrect. To be categorized as correct, the participant had to group together the three scenarios illustrating fallacious reasoning and provide a rationale illustrating that the scenarios were categorized based on the fallacious reasoning pattern. Thus, a participant received a correct rating only if she had both components. She received an incorrect rating if she either did not group the cards correctly or if she did not provide an appropriate rationale. Table 3 illustrates the scoring categories and examples of responses.

Two raters (independent of one another and blind to condition) scored the responses. Cohen's kappa (κ) was calculated to determine level of interrater reliability. The analysis revealed $\kappa = 0.67$, indicating substantial agreement (percent agreement among raters was 84%). Responses that were not unanimously scored were scored by a third rater who was blind to condition (the third rater served as arbiter). Table 4 provides examples of sort rationales that were disagreed upon and required arbitration.

TABLE 3
Scoring Rationale and Examples

Criteria for Correct	Examples
Hasty generalization	
Justification suggests one group made a conclusion based on a small sample while the other group made a conclusion based on a large sample	Correct examples • In one group, a decision is based on a series of occurrences that support the argument. In the other group the decision is based on one occurrence. • One of the categories is decisions based on informed data with a wide range of information to support it. The other category is decisions based on one instance or not enough information to generalize. Incorrect examples • One group is based more on statistics and fact while the other group is based more on opinion or chance. • One category has events that have relationship with their physical (e.g., alcohol and caffeine effect a human's body). Those events are actual. The other events relate with their mental –depending on their opinion or thought.
False cause	
Rationale indicates that in one group a causal link is indicated (even though it is based on an association) and in the other group the author is cautious –stopping short of claiming causality.	Correct examples • In one group the argument is a claim based off a onetime experience with no actual proof, the other is suggesting one thing causes the other but is aware it also may not. • Category 1: cause and effect –arguer believes change has occurred because of the discussed variable. Category 2 –arguer mentions a change but believes end result has to do with a different condition. Incorrect examples • I divided them based on social life conversations and sports or body activities conversation. • One category of cards deals with negative outcomes (food poisoning, baby rash, not doing well in the race). The other category deals with positive outcomes (computer software, dance performance, winning soccer team).

Results and Discussion

In the Control group 27% of participants were able to correctly perform the categorization task (successful sort and rationale), whereas 48% of participants succeeded in the Targeted Cues group. Participants in the Targeted Cues group performed significantly better on the categorization task than Control participants, χ^2 $(1, N = 88) = 4.2$, $p = .047$. Odds ratio analysis can be used to assess the strength of this association (effect size). Participants with Targeted Cues were 2.47 times more likely (relative to Control) to successfully recognize a fallacious reasoning pattern. Results within each fallacious pattern also illustrated better performance among the Targeted Cues group. With each of the fallacious patterns, the

TABLE 4
Examples of Disputed Responses

Response	Final Score From Blind Third Rater
Hasty generalization	Correct
The arguments with Gina and Gwen, Jane and Ron, and Tom and Bill are all decisions being made or given with personal experience being the deciding factors. The other three are more thoughtful and not completely agreeing or disagreeing but rather giving an idea while not going to the extreme.[a]	
I split the cards into two groups based on the legitimacy of the arguments. One group is all based on personal opinion while the other is based on fact.[a]	Incorrect
False cause	
One set of arguments used more in-depth and clear reasoning. The other set just refuted a previous point. Maybe you did well because of "b" not "a" argument.[b]	Incorrect
First category—Implications of prior events led to questions put forth by subjects. Second category—A direct conclusion without any doubt was made based on prior experiences.[c]	Correct

Note: [a]The first and second examples under hasty generalization are very similar and are illustrative of the nuances that were found in participant responses. The first example was ultimately scored correct by the third rater because of the final sentence suggesting "going to the extreme." The two raters who viewed this as correct interpreted that language as referring to the generalization. The second example did not have a clear reference to the overgeneralization and was therefore not rated as correct by two of the raters.
[b]The first example under false cause was scored correct by one of the initial raters because of the reference to the logic of "'B' not 'A'." The other two raters did not see this as a clear enough explanation to illustrate false cause.
[c]The second example under false cause was not clear to one rater because of the language used to explain the first category. The other two raters found explanation of the second category as a strong enough fit with the scoring rubric to rate it as correct.

Targeted Cues group recognized the pattern 20% more often than the Control group. Participants in the Targeted Cues group detected the false cause pattern 55% percent of the time (11/20) compared to 35% (8/23) of the time in the Control group. Participants in the Targeted Cues group recognized the hasty generalization pattern 41% of the time (7/17) compared to 21% (6/28) of the time in the Control group.

Analysis of the content of participant responses also supported our hypothesis in that many participants were drawn to surface-level similarities rather than the reasoning patterns illustrated in the scenarios. Participants who did not sort correctly in the false cause group commonly classified the scenarios as *negative versus positive* (25% incorrect rationales). Other characterizations observed included *things that make performance better versus excuses for poor performance*, *observation versus superstition*, *physical body versus related to the mind*, and *social life versus conversations about kinesthetic activities*. Major themes from incorrect sorts in the hasty generalization group included *personal experience versus experience or opinions of others* (30%), *general statements versus specific statements* (13%), and *positive versus negative experiences* (13%). Other rationales included *group membership versus behavior, whether the person in the scenarios likes or*

dislikes the situation, and *whether the decision would hurt the person or not*. Importantly, not all incorrect responses focused on surface-level details of the scenarios. For example, two participants in the Targeted Cues group interpreted the false cause scenarios as making a hasty generalization whereas a third participant indicated that three of the scenarios included assumptions and three did not include assumptions. A number of responses were vague (15%) and provided little information about what the participants based their decisions on.

The findings from this study support our hypothesis that students struggle to recognize opportunities to apply their knowledge related to argumentative reasoning—and that strategic use and forms of language can help to overcome this transfer failure. We choose to interpret these findings cautiously, however. One limitation is that we did not compare the impact of targeted cues on performance using the exact same items. Therefore, despite the measures of reading ease, etc., it is possible that the item sets differed or influenced performance in ways that we did not intend. It is worth noting that given the differences that were built into the item sets, there was no way to make them align perfectly. Accordingly, the items represent a focus on optimally realizing the reasoning patterns rather than attempting to maximize alignment. A second limiting factor in the interpretation of these results is that we do not know specifically which aspect of the Targeted Cues condition led to better performance—only that the combination was effective. This is consistent with the general nature of our initial claim: That strategic use and forms of language can improve knowledge access. It is a clear direction for future work to determine what specific types or combinations of targeted cues are most effective.

While the present findings show that strategic cues and reduced distractors can enhance identification of reasoning patterns, it was just over half of the participants in the Targeted Cues group who were successful. Given the broader objective of enhancing argument quality in science education contexts, a higher success rate is an important goal. One possible approach would be to try to strengthen the level of support by exaggerating our manipulation of instructions and the characteristics of the scenarios. We were not convinced there was a likely way to dramatically boost the impact in this manner. Instead, we compare whether an additional type of contextual cue works effectively in concert with the existing targeted cues. Are students more likely to transfer their knowledge about reasoning and argument when asked to evaluate explicitly scientific settings?

EXPERIMENT 2

Argumentative norms are contextually bound to settings. These settings have the potential to generate diverse antecedent stimuli resulting in activation of different cognitive processing components; ultimately resulting in outcome variation. Differential performance across context was documented by Osborne et al. (2004) who found that students developed more complex arguments in socioscientific contexts compared to strict science contexts. We hypothesize that broad context also influences students' abilities to detect argumentative reasoning patterns. When learning science, students move from everyday language and explanations of phenomena to scientific language and explanations (Mortimer & Scott, 2003). Therefore, construction and critique of arguments in science classrooms are embedded in both everyday and scientific language. The intimate relationship of the rhetorical norms of both everyday and scientific language informs how one processes information. Intuitively, everyday settings may be easier to comprehend, they may avoid a possible sense of intimidation for some students, and they may benefit from familiarity and background knowledge. The use of scientific settings, on the other hand, may explicitly or implicitly connote a sense of intellectual rigor and invite students to engage in more formal cognitive

processing or more fully draw upon their knowledge bases associated with logic and reasoning. While we suspect that science contexts will cue reasoning, we recognize that the *quality* of reasoning is contingent on context; thus, there are no fallacious reasoning patterns per se (Walton, 1998). As such. we recognize the science context may not be optimal for promoting attention to *all* reasoning patterns.

In E2, we used a fully randomized experimental design. This design has the benefit of allowing for examination of treatment and increases the control over possible validity threats. As in E1, the dependent variables in the study were sorting task outcomes and the ability to provide an explanation showing that the basis of the sort was related to the nature of the reasoning in the examples. The study explores an additional path to providing support that promotes student attention to reasoning patterns: the role of broad context.

Method

Participants. Participants were similar to the demographics found in E1 (juniors and seniors from a variety of majors) as they were recruited from the same courses as E1. Seventy-four students (72% female, 28% Male) were randomly assigned to one of two groups (Everyday Context or Science Context) with both groups completing three categorization tasks. The Everyday Context was the same group as the Targeted Cues group from E1. As mentioned in E1, the group consisted of 37 participants of which 73% were female and 27% were male. The makeup of the Science Context group was similar, consisting of 37 participants: 70% female and 30% male. The everyday group were mostly seniors (73%), 11% were juniors, and 16% unidentified. The science group were slightly less dominated by seniors (65%), 19% juniors, and again 16% unidentified. Participants were drawn from similar major areas of study across the everyday and science groups (Table 5).

TABLE 5
Study 2 Participant Majors (Reported as Relative Frequencies)

	Everyday Group	Science Group
Natural sciences	27	41
Mathematics/engineering	14	3
Management	5	11
Humanities	35	30
Social sciences	19	14
Undeclared	0	3

Procedure. Student participants were asked upon arrival to complete a consent form. After the consent form was signed, participants were randomly assigned to either the *Everyday* or *Science* group and each participant was asked to complete three sorts set in the assigned context. The participant was given the first sort task (either false cause or hasty generalization—the pattern was alternated to counterbalance across all participants). After completing the sort task and writing a rationale for the sort task, the participant raised his/her hand and a graduate student recorded the sort task results and collected the written rationale. The participant then received an additional sort task (personal attack), and this process was repeated. A third sort task was then given to the participant (either false cause or hasty generalization depending on which reasoning pattern they received during the first

TABLE 6
Scoring Description and Examples of Personal Attack

Personal attack	
Rationale indicates that in one group the person disagrees with the argument because of a characteristic of the messenger whereas in the other group the person disagrees because of a characteristic of the argument itself (e.g., the evidence is weak)	Correct examples • One category is based on logic, evidence, and reason. In one decisions are based on stereotypes, prejudgments, and profiling. • Unbiased claims that directly address the research versus biased assumptions based on the researcher (not research). Incorrect examples • Pile 1 are not based on facts and very bias. Pile 2 while also opinions are more rational because they stem from fact. • One group's argument is based off of one other person's opinion. The other group's argument is based off of their own opinion.

sort), and the sort results and rationale were collected. Once all three sorts were completed, the participant was debriefed on the study and exited the lab.

Materials. The instructions and format of the categorization scenarios followed the Targeted Cues group from E1 and were held constant across contexts. Three fallacious reasoning patterns were targeted in the experiment across both everyday and science contexts: (a) false cause, (b) hasty generalization, and (c) personal attack (Tables S5–S8 in the Supporting information). False cause and hasty generalization were chosen because they are commonly found in science contexts (refer to E1 for descriptions). Personal attack (ad hominem) is a common pattern found across a variety of disciplines and social contexts, but is often perceived as a violation of science etiquette. The personal attack, as the name suggests, is when a perceived negative aspect of a person's character is used as a basis for persuasion. Each participant completed one categorization task for each of the aforementioned reasoning patterns. The order in which students engaged in the three fallacies was counterbalanced with personal attack always being completed second. Scenarios set in everyday contexts scored a 7.7 grade level and 68.5 on the Flesch–Kincaid reading statistic, whereas those in science contexts scored a 7.1 grade level and a 67.2 (indicating the sets of cards were comparable and accessible to the undergraduate participants).

Analysis. A two-tailed *t*-test analysis was used to detect the effect of context on sort performance. The dependent measure of sort performance was coded categorically as correct or incorrect as in E1. Two raters (independent of one another and blind to condition) scored the responses using the same scoring rubric as in E1 for false cause and hasty generalization sorts (please refer back to Table 3 for criteria and example rationales). Table 6 provides the scoring rubric for personal attack and example rationales. Cohen's kappa (κ) was calculated to determine level of interrater reliability. The analysis revealed $\kappa = 0.97$, indicating strong agreement (percent agreement among raters was 97%). Responses that were not unanimously scored were scored by a third rater who was blind to condition

(the third rater served as arbiter). Scoring of the personal attack pattern was unanimous (refer to Table 4 for examples of scoring differences for false cause and hasty generalization).

Results and Discussion

Aggregating across the three fallacious reasoning patterns, participants in the everyday context were successful in the categorization tasks 50% of the time. This performance level is consistent with the results of E1. Given the multiple measures per participant, we computed a summed performance score across the three reasoning patterns and conducted a parametric statistical test (two-tailed *t*-test). The Science group ($M = 1.92$, $SD = 0.93$) performed significantly better than the Everyday group ($M = 1.51$, $SD = 0.80$), $t(72) = 2.01$, $p < .05$, $d = 0.47$. We followed this up with individual chi-square tests that revealed strong evidence for group differences on the personal attack reasoning pattern, χ^2 (1, $N = 74$) = 7.633, $p = .006$. Specifically, personal attack was detected 54% of the time in the everyday setting (20/37) and 84% of the time in the science setting (31/37). Odds ratio analysis indicated that participants were 4.4 times more likely to recognize the personal attack fallacy in the science setting than in the everyday setting. There was no evidence of a reliable difference in performance on the sort task for the false cause reasoning pattern, χ^2 (1, $N = 74$) = 0.056, $p = .814$), nor the hasty generalization reasoning pattern, χ^2 (1, $N = 74$) = 1.367, $p = .242$. To be clear, these results do not call into question the advantage of targeted cues observed in E1; they suggest variation in the impact of the science context across the different fallacious patterns.

Participant responses that were scored as incorrect in the personal attack-everyday context group included the themes *opinion versus logic* (35%), *lack of trust versus lack of evidence* (18%), and *personal judgment versus other's opinions* (18%). Only six rationales were incorrect in the personal attack-science context group. Two of those six participants suggested that they sorted by *valid versus invalid reasoning* while another sorted by *somewhat skeptical reactions to findings versus view findings as completely false.* Other rationales were less clear, i.e., *low level of research participants versus low level of personal experience*. Major incorrect themes in the false cause-science context group included *certain versus uncertain conclusions* (24%), *finished research versus not finished research* (12%), and *subject and cause are in contact versus subject and cause are near each other* (12%). The themes in the hasty generalization-science context group included *small sample size versus large sample size* (38%) and *things that were experimented on had choices versus no choices* (15%). The themes found among incorrect responses in the everyday context group for false cause and hasty generalization were reported in E1.

Our findings suggest that the influence of context is intimately tied to the dialogic norms associated with the context. That is, personal attacks are not considered appropriate in making a scientific argument, but could be considered appropriate in certain everyday settings (e.g., a quarrel). This dialectic view of argumentative reasoning has been established by Walton (1998) who contends that there are no fallacious reasoning patterns because appropriateness of reasoning is context dependent. We build on Walton's dialectic perspective of argument in suggesting that fallaciousness in context is not categorical, but rather continuous. That is, while we recognize that all three of the targeted reasoning patterns (false cause, personal attack, and hasty generalization) are fallacious in science contexts, false cause and hasty generalization are more subtle fallacies than personal attack.

Finally, the design of the experiment allowed for exploration of a possible practice effect. While practice was limited to only three sorts in a short period of time, we were interested to see whether students would perform better on the third sort than on their first sort.

This prediction was not supported as there was no detectable relationship between sort performance and practice (first sort vs. last sort), χ^2 $(1, N = 74) = 0.433$, $p = .511$.

GENERAL DISCUSSION

Promoting interest in argumentative reasoning and supporting its development gives students additional tools to engage in the science learning process. The two experiments reported in this study explored a context-dependent account of human argumentative reasoning. The results of E1 show that more strategic conceptual cues and reduction of extraneous information can enhance undergraduate students' abilities to recognize fallacious reasoning patterns. The findings from E2 were less clear. While participants were able to recognize fallacious reasoning patterns more in science contexts, the difference in performance was largely driven by the personal attack reasoning pattern. No differences in performance were detected with the false cause and hasty generalization patterns. These discrepant findings may be contingent upon the extent to which the reasoning pattern is masked by the science or everyday context (Walton, 1998). Taken together, the results of E1 and E2 support a context-dependent as opposed to knowledge-deficiency account of human argumentative reasoning. Participants did not receive any formal instruction about argumentative reasoning structures; yet they were able to detect more appropriate reasoning patterns with targeted contextual cues. This suggests that undergraduate students have cognitive resources that are appropriate and applicable to argumentative reasoning. We believe the findings highlight a key role for task framing, i.e., success in accessing existing appropriate knowledge bases. Participants in our study held some level of reasoning that was not drawn upon when scenarios contained a considerable amount of extraneous details. Identification did increase as the relative amount of details decreased. We argue that the extraneous information in the Control group scenarios resulted in a minor reframing of the task for many students. While the study is limited by the undergraduate sample and the laboratory-based context of the study, the findings lend support to recent classroom-based studies that illustrate the contextual nature of argumentative reasoning, specifically that the perception of the task can influence participants argumentative abilities (Berland, 2011; Berland & Hammer, 2012; Berland & Lee, 2012; Felton et al., 2009; Garcia-Mila et al., 2013; Gilabert et al., 2012). These recent studies suggest that teachers place considerable emphasis on providing conditions that clearly establish the purpose of argument and strategically focus attention toward reasoning patterns within argument activities. The results of this study, although limited by the sample and study context, appear to support similar practical implications.

Importantly, even though the lab-based study is not representative of actual classrooms—particularly those that engage students in verbal argumentation around science phenomena, we suggest that the sterile nature of the study's environment actually may contribute to an underestimation of the importance of contextual cues and recognition of when to enact cognitive resources related to argumentative reasoning. That is, in the current study there were far fewer distractors and information available (compared to an actual classroom) to seduce attention away from knowledge bases related to argumentative reasoning. Thus, it is reasonable to suggest that authentic classroom environments may be playing an even greater role in student abilities to enact or access existing, yet inert knowledge related to argumentative reasoning.

Despite limitations in the present study, it is worth speculating about potential implications for classroom application. Most directly, this is an alternative basis for explaining poor performance in argumentative reasoning. A student may actually possess the appropriate knowledge bases to be successful in engaging in high-quality argumentation, but requires a

context that supports timely recruitment of those resources. This suggests that assessment should be pursued with due consideration to the context provided to the student and that pedagogy should be directed, as appropriate, either toward improving the knowledge bases or toward improving access to the knowledge bases. We are intrigued by the possibility that appropriate use of context in the classroom can serve to elevate both performance and learning in the domain of argumentative reasoning – serving as a kind of shortcut toward more expert-like ability.

While this study supports the recent context-based view of argumentative reasoning, it also may inform our understanding of previous work in argumentation abilities and interventions. First, our study along with the recent context-based studies of argumentation suggests that earlier studies may be underestimating students' argumentative reasoning abilities. Importantly, we are not the first to make such an assertion (see Koslowski, 1996; Mercier & Sperber, 2011). Perhaps more importantly, a framing perspective also may help explain findings from previous studies reporting gains from explicit instruction and practice (Kuhn, 2010; McNeill et al., 2006; Osborne et al. 2004). Explicit instruction in argumentative reasoning and practice (or simply practice itself) may be increasing the likelihood that students' frame tasks as argumentative—more commonly cueing knowledge bases related to argumentative reasoning. Indeed, if a school science class is continually emphasizing evidence-based claims, counterclaims, and rebuttals, it is reasonable to suggest that students will recognize those expectations and there will be increased attention and subsequent performance of argumentative reasoning tasks (Kuhn, 2010; Kuhn, Zillmer, Crowell, & Zavala, 2013). The current study also raises a number of specific questions. For example, how do the findings illustrated in the current study translate for school-aged students in complex classroom contexts and what classroom conditions trigger attention toward the use of cognitive resources related to argumentative reasoning?

In summary, recent studies in argumentative reasoning, including the current study, appear to be contributing to a subtle but important shift in the field. The findings of E1 and E2 do not contradict existing approaches, but instead advance the claim that access to existing cognitive resources is a causal factor, and, furthermore, it is a factor that may be readily influenced to produce a positive impact.

The authors would like to thank Katy Kam and Michelle Tao for their contributions to this study and manuscript as well as Dr. Richard Lamb, Director of the WSU Neurocognitive Science Laboratory, and Dr. Olusola Adesope for their considered feedback.

REFERENCES

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychological Bulletin, 128(4), 612–637.

Berland, L. K. (2011). Explaining variation in how classroom communities adapt the practice of scientific argumentation. Journal of the Learning Sciences, 20(4), 625–664.

Berland, L. K., & Hammer, D. (2012). Framing for scientific argumentation. Journal of Research in Science Teaching, 49(1), 68–94.

Berland, L. K., & Lee, V. R. (2012). In pursuit of consensus: Disagreement and legitimization during small-group argumentation. International Journal of Science Education, 34(12), 1857–1882.

Cavagnetto, A. R. (2010). Argument to foster scientific literacy: A review of argument interventions in K–12 contexts. Review of Educational Research, 80, 336–371.doi:10.3102/0034654310176953

Cavagnetto, A. R., & Hand, B. (2011). The importance of embedding argument within science classrooms. In K. Myint (Ed.), Perspectives on scientific argumentation: Theory, practice and research (pp. 39–53). Dordrecht, The, Netherlands: Springer.

Cavagnetto, A. R., Hand, B., & Norton-Meier, L. (2010). The nature of elementary student science discourse in the context of the science writing heuristic approach. International Journal of Science Education, 32, 427–449. doi:10.1080/09500690802627277

Chi, M. T. H., Feltovitch, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121–152.

Day, S., & Goldstone, R. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. Educational Psychologist, 47, 153–176.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. Science Education, 84(3), 287–312.

Engel, S. M. (1995). With good reason: An introduction to informal fallacies (5th ed.). New York, NY: St. Martin's Press.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. Science Education, 88(6), 915–933.

Felton, M., Garcia-Mila, M., & Gilabert, S. (2009). Deliberation versus dispute: The impact of argumentative discourse goals on learning and reasoning in the science classroom. Informal Logic, 29(4), 417–446.

Galbraith, D. (1999). Writing as a knowledge constituting process. In M. Torrence & D. Galbraith (Eds.), Knowing what to write: Conceptual processes in text production (pp. 139–160). Amsterdam , The Netherlands: Amsterdam University Press.

Garcia-Mila, M., Gilabert, S., Erduran, S., & Felton, M. (2013). The effect of argumentative task goal on the quality of argumentative discourse. Science Education, 97(4), 497–523. doi:10.1002/sce.21057

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability and inferential soundness. Cognitive Psychology, 25, 524–575.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. Cognitive Psychology, 12, 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15, 1–38.

Gilabert, S., Garcia-Mila, M., & Felton, M. K. (2012). The effect of task instructions on students' use of repetition in argumentative discourse. International Journal of Science Education, 35(17), 2857–2878. doi:10.1080/09500693.2012.663191

Grooms, J., Sampson, V., & Golden, B. (2014). Comparing the effectiveness of verification and inquiry laboratories in supporting undergraduate science students in constructing arguments around socioscientific issues. International Journal of Science Education, 36(9), 1412–1433.

Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In J. P. Mestre (Ed.), Transfer of learning from a modern multidisciplinary perspective (pp. 89–119). Greenwich, CT: Information Age Publishing.

Jiménez-Aleixandre, M. P., Bugallo Rodríguez, A., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in high school genetics. Science Education, 84(6), 757–792.

Kelly, G. J., & Takao, A. (2002). Epistemic levels in argument: An analysis of university oceanography students' use of evidence in writing. Science Education, 86(3), 314–342.

Koslowski, B. (1996). Theory and evidence: The development of scientific reasoning. Cambridge, MA: MIT Press.

Kuhn, D. (1991). Education for thinking. Cambridge, MA: Harvard University Press.

Kuhn, D. (2010). Teaching and learning science as argument. Science Education, 94(5), 810–824. doi:10.1002/sce.20395

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). The development of scientific thinking skills. Orlando, FL: Academic 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.

Kurtz, K., & Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one. Memory & Cognition, 35(2), 334–341.

Lamb, R. L., Vallett, D. B., Akmal, T., & Baldwin, K. (2014). A computational modeling of student cognitive processes in science education. Computers & Education, 79, 116–125.

Loewenstein, J. (2010). How one's hook is baited matters for catching an analogy. In B. Ross (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 53, pp. 149–182). New York, NY: Academic Press.

McNeill, K., Lizotte, D., Krajcik, J., & Marx, R. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. Journal of the Learning Sciences, 15(2), 153–191.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. Behavioral and Brain Sciences, 34(11), 57–111.

Mortimer, E. F., & Scott, P. H. (2003). Meaning making in secondary science classrooms. Philadelphia, PA: Open University Press.

Nussbaum, E. M., Sinatra, G., & Poliquin, A. (2008). Role of epistemic beliefs and scientific argumentation in science learning. International Journal of Science Education, 30(15), 1977–1999.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. Journal of Research in Science Teaching, 41(10), 994–1020.

Ross, B. H. (1984). Reminding and their effects in learning a cognitive skill. Cognitive Psychology, 16, 371–416.

Rottman, B., Gentner, D., & Goldwater, M. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. Cognitive Science, 36, 919–932.

Ryu, S., & Sandoval, W. A. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. Science Education, 96(3), 488–526.

Sadler, T. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. Journal of Research in Science Teaching, 41(5), 513–536.

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naive similarity to ecological knowledge. Journal of Experimental Psychology: Learning, Memory, & Cognition, 29, 641–649.

Smith, J. I., Combs, E. D., Nagami, P. H., Alto, V. M., Goh, H. G., Gourdet, M. A. A., ... Tanner, K. D. (2013). Development of the biology card sorting task to measure conceptual expertise in biology. CBE–Life Sciences Education, 12(4), 628–644. doi:10.1187/cbe.13-05-0096

Spencer, R. M., & Weisberg, R. W. (1986). Context-dependent effects on analogical transfer. Memory & Cognition, 14, 442–449.

Stains, M., & Talanquer, V. (2008). Classification of chemical reactions: Stages of expertise. Journal of Research in Science Training, 45, 771–793.

Walton, D. (1995). A pragmatic theory of fallacy. Tuscaloosa, AL: University of Alabama Press.

Walton, D. (1998). The new dialectic: Conversational contexts of argument. Toronto, Canada: University of Toronto Press.

Whitehead, A. N. (1929). The aims of education and other essays. New York, NY: Free Press.