Published in final edited form as:

J Exp Educ. 2009 July 1; 77(4): 339.

Improving Students' Evaluation of Informal Arguments

AARON A. LARSON,
Northern Illinois University

M. ANNE BRITT, and
Northern Illinois University

CHRISTOPHER A. KURBY
University of Memphis

Abstract

Evaluating the structural quality of arguments is a skill important to students' ability to comprehend the arguments of others and produce their own. The authors examined college and high school students' ability to evaluate the quality of 2-clause (claim-reason) arguments and tested a tutorial to improve this ability. These experiments indicated that college and high school students had difficulty evaluating arguments on the basis of their quality. Experiments 1 and 2 showed that a tutorial explaining skills important to overall argument evaluation increased performance but that immediate feedback during training was necessary for teaching students to evaluate the claim-reason connection. Using a Web-based version of the tutorial, Experiment 3 extended this finding to the performance of high-school students. The study suggests that teaching the structure of an argument and teaching students to pay attention to the precise message of the claim can improve argument evaluation.

Keywords

argumentation; argument evaluation; argument structure; claim predicates; immediate feedback

The ability to evaluate arguments is a fundamental aspect of social interaction and human decision making. Argumentation skills are also important in academic endeavors. Students in the humanities are expected to analyze arguments presented in texts and present a reasoned analysis of that information (National Center for History in the Schools, 1996). In the sciences, argumentation skills can help students acquire knowledge (Driver, Newton, & Osborne, 2000; Schwarz, Neuman, Gil, & Iiya, 2003; Zohar & Nemet, 2002), and it is central to hypothesis testing; that is, students need to be able to determine whether empirical evidence provides support for a stated claim (e.g. Duschl, 2000; Kuhn, 1993; Lawson, 2003; Lederman, 1992). Argumentation skills are also required for professional training as indicated by its inclusion on most entrance exams (e.g., SAT, GRE, LSAT, and MCAT).

Although current standards assume that students become skilled in argumentation during their schooling (American Association for the Advancement of Science, 1993; National Center for History in the Schools, 1996; National Science Education Standards, 1996; No Child Left Behind Act, 2001), many students leave high school unable to understand, evaluate, or write arguments (National Assessment of Educational Progress, 1996, National Assessment of Educational Progress, 1998, National Science Board, 2006). Students also seem to have special

difficulty with scientific arguments (Duschl, Ellenbogen, & Erduran, 1999; Duschl & Osborne, 2002; Forman, Larreamendy-Joerns, Stein, & Brown, 1998; Jimenez-Aleixandre, Rodrigues, & Duschl, 2000; Kelly, Druker, & Chen, 1998; Kuhn, 1993; Osborne, Erduran, & Simon, 2004; Takao & Kelly, 2003). Perkins, Farady, and Bushey (1991) and Kuhn (1991) have concluded that these failings may be a result of a lack of exposure to arguments and explicit instruction and practice in the skills of argumentation.

Although there are several components of argumentation skills, perhaps the most fundamental component is the ability to distinguish warranted from unwarranted arguments. A student who cannot distinguish between a claim supported by a valid reason and one supported by a disassociated reason, or not supported at all, will be able neither to critically analyze arguments nor to produce them with much proficiency. Consider arguments 1a through 1c:

- 1a. People should be allowed to have only two biological children.
- 1b. People should be allowed to have only two biological children because children are small.
- 1c. People should be allowed to have only two biological children because it would help stabilize population growth.

Argument 1a is an *unsupported assertion*. It does not fit the definition of a *minimal argument* because it does not have a reason (Toulmin, 1958; Voss & Means, 1991). Argument 1b also does not have the structure of a minimal argument. Although there is a reason provided (i.e., because children are small), the reason does not support the claim. Logically speaking, an argument with an unrelated reason has the same structure as does one that provides no reason. Items such as 1b are referred to as *unwarranted arguments* because the reason does not warrant the claim. The final argument, 1c, is a structurally acceptable argument because there is a reason (i.e., because it would help stabilize population growth), and the reason supports the claim. The ability to assess whether a statement meets these minimal requirements of an argument is a necessary step in comprehending, evaluating, and producing arguments. It is also a skill that researchers expect schools to impart to students. As Siegel (1995) argues, "When we say that rationality is a fundamental educational ideal, we are claiming that education ought to strive to foster in students the skills and abilities which will enable them competently to assess reasons" (p. 161).

Through the present series of experiments, we examined high school and college students' ability to distinguish well-structured arguments (e.g., 1c) from poorly structured arguments, when there was no reason provided (e.g., 1a) or when the reason did not support the claim (e.g., 1b). These experiments provided a baseline level of performance for high school and college students and tested the effectiveness of a tutorial that we constructed to improve their application of this skill.

Simple informal arguments

A minimal argument must have a claim that is supported by at least one reason (Toulmin, 1958; Voss & Means, 1991). The *claim* is a statement of belief that is presumed to be open to debate, thus requiring support. For example, the claim in Argument 2 is "handguns should be banned."

Argument 2: Handguns encourage criminal behavior; therefore, handguns should be banned.

The *reason* in Argument 2 is "handguns encourage criminal behavior." For the argument to be well structured, the reason must support, or lend credence to, the claim through an implicit or explicit inference from another, already accepted belief. One such inference is the *warrant*.

Warrants are unstated assumptions that arguer assumes are already acceptable to the audience (Toulmin, 1958). For example, the unstated warrant for Argument 2 is something like “the government should prohibit things that encourage criminal behavior.”

The warrant—Philosophers since Aristotle have proposed that informal arguments (e.g., Argument 2) require the reader to supply unstated information to form a complete argument. According to Aristotle, “The enthymeme must consist of a few propositions, fewer often than those which make up the normal syllogism. For if any of these propositions is a familiar fact, there is no need even to mention it; the hearer adds it himself” (1973, p 735). This corroborates Grice’s maxim of Quantity: Speakers should state what is required, but no more, to make one’s point understood (Grice, 1975). The most common unstated element in these informal arguments is the warrant. It is “a general, hypothetical statement, which can act as a bridge, and authorize the sort of step to which a particular argument commits us” (Toulmin, 1958, p 98).

Although there is some disagreement over the exact form of argument warrants (Mans & Preyer, 1999; Grennan, 1997; Levi, 1995), it is clear that the reason must support the claim if it is to form a well-structured argument. For example, the claim “handguns should be banned” can be supported by the reason in 3a but not by the reason in 3b:

3a. Handguns encourage criminal behavior, so handguns should be banned.

3b. Ninety percent of handgun purchases are now subject to instant FBI criminal background checks, so handguns should be banned.

To represent each of these arguments, readers must form a claim-reason connection and test whether that connection is warranted (i.e., the reason provides relevant support for the claim). Skilled readers of arguments need to be able to appreciate the differences among arguments such as 1a–1c and 3a–3b in terms of strength and acceptability.

Blair & Johnson (1987) have proposed a set of criteria that are required of argument premises. These include *acceptability* (the truth status of the premises), *relevance* (whether the reason warrants the conclusion), and *sufficiency* (whether the reason provides enough evidence considering everything known). Although students’ skill at considering each of these elements is important, the present research focuses on relevance.

Role of claim predicates in determining relevance

The claim of any argument can be decomposed into various components, the most important of which are the *theme* and *predicate*. The *theme* is the topic of the argument, such as “cell phone use,” “marijuana use,” or “car insurance rates.” In claims 4a–4c, the theme is “government regulation of tobacco companies.” The *predicate* refers to the main verb or predicate adjective in the claim. Claims 4a–4c differ with respect to the predicate asserted: “sued,” “was right to sue,” and “should sue.”

4a. The government recently sued the tobacco companies.

4b. The government was right to sue the tobacco companies.

4c. The government should sue the tobacco companies.

Differences among predicates can drastically alter the intent of the argument with a change of only a few words, as demonstrated in examples 4a–4c. *Factual* or *causal predicates*, such as 4a, assert that a state of affairs existed, exists, will exist, or caused something else to exist (e.g., “sued,” “will sue,” “sued because”). These types of claims are common in science and are intended, primarily, to change a belief. *Value predicates*, such as 4b, assert that some action or state is valuable, desirable, or moral (e.g., “was right to sue,” “is unfair to sue,” “needed to

sue”). The goal of value claims is to change a belief or attitude. *Policy predicates*, such as 4c, assert the need or favorability of a behavior (e.g., “should sue,” “should not be allowed to sue,” “should encourage suing”). They are frequently used to argue for a particular course of action or solution to a problem.

Precisely representing the predicate and noticing the differences among types of predicates is important because different predicates license different forms of evidence or reasons. For example, if claim 4a were in dispute, one would require court documents or some other evidence of the lawsuit to form a warranted argument. In contrast, claim 4b would require reasons that invoke specific moral standards or shared values, such as protecting the citizenry. Last, claim 4c would require a statement of an expected benefit or mitigation of an existing harm, such as recouping the cost of treating lung cancer patients.

Without careful attention to the precise predicate of the claim, readers will not be able to correctly evaluate the structural quality of arguments. In other words, they will fail to accept warranted arguments, such as 5a, or will incorrectly accept unwarranted arguments, such as 5b:

5a. The death penalty is immoral because DNA has been used to prove that many innocent people have been sentenced to death.

5b. The death penalty is ineffective because DNA has been used to prove that many sentenced to death were innocent.

Researchers have recently found that undergraduates have surprising difficulty recalling the precise wording of claim predicates, but those who can recall precisely are better able to judge whether the argument is warranted (Britt, Kurby, Dandotkar & Wolfe, 2008){Now Published, Changed to Reflect}. In three experiments, undergraduates read several simple claim-reason arguments such as 6a and 6b. After reading each argument, they rated their agreement with the argument and then it was removed from the screen and they typed in their recall of the claim of the argument.

6a. *Policy argument*. Recycling should be federally mandated because it helps protect the environment.

6b. *Value argument*. Recycling is very beneficial because it helps protect the environment.

Britt,(2008) found that undergraduates were not very accurate at recalling the predicate of the claim ($M = 76\%$, $SD = 21.6$), but were much more accurate at recalling the theme ($M = 95\%$, $SD = 7.6$). Twenty-five percent of the time, participants could not precisely state the predicate of the claim to which they had just agreed or disagreed. Instead, students seemed to rely on a gist, rather than verbatim, representation of the claim predicate (Brainerd & Reyna, 1998, 1992; Kintsch, 1988; Kintsch, 1998; Kintsch & van Dijk, 1978). It is understandable, then, that students would have difficulty judging whether a reason supports a claim if they were relying on gist representations when making such judgments. For example, a reader would need a verbatim representation of the predicate to notice that the claims in 6a and 6b are warranted from the reason “because it helps to protect the environment,” but the claim in 6c is not. One who is not attending to the precise claim predicate would have difficulty judging the quality of these three arguments.

6c. Recycling is cost-effective because it helps protect the environment.

Training evaluation of argument quality

Just as bridging inferences (McNamara, 2004) and deductive reasoning (Lehman & Nisbett, 1990) can be taught, one may expect that students can be taught to evaluate arguments on the basis of the quality or presence of a reason. Presently, there has been little research concerning

the promotion of argument evaluation. Klein, Olson, and Stanovich (1997) examined elementary school students' evaluation of deductive arguments. They found that their classroom intervention successfully improved students' ability to distinguish valid and invalid arguments. In the present studies, we attempted to improve students' evaluation of informal, non-deductive arguments. In this attempt, we designed and based our tutorial on four principles from cognitive and educational theories: active learning, motivation, task decomposition, and feedback.

Active learning and self-generation—Research has shown that students remember information when they actively complete or construct their own information (Marsh, Edelman, & Bower, 2001; Slamecka & Graf, 1978; DeWinestanelly & Bjork, 2004). This self-generation effect was extended to requiring students to generate self-explanations while learning, and this learning strategy has been shown to help students acquire deep, conceptual understanding (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & LaVancher, 1994). The constructivist approach has proposed that learners who are more engaged and who process the information more actively will learn the information better (Bransford, Brown, & Cocking, 1999; Bruer, 1993; Mayer, 2003). Active learning may be particularly useful for students who are having difficulty (Watson, Kessler, Kalla, Kam, & Ueki, 1996).

The tutorials we created use self-generation and active learning by presenting an interactive text. There are several examples in the text, and rather than just presenting them as worked examples, students are first challenged to fill in their own answers. Then they are given the correct answer along with a detailed explanation. Furthermore, students are told that they are not simply reading a text but that they will have to apply this knowledge later. Application of the skill during the learning process should support future usefulness of this skill.

Motivation and illusions of competence—Motivation and attention to instructional materials is a critical element to the success of any educational program (Deci & Ryan, 1985; Deci, 1971). If students are not motivated, they will not actively engage in learning activities. In addition to the regular sense of motivation, there is also the problem that students will not attend to the tutorial because they already believe they know what the tutorial is teaching. Bjork refers to this as an illusion of competence (Bjork, 1999; Koriati & Bjork, 2006). Indeed, in previous research, we found only a .10 correlation between performance and one's judgment of one's own argument skill. As with other tasks, participants may not realize that training will help them if they think they are already skilled or if they confuse an understanding of the words in the tutorial with an ability to use the skill the tutorial teaches.

To help students read the tutorial in a more trainable state, we began by presenting several simple arguments and asking participants to commit to a forced-choice flawed or not-flawed judgment. On the second page, the tutorial explained that all of the preceding arguments had a flaw and if they had thought that the arguments were acceptable, "They fell for common tricks that people use to persuade others." This technique was designed after a method found effective for helping dispel students' illusions of invulnerability to illegitimate persuasive techniques (Sagarin, Cialdini, Rice, & Serna, 2002). Because the majority of our participants answered incorrectly on at least one of these items, this activity gave users the opportunity to experience the need for training.

Task decomposition—The current approach to determining what should be taught is based on an information processing model that assumes that complex cognitive skills can be understood by decomposition into subprocesses (Anderson, 1993; Ericsson and Simon, 1993; Newell, 1991). This approach begins with an expert model of performance on the skills to be taught (Chi, Glaser, & Rees, 1982; Lesgold, Feltovich, Glaser, & Wang, 1981). A detailed

task analysis is then conducted to identify the component processes, and an optimal sequence of these cognitive processes is modeled (Jonassen, Tessmer, & Hannum, 1999).

We identified several information processing skills that an expert must complete during the argument evaluation. One such skill is remembering the exact wording of the claim. In determining whether a reason supports the claim, the reader must precisely represent the theme and predicate of the claim. Britt et al. (2008) found that many readers had difficulty recalling the claim predicate after making an agreement judgment. They also found that those participants who were more skilled at recalling the precise predicate were more skilled at judging whether an argument was warranted or not. Thus, attending to the predicate of the claim may be an important initial step in making a quality evaluation judgment. Immediately after reading a possible reason, the reader must reactivate or maintain his or her representation of the claim predicate so that it is available when connecting the reason to form a warranted argument. In a follow-up study, Kurby, Britt, & Dandotkar (2006) used a probe recognition paradigm to test activation of the claim predicate immediately after reading the claim and reading the supporting reason. Kurby et al. found that skilled reasoners (as measured by the logical reasoning test of the Law School Admissions Test) maintained an active representation of the predicate even after reading the reason. Less skilled reasoners were less likely to correctly detect the claim predicate after reading the reason than they were after reading the claim. This suggests the claim predicate may not be active when it is needed to make the relevance judgment. This will make it impossible for the reader to make an accurate evaluation of whether the reason supports the claim. The tutorial we constructed helps students learn to attend to claim predicates and to maintain this information even after reading intervening text. It also teaches students to search for the presence of a reason and test whether the reason does, in fact, support the claim.

Feedback—Feedback has also been shown to motivate (e.g., Deci, 1971; Harackiewicz, 1979) and aid learning (Corbett & Anderson, 2001; Kulhavy & Stock, 1989; Kulik, & Kulik, 1988). Kulhavy & Stock found that students can benefit from feedback that provides a verification of the accuracy of their answer and provides corrective information that explains the answer. A meta-analysis of the effects of feedback revealed that immediate feedback led to better learning gains for content in the classroom (Kulik, & Kulik). Immediate feedback has even been shown to be beneficial in situations where testing has led to students to make an error (Pashler, Cepeda, Wixted, & Rohrer, 2005). Rather than strengthening the association with the error, the benefits of feedback were greatest when an error was made during initial learning. Last, Corbett & Anderson found that immediate feedback led to more efficient learning in a tutorial environment.

The tutorial we created provides several examples throughout the main instruction and many more examples during a practice section. For all examples, students received the correct answer and an explanation for why it was correct. The tutorial for Experiment 1 was presented in paper booklets so that immediate and individually tailored feedback could not be given. Feedback for many items was presented in blocks. In Experiments 2 and 3, the tutorial was presented on the computer so that individualized immediate feedback could be provided.

Overview of the experiments

In a series of three experiments, we assessed students' proficiency at evaluating the quality of arguments, and we tested the effectiveness of tutorials designed to help improve this skill. In Experiment 1, we tested the effectiveness of a paper and pencil tutorial. In Experiment 2 we created a computer version and tested the need for immediate feedback. Then we created a Web-based tutorial and tested its effectiveness for teaching high school students (Experiment 3). The training and practice items were intended to be varied, with a range of topics and

perspectives in order to assess students' general skills. Clearly, personal beliefs make a difference in students' ability to evaluate arguments (Stanovich & West, 1997), but critical thinking is the ability to apply skills of argumentation independent of personal beliefs (Baron, 1991, Voss, Perkins, & Segal, 1991). Therefore, we endeavored to assess students' skill at argument evaluation across beliefs. Likewise, it is important to be able to apply argumentation skill to topics students are interested in as well as those they are less interested in. A student who can only argue about topics in which he or she is interested would be as functional as a student who could only write essays on topics in which he or she is interested. Therefore, we selected topics that students would have knowledge about, but for which they would differ on agreement and interest. One may expect students to perform slightly better when they are interested in a topic, but we wanted to measure performance across these dimensions rather than examine an ideal case.

We also acknowledge that argumentation often occurs in the context of an interactive dialogue with one or more individuals (Chinn & Anderson, 1998; Kuhn & Udell, 2003; Reznitskaya, Anderson, McNurley, Nguyen-Jahiel, Archodidou, & Kim, 2001; Veerman, Andriessen, & Kanselaar, 2002). It also occurs in more authentic tasks such as writing a research or term paper. However, the goal of this research was to assess ability in one important component skill: argument evaluation. By *argument evaluation*, we mean detecting minimally acceptable arguments from those that are not acceptable (unsupported assertions and unwarranted arguments). This skill is at the heart of any form of argumentation. Skilled quality evaluation is necessary for guiding one's comprehension and production of arguments. Once the component skill is better understood and a tutorial is available for helping improve this skill, it will then be possible to extend to more naturalistic materials and more complex situations (e.g., debate or discussion in science classrooms).

Experiment 1

The goal of Experiment 1 was to provide a baseline measure of undergraduates' ability to evaluate argument quality (i.e., distinguish unwarranted and unsupported claims from acceptable ones) and to test the effectiveness of a tutorial to improve this skill by training students to represent claims more precisely and to evaluate the reasons provided. In this experiment, half of the participants were provided instructions followed by several practice items. Then all participants were given a paper-and-pencil, flawed judgment test.

Method

Participants—Fifty-seven native English-speaking students from an introductory level psychology class at Northern Illinois University participated for partial course credit. Most of the participants were female (67%) and most were freshman (freshman 56%, sophomore 32%, junior 11%, senior 2%).

Flawed judgment test—We constructed a forced-choice flawed argument judgment test to detect whether participants were sensitive to argument quality. We manipulated 24 stem items in terms of their Quality type: acceptable, unwarranted, unsupported. The sample item presented below illustrates the *acceptable* version (7a), which has a reason that does provides support for the claim; the *unwarranted* version (7b), which has a reason that does not provide support for the claim; and the *unsupported* version (7c), which is a claim with no supporting reason.

7a. *Acceptable*. Banks shouldn't charge ATM fees because the fees make many customers unhappy.

7b. *Unwarranted*. Banks shouldn't charge ATM fees because banks are financial institutions.

7c. *Unsupported*. Banks shouldn't charge ATM fees.

For each argument, participants were asked to mark the argument that was logically or structurally flawed as *flawed* and the argument that was logically or structurally acceptable as *OK*. Participants were instructed to make this judgment only on the basis of the structure of the argument and not to consider whether they agreed with it or not or how convincing it was, given everything they knew. The correct answer for the unsupported and the unwarranted versions was *flawed*, whereas the correct answer for the acceptable versions was *OK*. Participants received an equal number of items from each quality type condition and received only one thematic version of each item. To balance the number of *flawed* and *OK* responses, an additional forty control items were created, twenty-four of which were acceptable (*OK* response) and sixteen were flawed. There was only a single version of the filler items (i.e., acceptable) and they were not included in the analysis. Reliability for the flawed judgment task was verified using Cronbach's alpha for each of the individual sections. The resultant Cronbach's alphas were .80 for the unsupported items, .73 for the unwarranted items, .64 for the changed theme items, and .59 for the good items.

To ensure these versions of the items were, in fact, of each type, 11 upper level psychology graduate students who had completed at least 2 years of graduate training in psychology were given the same test. As shown in Table 1, the graduate students (Experts), who are expert at reading research articles that contain scientific arguments, were quite accurate at correctly rejecting unsupported and unwarranted arguments and correctly accepting the acceptable arguments. Thus, skilled readers of arguments did judge the flawed arguments as flawed and the acceptable arguments as acceptable. The graduate student data also establishes that it is possible to perform accurately on the items in this test.

Tutorial—We created an eight-page, paper-and-pencil tutorial that addressed four important points: *what* is the skill, *how* to apply it, *why* it should be applied, and common *problems* in applying it. (See the appendix for the Web-based tutorial used in Experiment 3 or go to <http://think.wcp.muohio.edu:9006/mg/index.html> and type “mg1” in the login box.) The tutorial began by addressing why the tutorial was necessary for students to learn. To help motivate students to read the tutorial, page one presented a set of arguments and asked participants to state whether any of the arguments were structurally acceptable. Then, after turning to the next page, the tutorial explained that each of the preceding arguments had a flaw and this tutorial would help them better judge arguments so they would not be manipulated as easily in the future. Next, the tutorial explained what the skill was and how to do it. This section began by defining the key elements of an argument: claim, reason, and warrant. Then it provided predicate precision training by explaining the importance of attending to the predicate when evaluating whether an argument is warranted (i.e., determining whether the reason supports the claim). Many examples were provided to illustrate the process of distinguishing flawed from acceptable arguments. Next, the tutorial warned about common problems of applying this skill. The tutorial explained that, the most common problem was focusing on the argument's consistency with one's personal beliefs and whether it was the best argument given the totality of evidence possible. Students were instructed that a quality judgment does not consider those factors and that it is important to learn to separate these judgments.

After the instructional portion of the tutorial, students completed the practice portion of the tutorial. They were given 25 practice items in which they were asked to underline the predicate of the claim and to decide whether the item was an unwarranted argument, unsupported opinion, or acceptable argument. On the final page of the practice section, students were provided with an answer and a detailed explanation for each answer. They were then asked to self-check their work.

Procedure and Design—Participants were randomly assigned to Instructional condition, with half receiving no tutorial (Baseline condition) and the other half receiving the tutorial (Tutorial condition). Participants who were assigned to the tutorial condition were first given the tutorial, which took about 15 minutes to complete. Then all participants completed the forced-choice flawed judgment test. Participants in the control condition were given the tutorial after the judgment test to equate time spent in the experiment and to allow all participants to learn the material.

Results and Discussion

The average proportion values for accuracy on the flawed judgments task are presented in Table 1. The expert graduate student proportions are included only as a measure of ceiling performance and were not included in the analysis. A 3 Quality type (Unsupported, Unwarranted, Acceptable) $\times$ 2 Instructional condition (Control vs. Tutorial) mixed analysis of variance (ANOVA) was conducted with Quality type as a within subjects factor. The effect of Instructional condition was not significant, but there was a trend, $F(1, 57) = 3.134$, $MSE = 0.067$, $p = 0.08$. There was a significant Quality type $\times$ Instructional condition interaction, $F(2, 114) = 4.313$, $MSE = .049$, $p < .05$, $\eta^2 = .07$. Simple effect t tests found that there was no difference between the tutorial and baseline conditions on the Acceptable arguments, $t(57) = .748$, $p > 0.05$, or Unwarranted arguments, $t(57) = .524$, $p > 0.05$. There was, however, a difference for the Unsupported arguments, $t(57) = 3.12$, $p < 0.01$.

These results show that untrained undergraduates are not skilled at detecting flawed arguments. While the 66% baseline accuracy is significantly better than chance performance (50%), it is far below skilled performance (80%). These results also show that a short, 15-minute tutorial can help students improve their rejection of unsupported opinions as acceptable arguments. They were able to improve from 61% to 81% on checking for a reason before deciding an item was a complete argument. Training was not successful, however, in improving participants' rejection of unwarranted arguments. An inability to distinguish whether a reason provides support for a claim created a serious problem with argument production, comprehension, and evaluation. An important finding is that students have difficulty with this skill and that it is difficult to improve.

In an attempt to improve the tutorial, we examined whether students followed the instructions to self-check their answers against the feedback sheet. Even though feedback was given for the practice items, it was presented on the following page and required self-checking. A detailed review of the tutorials showed that there was a compliance problem in both the completion of the practice items and the self-checking of feedback. In fact, about half of the participants failed to check and correct their errors. In addition, informal interviews found three explanations for this lack of compliance. First, some participants said that they did not check or correct their errors because they thought the items were easy and they had confidence in their answers. However, it was generally not true that uncorrected items were correct. Second, some participants claimed that they did check a couple of items but that they had difficulty remembering why they had responded the way they did; therefore, they quit because they found it was not helpful. Third, some participants said that they did not think checking their answers was worth the effort. These results made it clear that the lack of immediate feedback was a possible reason for the lack of improvement through the tutorial.

Experiment 2

The goal of Experiment 2 was to examine whether immediate feedback during practice trials would improve undergraduates' ability to evaluate argument quality (unwarranted and unsupported vs. acceptable claims). The tutorial was modified for computer presentation so that immediate feedback could be provided. Thus, Experiment 2 tested whether immediate

feedback was helpful during practice trials beyond the effectiveness of the interactive tutorial. There were three conditions: tutorial without feedback, tutorial with immediate feedback, and no tutorial baseline. We expected that for the unsupported claim items, there would be no difference between the tutorial without and the tutorial with immediate feedback, which should have both been better than the baseline condition. For the unwarranted items, it was expected that the immediate feedback would produce the largest learning gains, and the tutorial without feedback would be equally accurate as the baseline condition. If immediate feedback was not required for learning this skill, then both tutorial conditions would perform better than would the baseline condition that did not receive the tutorial.

Method

Participants—Eighty-three English-speaking undergraduates from an introductory level psychology class at Northern Illinois University participated for partial course credit. Thirteen of the participants (or 15.7%) failed to fill in their gender and year in school. Of those students who did fill these in, most were male (57%) and most were freshman (86%, sophomore 7%, junior 6%, senior 1%).

Materials—We manipulated forty stem items in terms of their Quality type: acceptable (7a), unwarranted (7b), unsupported (7c), and changed theme (7d). The new category, *changed theme* (see 7d), was added to provide another test of the extent to which undergraduates monitor or coordinate information across claim-reason clauses. Whereas the unwarranted items require one to detect the connection of the reason as support for the claim, the changed theme items only require one to detect whether the theme or topic is related. It is worth noting that in item 7d, the theme of the claim is *banks*, whereas the theme of the reason is *the government*. These items were constructed so that if the theme was consistent, the reason would have warranted the claim.

7d. *Changed theme*. Banks shouldn't charge ATM fees because the government already takes too much money from its citizens.

Tutorial—The tutorial from Experiment 1 was modified for presentation on the computer so that immediate feedback could be provided. The tutorial was presented using E-Prime (Psychology Software Tools, 2002). After the tutorial, students were given 31 practice items in which they were asked to decide whether the item was structurally flawed or acceptable. There were two practice conditions following the tutorial. In the Feedback condition, participants were given feedback immediately after answering. This feedback included an explanation of why the answer was correct. In the No-feedback condition, participants were not given feedback following their responses.

Procedure and Design—Participants were randomly assigned to one of three Instructional conditions: baseline (no tutorial), tutorial with feedback, and tutorial with no-feedback. Participants assigned to the tutorial conditions were first given the tutorial and then received practice items either with or without feedback. The tutorial took about 15 minutes to complete. Next, all participants completed the flawed judgment test. Participants in the Baseline condition completed the flawed judgment test first and then were given the tutorial to equate time spent in the experiment and to allow all participants to learn the material.

Results and Discussion

The average proportion accuracy on the flawed judgment test are presented in Table 2. A 4 Quality Type (unsupported, unwarranted, changed theme, and acceptable) $\times$ 3 Instructional Condition (baseline, no-feedback tutorial, and feedback tutorial), mixed ANOVA was conducted with Quality Type as a within-subjects factor. There was a significant main effect

of Quality Type, $F(3, 240) = 6.138$, $MSE = .0394$, $p < .01$, $\eta^2 = .071$. Unsupported items were answered most accurately ($M = .76$, $SD = .25$), followed by acceptable ($M = .68$, $SD = .22$), unwarranted ($M = .65$, $SD = .25$), and changed theme ($M = .64$, $SD = .22$). There was also a main effect of Instructional Condition, $F(2, 80) = 7.332$, $MSE = .089$, $p < .01$, $\eta^2 = .155$. Overall, the group that was given the feedback tutorial made the most accurate evaluations (.76), followed by the no-feedback tutorial (.68), and, last, the group not given training (.61). These main effects were qualified by a significant Quality Type $\times$ Instructional Condition interaction, $F(6, 240) = 2.667$, $MSE = .0394$, $p < .05$, $\eta^2 = .062$. Simple effects tests of Instructional Condition for each Quality Type found that there was no difference on the flawed judgment test for the three conditions on the acceptable arguments, $F < 1$. There was, however, a difference for each of the other Quality Types: unsupported items, $F(2, 80) = 9.375$, $MSE = .0535$, $p < .01$, $\eta^2 = .190$; unwarranted arguments, $F(2, 80) = 4.005$, $MSE = .0603$, $p < .05$, $\eta^2 = .091$; changed theme arguments, $F(2, 80) = 5.104$, $MSE = .0423$, $p < .01$, $\eta^2 = .113$. Bonferroni post-hoc tests were conducted for the three condition means in each of the simple effects test. The post-hoc tests revealed that both tutorial conditions improved participants' accuracy in judging unsupported items as flawed. For unwarranted arguments and changed theme arguments, however, only the feedback tutorial significantly improved accuracy over the no-training baseline condition.

These results, like those from Experiment 1, show that untrained undergraduates are not highly skilled at detecting flawed arguments. Baseline performance for Experiment 2 (61%) was similar to that found in Experiment 1 (66%). The results further show that students can easily be taught to reject unsupported assertions (claims without reasons) as acceptable arguments. Students were able to improve from 61% to 81% and 86% on checking for a reason before deciding an item was a complete argument. For these items, the presence of the tutorial and practice was important, but feedback was not required.

We then conducted a more stringent test. Because the unsupported items were relatively easy to learn (as shown by Experiment 1), this more stringent analysis focused only on the items that required an evaluation of the connection between the claim and the reason (unwarranted, changed theme, and acceptable items). This was considered a more stringent analysis because of the more complex skill it required of participants. For this analysis, the dependent variable was accuracy averaged across acceptable, unwarranted, and changed theme arguments. A simple one-factor ANOVA was carried out. There was a significant effect of Instructional condition, $F(2, 80) = 4.793$, $MSE = .023$, $p < .05$, $\eta^2 = .107$. Bonferroni post-hoc tests revealed that participants in the tutorial with feedback condition ($M = .73$, $SD = .14$) were more accurate than those in the Baseline condition ($M = .61$, $SD = .11$). Participants in the Tutorial with No-feedback condition did not significantly improve in accuracy ($M = .64$, $SD = .19$) over those in the baseline condition. Thus, the tutorial and practice items were not effective unless students were given feedback during training. Experiment 1 showed that the tutorial with practice was not effective if students failed to look at the feedback. Experiment 2 replicated this finding by not providing feedback.

Experiments 1 and 2 show that the difficulty in teaching students to evaluate arguments lies in teaching them to test the connection between the reason and the claim, either in terms of support for the specific predicate or for the specific theme. Teaching this skill requires immediate feedback during training. The tutorial with immediate feedback improved participants' skill in evaluating the claim-reason link, whereas the no-feedback condition did not significantly improve students' skill. The results also suggest that this tutorial would not be effective as a paper-and-pencil class activity. Some form of computer-based presentation may be necessary to provide immediate feedback and ensure compliance.

These results point specifically to the need for practice with immediate feedback for teaching students to distinguish warranted and unwarranted arguments. It provides support for Kuhn's (1991) claim that a lack of exposure and practice accounts for a significant aspect of their poor argumentation skill. One interesting finding was that practice with feedback was not required to improve the participants' rejection of the unsupported claims as arguments. This is a relatively easy task of simply detecting the presence of a reason. Once it was made clear to the participants that minimal arguments required a reason, they were mostly able to do it successfully without much practice. In contrast, learning to examine the relation between the claim and the reason and to determine whether the reason actually supports the claim is more complex and requires practice with immediate feedback.

Experiment 3

The goal of Experiment 3 was to test the effectiveness of the tutorial in a more authentic, non-experimental context. In order to make the tutorial more accessible to students and teachers, the computer version of the tutorial was converted for Web-based presentation. Experiment 3 provided an assessment of high school students' baseline ability to evaluate arguments, and it tested the effectiveness of an abbreviated Web-based tutorial for a high school classroom. The Web-based tutorial and judgment test from Experiment 2 were condensed because the tutorial, practice, and judgment test had to be administered in a single class period. The teacher, without the presence of the experimenter, had students access the tutorials remotely via computers in their school lab.

Method

Participants—High school students came from two AP psychology classes from a suburban, middle-class school outside of Boston. Both classes had the same teacher. One class had 26 students and the other had 24 students. Four students from each class were not present for one of the two sessions, so their data were not used. Slightly more than half of the students were female (59%). The testing came at the end of the third quarter of the school year and was carried out by the teacher over two class periods.

Flawed judgment test—Because the number of students was limited, a pretest–posttest design was employed. Two versions of the flawed judgment test were created with both the pretest and posttest having eight acceptable arguments, eight unsupported claims, and eight unwarranted arguments. Eight filler arguments were also included to equate the number of *yes* responses.

Tutorials—The immediate feedback tutorial from Experiment 2 was modified for presentation over the Web. Because the tutorial and the posttest flawed judgment test had to be given within the 45-minute class period, the tutorial was shortened by removing about one third of the words. The number of examples and the number of practice items remained the same.

Procedure and Design—The teacher alone directed students through the study. One of the classes was randomly assigned to receive the tutorial condition, and the other class served as a no-treatment control. Students completed the pretest version of the flawed judgment test during a regular class period. One week later, the tutorial condition was given the tutorial in the school computer lab. After 35 minutes, all students were stopped and given the post-test version of the flawed judgment test. Students in the baseline class were given the tasks in the reverse order (i.e., post-test flawed judgment test followed by the tutorial). In the following class period, they were presented with the goals of the study, the hypothesis, the experimental design, and the results to demonstrate how research is conducted.

Results

The average proportion values accuracy on the flawed judgments test are presented in Table 3. A 2×2 , Testing occasion (Pretest vs. Posttest) $\times$ Instructional condition (Baseline vs. Tutorial), mixed ANOVA was conducted with Testing Occasion as a within-subjects factor. There was a significant main effect of Testing Occasion, $F(1, 40) = 10.72$, $MSE = .016$, $p < .01$, $\eta^2 = .211$. Overall students were more accurate on the posttest ($M = .79$, $SD = .13$) than on the pretest ($M = .70$, $SD = .13$). There was also a main effect of Instructional Condition, $F(1, 40) = 14.935$, $MSE = .016$, $p < .01$, $\eta^2 = .272$. The class given the tutorial was more accurate ($M = .80$, $SD = .13$) than the Baseline condition ($M = .69$, $SD = .14$). These main effects were qualified by a significant Testing Occasion $\times$ Instructional Condition interaction, $F(1, 40) = 7.887$, $MSE = .016$, $p < .01$, $\eta^2 = .165$. Bonferroni posthoc tests show a significant improvement on the posttest for the Tutorial group but no improvement for the Baseline group. The same pattern of improvement in the posttest for the tutorial group ($M = .88$, $SD = .12$) over the Baseline group ($M = .70$, $SD = .14$) was found when pretest scores were used as a covariate, $F(1, 39) = 21.190$, $MSE = .017$, $p < .01$, $\eta^2 = .352$.

The results show that high school AP students, like the undergraduates, had difficulty distinguishing arguments on the basis of structural quality ($M = .70$, $SD = .13$). This difficulty with analyzing the structural quality of arguments is even more surprising because the students were near the end of the year in an AP class that has a goal of helping students think, read, and write critically. The results also show that it is practical for a teacher to use a Web-based tutorial in a class period to improve high school students' ability on this task ($M = .80$, $SD = .13$)¹. We suspect that multiple sessions would help even more.

Although, it may be tempting to compare the high school students' performance to that of the college students, the changes in the test make that comparison impossible. Because of the need for a pretest–posttest design and the limited class time in the high school study, the flawed judgment test was shortened and items were randomly assigned to version. Therefore, it is inappropriate to draw any conclusions about the baseline level of skill across the two groups.

Discussion

Three experiments were conducted to assess students' ability to distinguish structurally acceptable arguments from unacceptable arguments, and to test the effectiveness of instructional interventions. In these experiments, participants evaluated the quality of items that were either complete arguments (claim-reason pairs), unsupported claims (no reason), or unwarranted arguments (reason does not support claim). We found that without an argument tutorial, both high school and college students frequently failed to distinguish acceptable arguments from structurally flawed arguments. We also found that a short tutorial improved students' ability to reject unsupported assertions but students needed immediate feedback during training to improve their detection of unwarranted arguments. College students' accuracy at argument detection significantly improved with feedback.. High school students also improved with feedback.

Although the tutorial was effective, it was a very brief exposure (15–35 minutes). It is not surprising that students may need multiple exposures to help them reach mastery level performance on this skill. In Experiment 2, for example, about half of the students did not reach a minimal mastery level of 80% accuracy². Expertise takes time to develop (Ericsson, Krampe,

¹The high school students rated each instructional elements of the tutorial as helpful (median rating of 7 for each component, with 1 being very unhelpful and 9 being very helpful). The students, however, thought the presentation style could be improved (median rating of 4 on aesthetics, capturing interest, and enjoyment). We are currently working to make the tutorial more interesting and enjoyable without negatively impacting the learning gains.

& Tesch-Romer, 1993; Feltovich, Prietula, & Ericsson, 2006), and although true expertise is not the goal, mastery may take more than a single exposure for some students. In fact, students frequently have trouble learning reasoning skills even after an entire term of logic (Leighton, 2006). therefore, it is reasonable to expect that multiple exposures could improve the chance of long-term retention of the skill and eventual transfer of the skill to other situations.

An added benefit of multiple exposures is the opportunity for reaping the benefits of distributed learning. Research has shown that distributed learning leads to better retention (Dempster, 1987; Glenberg, 1979; Reder & Anderson, 1982) and longer intervals between exposures (e.g., 1–2 months) leads to better long-term retention than shorter inter-session intervals (Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Bahrick & Hall, 2005; Pashler, Zarow, & Triplett, 2003). Thus, average students may need multiple appropriately spaced sessions to achieve mastery of these skills. Additional gains in argument skill may also be accomplished by adapting the tutorial to provide more information regarding specific argumentative moves, such as counterarguments or rebuttals. Instruction in these areas may lead students to further realize the importance of attending to the warrant of the argument.

The positive effects of training students to evaluate the quality of arguments have implications for improving students' abilities in other areas. For example, students are frequently required to write research papers and essay exams. History students may need to write an essay taking a stance on whether Andrew Carnegie was responsible for the breaking of the union in homestead Pennsylvania. Psychology students may need to write a term paper that makes a case that antidepressant drugs are beneficial to adolescents even though they are associated with an increase in suicidal ideation. In addition, biology students may need to write a research paper arguing that global warming is caused by human activity.

An analysis of the component skills involved in writing such papers include comprehending other's arguments in the documents they read, deciding on a main claim (or thesis), selecting relevant information, integrating that information into a coherent argument, and writing the paper. Each of these component skills can be influenced by an improved ability to accurately assess whether arguments are complete and warranted. In the following section, we speculate on the impact of the skill of evaluating argument quality for each of these component skills.

Comprehending other's arguments

The students' task of writing a research paper requires students to accurately comprehend the arguments presented in the sources they read to include in their paper. In doing so, students must detect an author's main assertion (or claim), paying special attention to the predicate of the claim. It is a nontrivial task for students to comprehend naturalistic arguments. Larson, Britt, & Larson (2004) found that college students were only about 30% accurate when identifying authors' claims and supporting reasons in short arguments (about 400 words each). Furthermore, many of the reasons they identified could not have supported the claim they identified. Training students to check whether the reasons support the claim should help them correctly identify an author's reasons for a claim. Also, the tutorial we used, with its focus on precise wording, should help students attend to the predicates that mark assertions. There are many lexical, semantic, and pragmatic cues that can indicate whether a statement is an assertion that requires support. As Britt and Larson (2003) found, students are able to use modals (e.g., should) and uncertainty markers (e.g., probably) to identify claims. Therefore, students taught to attend to predicates should be better attuned to statements that require support. The tutorial also should help students appreciate the need to support these assertions and evaluate assertions

²We have selected 80% because it roughly corresponds to 2 standard deviations above the mean performance by control participants and is a typical cutoff for "B" level of performance in academic settings.

differently if no evidence is provided. Thus, the tutorial should help students read, represent, and evaluate others' arguments—skills necessary for conducting research to include in a research paper.

Deciding on a main claim

The tutorial will not be as directly helpful for students in deciding on a thesis or main claim. However, it should stress the importance of selecting a thesis for which they can find support and maintaining the focus on that main claim.

Selecting and integrating information

A research paper also requires students to select relevant information from the articles they read and integrate this information into a coherent and persuasive argument. This involves making sure that each reason supports the main claim (warranted) and that supporting reasons are also not a series of unsupported assertions. The tutorial should help students understand the need for supporting controversial assertions that one is trying to get the reader to believe. Several important aspects of the task of selecting and integration information not addressed by the tutorial are evaluating the relative strength of potential reasons and making sure that all well-establish facts are presented. The tutorial can be enhanced by teaching students to evaluate the sufficiency of reasons given the totality of the known evidence (Blair & Johnson, 1987). Students' ability to integrate information into an argument and the processes involved in successful integration is an area that should be researched further.

Writing the research paper

Last, there is the actual writing process. The tutorial may be helpful to the extent that this tutorial helps students appreciate the need for support for all assertions and to attend to the precise wording of the predicates of each assertion. It may also help students to organize the support and make all connections between claims and reasons as explicit as necessary for the audience. This may require making the warrant explicit and even providing support for it.

Possible implications of improving students' ability to evaluate arguments may also prove important beyond the classroom. First, from an economic perspective, it is harmful for consumers to be easily persuaded by false evidence. Many companies advertise through nonrational means (Sagarin et al., 2002), such as using actors who play doctors as experts on medical products. Because the actor is not actually an expert, his or her testimony or opinion does not warrant the conclusion. Helping students question why this type of evidence supports the assertion would lead to more informed decision making.

Second, and perhaps more important, improving the electorate's ability to evaluate argument quality is important in a democracy. The key to a healthy democracy is its citizens' being are knowledgeable about public issues and able to reason about those issues. Citizens must be able to determine whether evidence is present or whether presented evidence warrants the conclusion. A more general implication of these results is that people may be unskilled at evaluating an answers from politicians to reporters' questions. For example, often when reporters ask politicians challenging questions, the artful politicians twist answers so they are not actually addressing what the reporters want and instead change the questions to ones they prefer to answer. The results of the present study suggest that not all listeners are able to detect whether politician actually answer questions. Much like the reasoners having to keep the claim predicate active while reading the reason, here the listener has to keep the predicate of the question active while listening to the question. Then they must detect whether the answer is relevant to the question asked.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H05133 to Northern Illinois University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We wish to thank James Voss for comments and helpful suggestions.

References

- American Association for the Advancement of Science. Benchmarks for Scientific Literacy. 1993. Retrieved July 20, 2006, from <http://nces.ed.gov/nationsreportcard/science/achieveall.asp>
- Anderson, JR. Rules of the Mind. Hillsdale, NJ: Erlbaum; 1993.
- Aristotle. Introduction to Aristotle. 2. McKeon, Richard, editor. Chicago: University of Chicago Press; 1973.
- Bahrack HP, Hall LK. The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language* 2005;52:566–577.
- Bahrack HP, Bahrack LE, Bahrack AS, Bahrack PE. Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science* 1993;4:316–321.
- Baron, J. Beliefs about thinking. In: Voss, JF.; Perkins, DN.; Segal, JW., editors. *Informal reasoning and education*. Hillsdale, NJ: Erlbaum; 1991. p. 169–186.
- Bjork, RA. Assessing our own competence: Heuristics and illusions. In: Gopher, D.; Koriati, A., editors. *Attention and performance XVII: Cognitive regulation of performance Interaction of theory and application*. Cambridge, MA: MIT Press; 1999. p. 435–459.
- Blair JA, Johnson RH. Argumentation as dialectical. *Argumentation* 1987;1:41–56.
- Brainerd CJ, Reyna VF. Explaining “memory free” reasoning. *Psychological Science* 1992;3:332–339.
- Brainerd CJ, Reyna VF. When things that were never experienced are easier to “remember” than things that were. *Psychological Science* 1998;9:484–489.
- Bransford, JD.; Brown, AL.; Cocking, RR. *How people learn*. Washington, DC: National Academy Press; 1999.
- Britt MA, Kurby CA, Dandotkar S, Wolfe CR. I Agreed with What? Memory for Simple Argument Claims. *Discourse Processes* 2008;45:52–84.
- Britt MA, Larson A. Construction of argument representations during on-line reading. *Journal of Memory and Language* 2003;48:749–810.
- Bruer, JT. *Schools for thought: A science of learning in the classroom*. Cambridge, MA: MIT Press; 1993.
- Chi MTH, Bassok M, Lewis MW, Reimann R, Glaser R. Self-explanation: How students study and use examples in learning to solve problems. *Cognitive Science* 1989;13:145–182.
- Chi, MTH.; Glaser, R.; Rees, E. Expertise in problem solving. In: Sternberg, R., editor. *Advances in the psychology of human intelligence*. Vol. 1. Hillsdale, NJ: Erlbaum; 1982. p. 17–76.
- Chi MTH, de Leeuw N, Chiu M, LaVancher C. Eliciting self-explanations improves understanding. *Cognitive Science* 1994;18:439–477.
- Chinn CA, Anderson RC. The structure of discussions that promote reasoning. *Teachers College Record* 1998;100:315–368.
- Corbett, AT.; Anderson, JR. Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. *Proceedings of ACM CHI'2001 Conference on Human Factors in Computing Systems*; Seattle, WA. 2001 Mar. p. 245–252.
- Deci EL. Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology* 1971;18:105–115.
- Deci, EL.; Ryan, RM. *Intrinsic motivation and self-determination in human behavior*. New York: NY. Plenum; 1985.
- Dempster FN. Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology* 1987;79:162–170.
- deWinstanley PA, Bjork EL. Processing strategies and the generation effects: Implications for making a better reader. *Memory and Cognition* 2004;32:945–955.

- Driver R, Newton P, Osborne J. Establishing the norms of scientific argumentation in classrooms. *Science Education* 2000;84:287–313.
- Duschl, R. Making the nature of science explicit. In: Millar, R.; Leach, J.; Osborne, J., editors. *Improving science education: The contribution of research*. Philadelphia, PA: Open University Press; 2000.
- Duschl R, Osborne J. Supporting and promoting argumentation discourse in science education. *Studies in Science Education* 2002;38:39–72.
- Duschl, R.; Ellenbogen, K.; Erduran, S. Middle school students' dialogic argumentation. In: Komorek, M.; Behrendt, H.; Dahncke, H.; Duit, R.; Graber, W.; Kross, A., editors. *Research in science education: Past, present and future; Proceedings of the Second International Conference of the European Science Education Research Association*; Kiel. 1999. p. 420-422.
- Ericsson, KA.; Simon, HA. *Protocol analysis: Verbal reports as data*. 2. Cambridge, MA: MIT Press; 1993.
- Ericsson KA, Krampe RT, Tesch-Romer C. The role of deliberate practice in the acquisition of expert performance. *Psychological Review* 1993;100:363–406.
- Feltovich, PJ.; Prietula, MJ.; Ericsson, KA. Studies of expertise from psychological perspectives. In: Ericsson, KA.; Charness, N.; Feltovich, PJ.; Hoffman, RR., editors. *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge, England: Cambridge University Press; 2006.
- Forman EA, Larreamendy-Joerns J, Stein MK, Brown CA. "You're going to want to find out which and prove it": Collective argumentation in a mathematics classroom. *Learning and Instruction* 1998;8:527–548.
- Glenberg AM. Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition* 1979;7:95–112.
- Grennan, W. *Informal Logic: Issues and Techniques*. Montreal, Quebec: McGill-Queen's University Press; 1997. The role of arguer credibility in argument evaluation.
- Grice, HP. *Logic and Conversation*. In: Cole, P.; Morgan, J., editors. *Syntax and Semantics 3: Pragmatics*. New York: Academic Press; 1975.
- Harackiewicz J. The effects of reward contingency and performance feedback on intrinsic motivation. *Journal of Personality and Social Psychology* 1979;37:1352–1363.
- Jimenez-Aleixandre M, Rodrigues M, Duschl RA. 'Doing the lesson' or 'doing science': Argument in high school genetics. *Science Education* 2000;84:757–792.
- Jonassen, DH.; Tessmer, M.; Hannum, WH. *Task Analysis Methods for Instructional Design*. Mahwah, New Jersey: Erlbaum; 1999.
- Kelly GJ, Druker S, Chen C. Students' reasoning about electricity: combining performance assessments with argumentation analysis. *International Journal of Science Education* 1998;20:849–871.
- Kintsch W. The role of knowledge in discourse comprehension construction-integration model. *Psychological Review* 1988;95:163–182. [PubMed: 3375398]
- Kintsch, W. *Comprehension: A paradigm for cognition*. New York: Cambridge University Press; 1998.
- Kintsch W, van Dijk TA. Toward a model of text comprehension and production. *Psychological Review* 1978;85:363–394.
- Koriat A, Bjork RA. Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition* 2006;34:959–972.
- Kuhn D, Udell W. The development of argument skills. *Child Development* 2003;74:1245–1260. [PubMed: 14552396]
- Kuhn, D. *The skills of argument*. Cambridge, England: Cambridge University Press; 1991.
- Kuhn D. Science as argument: Implications for teaching and learning scientific thinking. *Science Education* 1993;77:319–337.
- Kulhavy RW, Stock W. Feedback in written instruction: The place of response certainty. *Educational Psychology Review* 1989;1:279–308.
- Kulik JA, Kulik CL. Timing of feedback and verbal learning. *Review of Educational Research* 1988;58:79–97.
- Kurby, CA.; Britt, MA.; Dandotkar, S. Representing Argument Claims: Availability and Accessibility of the Predicate and Implications on Argument Evaluation. Poster session presented at the 16th Annual Conference of the Society for Text and Discourse; Minneapolis, Minnesota. 2006 Jul.

- Larson M, Britt MA, Larson A. Disfluencies in comprehending argumentative texts. *Reading Psychology* 2004;25(3):205–224.
- Lawson A. The nature and development of hypothetico-predictive argumentation with implications for science teaching. *International Journal of Science Education* 2003;25:1387–1408.
- Lederman NG. Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching* 1992;29:331–359.
- Lehman DR, Nisbett RE. A longitudinal study of the effects of undergraduate training on reasoning. *Developmental Psychology* 1990;26:952–960.
- Leighton JP. Teaching and Assessing Deductive Reasoning Skills. *The Journal of Experimental Education* 2006;74:109–136.
- Lesgold, AM.; Feltovich, PJ.; Glaser, R.; Wang, Y. Technical report no. PDS-1. Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh; 1981. The acquisition of perceptual diagnostic skill in radiology.
- Levi DS. The case of the missing premise. *Informal Logic* 1995;17:67–88.
- Mans D, Preyer G. Reasoning and Argumentation. *Proto Sociology* 1999;13:3–13.
- Marsh EJ, Edelman G, Bower GH. Demonstrations of a generation effect in context memory. *Psychonomic Society* 2001;29:798–805.
- Mayer, RE. Learning and instruction. Upper Saddle River: Prentice Hall; 2003.
- McNamara DS. SERT: Self-Explanation Reading Training. *Discourse Processes* 2004;38(1):1–30.
- National Assessment of Educational Progress. NAEP 1994 U.S. History Report Card: Findings from the National Assessment of Educational Progress. Princeton, NJ: Educational Testing Service; 1996.
- National Assessment of Educational Progress. NAEP 1998 Writing Report Card: Findings from the National Assessment of Educational Progress. Princeton, NJ: Educational Testing Service; 1998.
- National Center for History in the Schools. National Standards for United States History: Exploring the American Experience. Los Angeles: University of California, Los Angeles; 1996.
- National Science Board. America's pressing challenge-building a stronger foundation. Jan. 2006 Retrieved July 20, 2006, from <http://www.nsf.gov/statistics/nsb0602/nsb0602.pdf>
- National Science Education Standards: An Overview. 1996. Retrieved July 20, 2006, from <http://darwin.nap.edu/html/nses/overview.html>
- Newell, A. Metaphors for mind, theories of mind: Should the humanities mind?. In: Sheehan, JJ.; Sosna, M., editors. *The boundaries of humanity: Humans, animals, machines*. Berkeley, CA, US: University of California Press; 1991. p. 158-197.p. xp. 274
- No Child Left Behind Act (2001).
- Osborne J, Erduran S, Simon S. Enhancing the quality of argumentation in science classrooms. *Journal of Research in Science Teaching* 2004;41:994–1020.
- Pashler H, Cepeda N, Wixted J, Rohrer D. When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 2005;31:3–8.
- Pashler H, Zarow G, Triplett B. Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 2003;29:1051–1057.
- Perkins, DN.; Farady, M.; Bushey, B. Everyday reasoning and the roots of intelligence. In: Voss, JF.; Perkins, DN.; Segal, JW., editors. *Informal reasoning and education*. Hillsdale, NJ: Erlbaum; 1991. p. 83-105.
- Psychology Software Tools. E-Prime (Version 1.1) [Computer software]. Pittsburgh, PA: 2002.
- Reder LM, Anderson JR. Effects of spacing and embellishment on memory for the main points of a text. *Memory & Cognition* 1982;10(2):97–102.
- Reznitskaya A, Anderson RC, McNurlen B, Nguyen-Jahiel K, Archodidou A, Kim S. Influence of oral discussion on written argument. *Discourse Processes* 2001;32:155–175.
- Sagarin BJ, Cialdini RB, Rice WE, Serna SB. Dispelling the illusion of invulnerability: The motivations and mechanisms of resistance to persuasion. *Journal of Personality and Social Psychology* 2002;83:526–541. [PubMed: 12219852]
- Schwarz BB, Neuman Y, Gil J, Iiya M. Construction of collective and individual knowledge in argumentative activity. *Journal of the Learning Sciences* 2003;12:219–256.

- Siegel H. Why Should Educators Care About Argumentation? *Informal Logic* 1995;17:159–176.
- Slamecka NJ, Graf P. The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory* 1978;6:592–604.
- Stanovich KE, West R. Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology* 1997;89:342–357.
- Takao AY, Kelly GJ. Assessment of evidence in university students' scientific writing. *Science & Education* 2003;12:341–363.
- Toulmin, SE. *The uses of argument*. Cambridge, MA: Cambridge University Press; 1958.
- Veerman A, Andriessen J, Kanselaar G. Collaborative Argumentation in Academic Education. *Instructional Science* 2002;30:155–186.
- Voss JF, Means ML. Learning to reason via instruction in argumentation. *Learning & Instruction* 1991;1:337–350.
- Voss, JF.; Perkins, DN.; Segal, JW. *Informal Reasoning and Education*. Hillsdale, NJ: Erlbaum; 1991.
- Watson DL, Kessler DA, Kalla S, Kam CM, Ueki K. Active learning exercises are more motivating than quizzes for underachieving college students. *Psychological Reports* 1996;78:131–134.
- Zohar A, Nemet F. Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching* 2002;39(1):35–62.

Biographies

Aaron A. Larson is affiliated with Northern Illinois University (NIU). His research interests include argument evaluation and the effects of knowledge on problem solving.

M. Anne Britt is an associate professor in the psychology department at NIU. She has developed tutorials for improving students' advanced literacy skills, such as argument comprehension, sourcing, and content integration.

Christopher Kurby is affiliated with psychology department at Washington University, St. Louis. His interests focus on how people create representations of events in narrative and the perceptual world.

Table 1

Average proportion correct on the flawed judgment task for each Tutorial condition and Quality Type for Experiment 1 (Standard Deviations in parentheses).

Instructional condition	Quality type			Average
	Acceptable	Unsupported	Unwarranted	
Baseline (No treatment)	.69 (.22)	.61 (.27)	.70 (.27)	.66 (.15)
Tutorial	.72 (.18)	.81 (.23)	.66 (.23)	.73 (.14)
Experts (Grad students)	.80 (.15)	.93 (.10)	.97 (.06)	.90 (.08)

Table 2

Average proportion correct on the flawed judgment task for Argument Type and *Tutorial* condition for Experiment 2 (Standard deviations in parentheses).

Instructional Condition	Quality Type			
	Acceptable	Unsupported	Unwarranted	Changed Theme
Baseline (No treatment)	.67 (.22)	.61 (.23)	.59 (.25)	.56 (.18)
No-feedback tutorial	.69 (.25)	.81 (.27)	.60 (.29)	.63 (.25)
Feedback tutorial	.69 (.19)	.86 (.18)	.76 (.19)	.74 (.18)

Table 3

Average proportion correct on the pretest and posttest flawed judgment task for each Instructional condition for Experiment 3(Standard deviations in parentheses).

Instructional condition	Test occasion	
	Pretest	Posttest
Baseline	.68 (.13)	.70 (.14)
Tutorial	.71 (.13)	.88 (.12)