Three experiments were conducted to examine how precisely readers recall the claims of arguments that they have just read. Participants read simple, 2-clause arguments such as, “The U.S. is right to intervene in other countries’ affairs because local events can catastrophically impact the entire world.” Participants then evaluated each argument with respect to agreement or fit and recalled either the claim or the complete argument. The first 2 experiments found that readers have difficulty precisely recalling the main predicate of the claim (e.g., “should intervene” or “is right to intervene”) as accurately as they do the theme of the argument. Furthermore, a comparison with predicates of short narrative statements indicates that this recall difficulty is specific to argument predicates. Experiment 3 found that skilled readers’ recall was more accurate than less-skilled readers. It also showed that the predicate recall problem occurred with both skilled and less-skilled readers and that it is related to one’s ability to detect poorly formed arguments.

Cognitive research tells us that when people read they create both a temporary, verbatim representation of the words in a text and a more durable, gist representation of the text’s meaning (Brainerd & Reyna, 1998; Johnson-Laird, 1983; Kintsch, 1988, 1998; Kintsch & van Dijk, 1978; Roediger & McDermott, 2000). For many reading tasks, the quick loss of the verbatim representation is not problematic because comprehension and even problem solving can proceed based on the gist rep-
resentation alone. There are other types of reading tasks, however, such as reading argumentative prose for which depending on a gist representation might cause difficulties. Detecting flaws in an argument and generating counter-arguments often requires making subtle distinctions of meaning and using words precisely—skills that depend on remembering the exact wording of a text.

In this article, we present three experiments that examine readers’ ability to precisely recall argument claims and the impact this precision has on their ability to evaluate arguments. In Experiments 1 and 2, we find that immediate recall of claim predicates is impaired relative to recall for other claim elements and non-claim predicates. Experiment 3 then relates this deficit to reading ability: Skilled readers recalled predicates better than less-skilled readers, but both still showed poorer recall for claim predicates than claim themes. This experiment also found that skill in recalling claim predicates is associated with poorer detection of faulty arguments.

**READING ARGUMENTS**

An argument is an attempt to persuade a reader or listener to accept an assertion (referred to as the claim) by providing one or more supporting reasons. Analysis of argument forms by philosophers and rhetoricians dates back to ancient times, but only recently has this genre begun to receive attention from text researchers. Voss and Means (1991) used a model of argumentation in jurisprudence developed by Toulmin (1958) in their examination of students’ ability to evaluate arguments. Toulmin described arguments in terms of the relations among key elements, identified as the claim, reason (“datum” in Toulmin’s terminology), warrant, qualifier, rebuttal, and backing. Argumentation involves supporting a claim by providing a set of reasons and possibly modifying the claim–reason relation by providing warrants, qualifiers, backing and rebuttals. To illustrate, consider Argument 1:

1. Banning cell phone use while driving is unfair because it’s only a few drivers who can’t handle a phone conversation behind the wheel.

The claim in this example is, “Banning cell phone use while driving is unfair.” The reason, “it’s only a few drivers who can’t handle a phone conversation behind the wheel,” provides the required support for the claim. Claims and reasons are connected by warrants—general principles that permit the conclusion given the reason. Warrants are sometimes explicitly stated, but usually they are not. The implied warrant for Argument 1 would be something like, “Banning something because a few people can’t handle it is unfair.” Toulmin described the way that warrants connect claims and reasons as a kind of general enablement or licensing. Our use of the concept in this article is more in line with that of a general statement that pro-
vides a simple logical connection between the claim and its support that can be inferred by a reader or listener.

Claims have a special status in arguments. The claim not only establishes the point of contention, it also acts as an organizer for the subsequent discourse. Britt and Larson (2003) found that the order in which argument elements were presented influenced participants’ reading time and recall of simple, two-sentence arguments. Arguments presented in a claim–reason order were read faster and recalled more accurately than the identical argument presented in a reason–claim order. This effect is likely due to the claim’s role in establishing the logical relations among argument elements that are required for text coherence. Reading in a claim-first order enables constraints and expectations for those connections to be set up prior to encountering them in the text.

The claim also guides activation of relevant knowledge and beliefs. According to the three-phase processing model of argument comprehension developed by Voss, Fincher-Kiefer, Wiley, and Silfies (1993), reading a claim initially activates one’s attitude toward the theme of the claim. For instance, reading the claim, “The death penalty is immoral,” leads to the activation of other beliefs about the death penalty. Strongly held beliefs are activated more quickly than moderately held beliefs. Next, the claim activates related reasons, with an advantage for attitude-consistent, “pro” reasons. Finally, the values associated with the reason–claim connection (warrant) become activated. According to this model, argument-related activation is claim-centered, and claim-related elaborations are possible early in processing.

Claims themselves can be decomposed into elements. In our research we have looked at three such components: theme, predicate, and side. The theme is the topic of the argument such as cell phone use, marijuana use, and car insurance rates. It is frequently, but not always, the subject noun phrase of the claim. In claims 2a through 2c the theme is “cell phone use while driving.” The predicate refers to the main verb or predicate adjective in the claim. Claims 2a through 2c differ with respect to the predicate asserted: “should not be banned,” “is unfair,” and “is unnecessary”:

2a. Cell phone use while driving should not be banned.
2b. Banning cell phone use while driving is unfair.
2c. Banning cell phone use while driving is unnecessary.

Policy predicates, such as “should not be banned” in 2a, assert actions and behaviors as solutions to problems; whereas value predicates, such as “is unfair” in 2b and “is unnecessary” in 2c, address the desirability or morality of things or actions. The side of an argument represents whether the author is arguing for or against the proposal. Claims 2a through 2c all argue for the same general side of the controversy (against banning cell phone use while driving). Because the theme and side
of each claim are the same, the mental representation of each of these claims would be different only for readers who precisely represent the predicate.

Two prior studies have suggested that imprecise representations might be related to comprehension and evaluation problems with arguments. Larson, Britt, and Larson (2004) examined undergraduates’ ability to identify argument elements in complex argumentative texts. They found that participants accurately identified only 30% of the claims and reasons in these texts. An analysis of the identification errors indicated that most of the identified reasons could not actually provide support for the claim as stated by the participant. For example, one student identified the claim as, “The death penalty is an ineffective punishment,” and then identified the reasons to be: “Death penalty is not fair” and “It discriminates by race.” Although the texts were in front of them, some participants did not maintain a sufficiently precise representation of the claim to recognize that the statements they were identifying as reasons did not actually support the claim. Such a result might occur if a reader’s representation of a claim consisted of only theme and side. A reader whose representation of a claim was something like “against death penalty” might easily identify as reasons statements that were generally supportive of this impoverished claim (e.g., “the death penalty discriminates by race”) but not supportive of the specific claim that included the claim’s predicate.

Britt and Kurby (2005) found a similar result on a task that had undergraduate and graduate students explicitly judge whether a reason supported a claim. In this study, participants were presented with a series of arguments that were either well-structured (such as 3a) or structurally flawed (such as 3b):

3a. Warranted argument: Banks shouldn’t charge ATM fees because the fees make their customers unhappy.
3b. Unwarranted argument: Banks shouldn’t charge ATM fees because banks are financial institutions.

They found that undergraduates were significantly less able to reject unwarranted arguments than were graduate students (68% vs. 97%, respectively). Like the Larson et al. (2004) result, this study suggests that either participants do not check the logical connection between the claim and reason when reading arguments or that they are using an insufficiently precise representation when doing this check. The graduate students, in contrast, appear to perform this check and have a sufficiently precise representation for this check to yield accurate results. Although both of these experiments hint at the importance of predicate memory, neither directly tested how accurately participants recalled claim predicates.

One reason why readers might fail to represent arguments precisely is related to the nature of reading and memory processes themselves. Influential theoretical work in the areas of text processing (Kintsch & van Dijk, 1978), reasoning (Brainerd & Reyna, 1998; Johnson-Laird, 1983), and false memories (Brainerd &
Reyna, 1998; Roediger & McDermott, 2000) suggests that people process information at multiple levels that differ in terms of how precisely the representation corresponds to the stimulus. It may be the case that students’ argumentation problems are a result of normal text processing.

MEMORY FOR TEXT

Models of text processing (Kintsch & van Dijk, 1978), reasoning (Brainerd & Reyna, 1998; Johnson-Laird, 1983), and false memories (Brainerd & Reyna, 1998; Roediger & McDermott, 2000) all propose that during language comprehension, readers create multiple, independent levels of representation simultaneously. One level is detailed, close to a verbatim copy of the text or utterance that preserves the surface form of the message. Another level is a less-detailed, gist representation that contains only the meaning of the message elaborated by inferences and other knowledge. Memory for the verbatim copy fades quickly, but memory for the gist or meaning-level representation persists (Kintsch, Welsch, Schmalhofer, and Zimny, 1990); and it is this representation that people tend to use when reasoning and solving problems (Brainerd & Reyna, 1992).

According to Kintsch and van Dijk (1978; see also Kintsch, 1988, 1998; van Dijk and Kintsch, 1983), people simultaneously form three different representations of a text while reading: a surface code, a textbase, and a situation model. The surface code is an unembellished representation of the verbatim text without abstraction. For example, although sentences 4a through 4d present the same general message, they use different syntactic forms or lexical items that would result in different surface representations for these sentences:

4a. The death penalty should be abolished.
4b. We should abolish the death penalty.
4c. We should get rid of the death penalty.
4d. The state of Illinois should change the law and stop using the death penalty as punishment.

The textbase level is a representation of the meaning or semantics of the text independent of the exact wording or syntactic structure used to convey that meaning. Example sentences 4a through 4c would lead to the same textbase representation; however, 4d would lead to a different textbase representation because it includes additional details and requires extra-textual knowledge to interpret. Finally, the situation model is a representation of the events described by the text with elaboration from the reader’s relevant background knowledge. Unlike the relatively encapsulated textbase and surface form, the construction of a situation model is assumed to be an integration of the information in the text with whatever back-
ground knowledge becomes activated. For example, someone living in Illinois might form the same situation model from reading all four of the previous example sentences.

This multilevel model was partially motivated by experiments in the late 1960s and early 1970s showing that readers are more accurate at detecting changes to the textbase than to the surface form. For example, Sachs (1967) found that participants could detect both surface changes (e.g., syntactic form or paraphrases) and semantic changes immediately after reading a sentence, but after a delay only the semantic changes were still detectable. These findings suggested that representation of meaning is distinct from verbatim memory and that although both types are initially represented, the representation of meaning persists longer. Similarly, Bransford and Franks (1971) gave participants a series of related sentences and asked them to judge whether a target sentence was old (in the original set) or new. Participants frequently accepted as old, sentences that were semantically consistent with the information in the original set. These results indicated that a gist representation of the meaning of a group of sentences is stored independently of the actual verbatim form of the presented material. Research from other labs further corroborated this distinction (e.g., Schmalhofer & Glavanov, 1986). In a more recent experiment, Kintsch et al. (1990) examined the time course of each type of representation. Immediately after reading, the surface level representation was available as indicated by a much higher recognition of originally presented items compared to paraphrased items. As the time delay increased, the difference between acceptance of verbatim items and paraphrased items decreased until it disappeared after a 4-day delay. Thus, participants could detect differences between exact wording of a statement and its meaning immediately after reading narratives but gradually only the meaning was available or accessible.

Recent research on false memory also supports the distinction between gist and verbatim memories (Brainerd & Reyna, 1998; Roediger & McDermott, 2000). Roediger and McDermott presented participants with lists of related terms such as bed, rest, dream, and blanket and then asked them to immediately recall the list. After all lists were presented, they were given a recognition test with an unpresented but highly associated word as a distractor (e.g., sleep). They found that participants frequently recalled this unpresented, strong associate incorrectly and recognized it at a comparable rate to the average presented item. In fact, participants actually stated that they remembered the item rather than just knowing it was presented. Similarly, Brainerd and Reyna (1998) found that items highly associated with the target set serve as strong retrieval cues for the list.

This gist–verbatim distinction has also been shown to influence reasoning, judgment, and decision making. One counter-intuitive finding in children’s reasoning about class inclusion problems, such as 5, is that memory for the details does not predict solution accuracy:
5. There are 7 cows, 3 horses, and 10 animals. Are there more animals or more horses?

Brainerd and Reyna (1992) found that for such problems specific memory for details (e.g., the number of horses) was not related to whether the child could solve the problem. This disconnect between how accurately participants recalled the details of a problem and their solution accuracy led to the conclusion that participants were solving these problems from a non-verbatim representation. To explain such findings across reasoning problem types, Brainerd and Reyna (1990) developed their Fuzzy-trace theory (Reyna & Brainerd, 1992). According to this theory, reasoners encode two independent representations, verbatim and gist, in parallel. The verbatim representation, like the surface form, is essentially an exact copy of the stimulus; whereas the gist representation, like the situation model, is an abstraction of the meaning of the stimulus that is connected to prior knowledge. Verbatim representations are less durable than gist representations and after prolonged delays, only the gist representation is accessible. As a result, gist representations are often relied on during reasoning or decision making (Reyna & Brainerd, 1995; Wolfe, 1995). Successful reasoning depends on using the most appropriate form of representation. For some types of reasoning problems, such as relational and class inclusion problems, the gist form makes salient the necessary relations, whereas the verbatim form is either distracting or misleading (Brainerd & Reyna, 1998; Reyna & Adam, 2003).

IMPORTANCE OF VERBATIM MEMORY FOR ARGUMENT PROCESSING

Although reliance on a gist representation may be acceptable for many tasks, such as comprehending narrative texts and reasoning through certain types of word problems, there are other types of tasks and texts for which retaining detail and precise wording is important (Brainerd & Reyna, 2005). Comprehension of expository texts (history and science texts), procedures (e.g., computer manuals), legal discourse (e.g., contracts and courtroom testimony), and argumentative texts all require one to remember details and sometimes make fine distinctions among nuances of meaning. For argument comprehension, a gist level of representation would be generally too vague to lead to accurate analysis. For example, if one reads a claim that “the death penalty is immoral,” the gist representation might be something like “the death penalty is bad” or “the death penalty should not be used.” Several other claims such as “the death penalty is ineffective” and “the death penalty is biased” can also result in the same general gist representation. Although this family of claims may all lead to a similar gist representation, they are not all synonymous. Not all of these claims will be warranted by the same reasons or backed by the same type of evidence. For example, the reason, “Many sentenced to death
have been found innocent using DNA,” can support the claim that the death penalty is immoral but not that it is ineffective. Readers who rely on a gist representation of arguments might represent the claim of Argument 6 as “the death penalty is bad” or “the death penalty shouldn’t be allowed”:

6. The death penalty is ineffective because many sentenced to death have been found innocent using DNA.

Then, when reading the reason, they will erroneously judge the argument as warranted because the reason does support the gist representation but not the precise representation. Thus, we would expect readers who rely on a gist representation when comprehending and evaluating arguments to have difficulty detecting flaws in arguments.

The experiments presented in this article examine whether participants are able to precisely recall the claim of an argument they have just read. There are three comparisons across these experiments that provide a measure of the extent to which claims are precisely represented. First, in Experiments 1 and 3, the recall of different claim elements (theme and predicate) is compared. Second, in Experiment 2, the recall of claim predicates is compared to the recall of narrative predicates to determine whether any problem recalling claim predicates was due to a general memory or motivational problems or something specific with predicates. Finally, in all three experiments, the type of predicate, policy, or value is manipulated to determine whether one type is recalled more precisely than another. Each claim has both a policy version (i.e., claims that assert actions and behaviors as solutions to problems) and a value version (i.e., claims that address the desirability or morality of things). An informal analysis of recall data from a previous study (Britt & Larson, 2003) found that, under delayed recall, there was a tendency for people to mistakenly recall value and factual claims as policy claims. The Britt and Larson study, however, was not designed to explicitly test recall differences between these types of claims. We suspected that this tendency might be due to the asymmetry in how these types of claims are used: Values and facts are used to support policy claims, but the reverse is not true. For example, one might argue that we should abolish the death penalty because it is immoral, but one would not argue that the death penalty is immoral because we should abolish it. Such an asymmetry in usage might make readers more likely to blur the distinction when presented with a value claim. The general implied warrant that “we shouldn’t do something that is bad” can apply to any value claim, converting it into a policy. This might be especially likely if readers are relying on only a gist representation of an argument. They may be more likely to recall policy claims accurately than value claims and errors will more likely be in the direction of falsely recalling a value as a policy than a policy as a value. The presence of such recall differences under immediate recall conditions would indicate that recall is based on a gist or situation model representation rather than a verbatim representation.
Experiment 1 examines recall of a claim’s theme and predicate for policy and value versions of the predicates after rating one’s agreement with the argument. A difference in recall accuracy between policy and value versions would indicate that participants did not use a verbatim representation of the claim at the time of recall. Experiment 2 changes the point of recall to occur immediately after participants read the claim and attempts to rule out general memory or motivation difficulties as the cause of recall errors by comparing claim recall with narrative event recall. Highly accurate event-predicate recall relative to claim-predicate recall would indicate that the recall problem is specific to predicates in claims. Finally, Experiment 3 tests whether skilled readers have better memory for claim predicates than less-skilled readers. It also tests whether those readers who more precisely recall claim predicates are also more accurate at detecting unwarranted arguments.

**EXPERIMENT 1**

Experiment 1 examined how accurately participants remember the precise wording of a claim under more “naturalistic” argument evaluation conditions. We consider this naturalistic because it is similar to real world tasks that readers engage in such as answering surveys or reacting to someone else’s argument. To assess the accuracy of recall for a claim that was just evaluated, Experiment 1 had participants state their opinion on an argument and then recall the argument they just evaluated. Participants read a series of two-clause arguments in a claim–reason order. One half of the claims contained policy predicates, and one half contained value predicates. After reading each argument, they rated how much they agreed or disagreed with it and then immediately recalled the argument. Given the shortness of the arguments, the relative immediacy of recall, and participants’ awareness that they would need to recall the claim, one would expect recall accuracy to be quite high and equal for themes and predicates. Furthermore, if participants still have access to their verbatim representation during argument evaluation, both policy and value predicates ought to be recalled with equal accuracy and near ceiling. If, however, the verbatim representation was not precisely encoded, was lost, or not accessed, then we might expect to see themes recalled more accurately than predicates and policy claims recalled more accurately than value claims.

**Method**

**Participants.** Twenty-nine undergraduates at Northern Illinois University participated in this experiment and received course credit for participating. All participants were native English speakers.
**Materials and design.** Twenty-four simple arguments were constructed. These arguments consisted of a claim followed by a reason. Two claim types, *policy* and *value*, were created for each stem. For example, Argument 7a employs a policy predicate, “should be federally mandated,” whereas Argument 7b employs a value predicate, “is beneficial”:

7a. *Policy argument:* Recycling should be federally mandated because it helps to protect the environment.

7b. *Value argument:* Recycling is very beneficial because it helps to protect the environment.

Reasons were selected that could support both the policy and value claims for each theme. Neither the theme nor predicate was repeated in the reason. The complete set of materials is presented in the Appendix. To equate for recency of mention, one half of the items mentioned the predicate prior to the theme in the claim, whereas the other half had the theme mentioned prior to the predicate.

The design manipulated *claim type* (policy vs. value) as a within-subjects factor. The two versions of each target argument were equated on the average number of characters per argument (*M* = 113.5, *SD* = 19.7 and *M* = 113.3, *SD* = 23.4 for policy and value versions, respectively). Items were assigned to one of two lists based on a Latin-square. Participants received each item in only one claim-type condition. The 24 arguments were presented in a different random order for each participant.

**Procedure.** Participants were run in individual rooms on a computer using E-prime software (Psychology Software Tools, Pittsburgh, PA). They began with 4 practice items to familiarize themselves with the procedure prior to receiving the 24 target items (12 policy and 12 value arguments). Participants read each argument and rated their agreement with it on a 6-point scale ranging from 1 (*strongly agree*) to 6 (*strongly disagree*). Then they pressed the spacebar and the argument was removed. They were then asked to type in the argument that they had just rated. Participants were instructed to use “as close to the exact words” as they could. They were instructed to read each argument at a normal rate of speed and not to spend additional time memorizing the arguments.

**Results and Discussion**

Recall responses were scored for both the accuracy of their theme and predicate. The theme was scored as the topic of the argument, such as “recycling.” The predicate was scored for the form of the verb. *Verbatim scoring* required an exact match with the target item. For example 7a, the participant would have to say “recycling” to receive verbatim credit for the theme and “should mandate” to get verbatim
credit for the predicate. Textbase scoring accepted both an exact match and close synonyms. For example, a participant including either “recycling” or the synonym “reuse” in their recall for 7a or 7b would receive textbase credit for the theme. Correct textbase recall for the policy predicate in 7a would include either “should mandate” or the synonym “should require.” Correct recall for the value predicate in 7b would include either “is beneficial” or the synonym “is valuable.” All other responses, such as “is necessary,” “is good,” “should be encouraged,” and “should be done” were scored as errors. There was high interrater reliability (agreement = 95%) for the textbase scoring between the judgments of 2 independent raters who were blind with respect to condition. Verbatim scoring was accomplished using a string matching function that first accommodated misspellings.

All analyses were conducted on the proportion correct recall. The average proportion recall for claim element (theme vs. predicate) for each claim type (policy vs. value) is shown in Table 1. Using the textbase criterion, themes ($M = .946$) were recalled better than predicates ($M = .754$), $F(1, 28) = 42.56$, $MSE = .025$, $p < .001$. Themes ($M = .851$) were also recalled better than predicates ($M = .634$) in the verbatim analysis, $F(1, 28) = 58.74$, $MSE = .023$, $p < .001$. Participants clearly had more difficulty recalling the precise claim predicate than they did recalling the theme. For the textbase analysis, policy predicates ($M = .83$) were recalled more accurately than value predicates ($M = .68$), $F(1, 28) = 19.79$, $MSE = .009$, $p < .001$. For the verbatim criterion, the same pattern was found: Policy predicates ($M = .68$) were recalled more accurately than value predicates ($M = .59$), $F(1, 28) = 5.612$, $MSE = .013$, $p < .025$. The interaction of claim element and claim type was significant only for the textbase analysis, $F(1, 28) = 19.81$, $MSE = .008$, $p < .001$. Inspect-

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th>Predicate</th>
<th>Theme</th>
<th>Predicate</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>All participants</td>
<td>Policy</td>
<td>.830 (.17)</td>
<td>.950 (.08)</td>
<td>.675 (.19)</td>
</tr>
<tr>
<td>Value</td>
<td>.676 (.28)</td>
<td>.943 (.08)</td>
<td>.592 (.28)</td>
<td>.842 (.11)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment 3</th>
<th>Predicate</th>
<th>Theme</th>
<th>Predicate</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less-skilled readers</td>
<td>Policy</td>
<td>.838 (.18)</td>
<td>.950 (.08)</td>
<td>.692 (.21)</td>
</tr>
<tr>
<td>Value</td>
<td>.667 (.18)</td>
<td>.971 (.07)</td>
<td>.553 (.19)</td>
<td>.861 (.14)</td>
</tr>
<tr>
<td>Mean</td>
<td>.753 (.16)</td>
<td>.961 (.07)</td>
<td>.623 (.19)</td>
<td>.850 (.10)</td>
</tr>
</tbody>
</table>

| Skilled readers | Policy | .942 (.07) | .991 (.03) | .807 (.15) | .905 (.08) |
| Value | .774 (.14) | .996 (.07) | .697 (.16) | .909 (.08) |
| Mean | .858 (.08) | .994 (.01) | .752 (.14) | .907 (.06) |
tion of the means shows that the interaction is due to a policy-value difference for predicates but not for themes. The presence of the interaction does not negate the general conclusion that predicates are recalled more poorly than themes. Bonferroni post-hoc tests \( (p < .05) \) found that predicates were recalled worse than themes for both policy items, \( t(28) = 5.72, p < .01 \); and value items, \( t(28) = 6.14, p < .01 \), for the textbase analysis. The same was true for the verbatim analysis. Predicates were recalled worse than themes for both policy items, \( t(28) = 6.03, p < .01 \); and value items, \( t(28) = 6.17, p < .01 \). This differential memory for the claim elements in addition to the overall level of accuracy suggests a lack of precision in the representation of claim predicates.

An examination of the recall errors found that most errors were due to mis-recalling the predicate rather than omissions. Participants tended to subtly, but significantly, change the meaning of the claim. These subtle changes can have an important influence on the meaning as demonstrated in the subtle distinctions between the factual (e.g., 8a and 9a), value (8b and 9b), and policy claims (e.g., 8c and 9c):

8a. \textit{Factual}: The U.S. has a right to intervene in other countries’ affairs.
8b. \textit{Value}: The U.S. is right to intervene in other countries’ affairs.
8c. \textit{Policy}: The U.S. should intervene in other countries’ affairs.
9a. \textit{Factual}: Car insurance companies charge higher rates to those under the age of 25.
9b. \textit{Value}: Car insurance companies are justified in charging higher rates to those under the age of 25.
9c. \textit{Policy}: Car insurance companies should charge higher rates to those under the age of 25.

The incidence of recall errors is shown in Table 2. Errors were divided into the following categories: recalling a different policy or value than that presented, changing the form of the predicate to a factual claim, or recalling a more general predicate (i.e., “is good” or “is bad”). For the policy claims, participants generally mis-recalled the predicate as an alternative policy claim (86%). Recall for the value claims was more varied. Although the highest category was an alternative value predicate (35%), there were also many policy, general, and factual errors. These errors seem to suggest that participants have a strong sense of a policy after reading policy claims but value claims are less stable as a category.

The results of this experiment show that participants made a substantial number of errors when recalling a claim that they had just read and to which they had just stated their agreement or disagreement. Furthermore, the comparison with the near ceiling level of recall of another claim element, the theme, indicates that the recall problem is specific to the claim predicate. An examination of the participants’ errors showed that they frequently changed the meaning of the predicate in the short
interval between reading and recalling. This lack of precision in predicate recall is important because at the time of making an agreement judgment, the predicate is a crucial element for evaluating the argument.

There are several possible explanations for this relatively poor recall of predicates. First, one may wonder whether belief consistency had an impact on precision of predicate recall. Wiley (2005) found that less-knowledgeable participants tend to recall significantly more arguments for their side of a controversy than arguments for the opposing side. Wiley argued that without knowledge, one relies on one’s position on the topic to guide recall, and this may lead to biased recall. Perhaps participants in our experiment were not attending to or not encoding the precise predicate, relying instead on heuristic processing in general (Chaiken, 1980; Chaiken, Giner-Sorolla, & Chen, 1996) such as recalling their own position on the argument. Alternatively, participants might have used systematic processing only for information with which they disagreed (Eagly, Kulesa, Chen, & Chaiken, 2001). They may process arguments they disagreed with more precisely and therefore have relatively better recall for those items than for the arguments with which they agreed.

Although these experiments were not specifically designed to examine this question, a post-hoc analysis looked at agreement and claim type. Items that a participant stated they agreed with (rated a 1–3) were classified as Participant Agreement, whereas those items that a participant stated they disagreed with (rated a 4–6) were classified as Participant Disagreement. The proportion recall was then recalculated for each participant, and a 2 Agreement (agree vs. disagree) x 2 Claim Element (predicate vs. theme) analysis of variance (ANOVA) was conducted on

<table>
<thead>
<tr>
<th>Type of Error</th>
<th>Policy</th>
<th>Value</th>
<th>Factual</th>
<th>Good/Bad</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presented as a Policy</td>
<td>86%</td>
<td>4%</td>
<td>0%</td>
<td>3%</td>
<td>2%</td>
</tr>
<tr>
<td>Presented as a Value</td>
<td>23%</td>
<td>35%</td>
<td>18%</td>
<td>23%</td>
<td>2%</td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presented as a Policy</td>
<td>92%</td>
<td>4%</td>
<td>0%</td>
<td>0%</td>
<td>4%</td>
</tr>
<tr>
<td>Presented as a Value</td>
<td>5%</td>
<td>57%</td>
<td>11%</td>
<td>16%</td>
<td>11%</td>
</tr>
<tr>
<td>Experiment 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less-skilled readers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presented as a Policy</td>
<td>86%</td>
<td>4%</td>
<td>4%</td>
<td>3%</td>
<td>4%</td>
</tr>
<tr>
<td>Presented as a Value</td>
<td>13%</td>
<td>47%</td>
<td>13%</td>
<td>23%</td>
<td>4%</td>
</tr>
<tr>
<td>Skilled readers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presented as a Policy</td>
<td>93%</td>
<td>5%</td>
<td>0%</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Presented as a Value</td>
<td>13%</td>
<td>48%</td>
<td>15%</td>
<td>22%</td>
<td>2%</td>
</tr>
</tbody>
</table>

TABLE 2
Average Percentage of Each Type of Recall Error for Policy and Value Claims for all Experiments

Downloaded By: [Northern Illinois University] At: 21:33 7 April 2010
the verbatim recalls.\(^1\) The only significant effect was Claim Element, \(F(1, 28) = 44.44, \text{MSE} = .036, p < .001\). No other effects were significant. In fact, simple \(t\) tests comparing theme and predicate recall for each agreement group were significant; showing poorer recall of predicates for items with which participants agreed (predicate recall: \(M = .669, S = 21.7\); theme recall: \(M = .855, S = 10.9, t(28) = 5.61, p < .001\)) and disagreed (predicate recall: \(M = .589, S = 31.3\); theme recall: \(M = .876, S = 14.8, t(28) = 4.78, p < .001\)).

A second factor that might have influenced recall is negation. Negative statements are known to be more difficult to process than positive statements. For instance, participants have more difficulty verifying negative information (Clark & Chase, 1972; Wason & Jones, 1963), have slower response times to negative statements (Millis & Just, 1994), and often recall the negative form with an affirmative synonym (Brewer & Lichtenstein, 1975). Again, the materials were not counter-balanced on negation, but there were a few negative items in each list (4 of 24 in one list and 2 of 24 in the other list). A re-analysis was performed after removing the negative items and recalculating the proportion correct for each subject on only the remaining affirmative items (20 in the first list and 22 in the second list). It found a significant effect of claim element, \(F(1, 28) = 38.56, \text{MSE} = .026, p < .001\); Claim Type, \(F(1, 28) = 10.10, \text{MSE} = .017, p < .001\); and a significant Claim Element × Claim Type interaction, \(F(1, 28) = 8.08, \text{MSE} = .018, p < .001\). Most important, Bonferroni post-hoc tests show that the predicate–theme recall difference was significant for both policy claims, \(t(28) = 3.32, p < .01\); and value claims, \(t(28) = 6.04, p < .001\). Therefore, the relatively poor recall of claim predicates occurs even when only affirmative items are presented.

Although this experiment addresses the question of how precisely readers represent claim elements after making agreement judgments, it raises the question of how soon readers lose (or fail to access) the verbatim representation of the predicate. According to text research, the surface code should still be active under conditions of immediate recall (Kintsch & van Dijk, 1978). In this experiment, recall was assessed under naturalistic conditions. Participants read the entire argument, made a judgment and then were asked to recall the claim they had just evaluated. This sequence may encourage more situation model processing. The final two experiments examine the accuracy of recall under more immediate conditions to test recall under a situation that encourages a reliance on verbatim processing. One would expect that predicate recall would be higher if recall were assessed immediately after reading the claim. To test this, Experiment 2 presented the argument in a reason–claim order and placed the recall task prior to the judgment task.

Experiment 1 also compared recall accuracy for the predicate with that of the theme. It is possible that people are simply much worse at recalling verbs or predi-

\(^1\)The focus of follow-up analyses are on the verbatim analysis because it is the most stringent criterion and most apt to show subtle differences.
cate adjectives than nouns and noun phrases. Experiment 2 addresses this concern by comparing recall of claim predicates with recall of narrative predicates.

**EXPERIMENT 2**

Experiment 2 reduced the delay between reading the claim and recalling it to allow for a more immediate assessment of whether readers maintain a precise representation of the argument. This was done by prompting recall immediately after reading the argument and by presenting arguments in a reason–claim order (10a and 10b) rather than in a claim–reason order:

10a. *Policy Full-argument*: Because recycling helps to protect the environment, recycling should be federally mandated.
10b. *Value Full-argument*: Because recycling helps to protect the environment, recycling is beneficial.

Presenting the claim last and immediately before recall provides a stronger test for whether a surface code or textbase representation includes the precise argument predicate.

Experiment 2 also examines whether any deficit in predicate recall was due to a problem of motivation or general memory for verbs. This was done by comparing recall of claim predicates with narrative predicates. Much of the early text processing research that observed a verbatim–gist distinction used narrative predicates. In an attempt to create a distinction for narratives that resembles the policy–value distinction for arguments we made narrative items whose predicate was either a high frequency event or a lower frequency synonym. For example, item 11a uses the high frequency verb “took” where its matched item 11b uses the lower frequency verb “snatched”:

11a. Since she had a horrible craving for sweets, she took a handful of candies from the jar in the entranceway.
11b. Since she had a horrible craving for sweets, she snatched a handful of candies from the jar in the entranceway.

We refer to items such as 11a as having a general event verb, and items such as 11b as having a specific event verb. This was done to eliminate any benefits of guessing a more general term. In addition to the event items, participants also received an equal number of arguments, evenly split between policy and value claims. All target elements were presented last (i.e., explanation–event and reason–claim).

Given the immediacy of recall, accuracy for event predicates should be at ceiling if participants are motivated and understand the task. If recall, however, is
equal for both event and argument predicates, then the poor predicate recall observed in Experiment 1 would be explained by a general difficulty recalling predicates. If recall accuracy is high for event predicates, then any difference for claim predicates would indicate that our participants did not have general memory or motivation problems or a problem with recalling predicates compared to nouns.

Method

Participants. Twenty-eight native English-speaking undergraduates at Northern Illinois University received course credit for participating in this experiment.

Materials. Sixteen event items were created to parallel the form of arguments from Experiment 1. To make sure participants could not guess the predicate, two versions of event predicates were used: a general event predicate (e.g., 11a) and a more specific event predicate (e.g., 11b). The two predicate types differed with respect to their frequency, with general predicates having a higher frequency ($M = 740.4, S = 714$) than the specific predicates ($M = 92.6, S = 120$). Word frequency was determined from the MRC Psycholinguistic Database (Wilson, 1988). For the 16 argument items, the same policy and value versions were used as in Experiment 1.

Design and procedure. Participants received 16 arguments from Experiment 1—one half were policy versions, and one half were value versions. In contrast to Experiment 1, these items were presented in a Reason–claim order to determine recall accuracy immediately following reading of the claim. They also received 16 explanations with event predicates—one half were general verbs, and one half were specific verbs. In addition to the target items, participants received 8 practice and 16 filler items. One half of these items were events, and one half were arguments. All of the filler items and one half of the practice items were created to be inconsistent to ensure a significant number of “no” responses for the judgment task. The 16 fillers and 32 target items were presented to participants in a different random order.

Participants read the first phrase (i.e., explanation or reason) and pressed the spacebar. The first phrase was replaced with the second phrase (i.e., event or claim). Immediately after pressing the space bar to remove the second phrase, participants were prompted to type in the most recently presented phrase (i.e., event or claim) in “as close to the exact words” as they could. Then they rated the consistency of the two phrases. Because it does not make sense to make an agreement judgment for events, participants were asked to make a consistency judgment after recall. They were asked to rate how well the two phrases “fit together” on a 6-point scale ranging from 1 (very low) to 6 (very high).
Results and Discussion

Responses were scored for the accuracy of their predicate recall. The verbatim analysis required an exact match between the target item and participant recall. The textbase analysis also allowed synonyms. There was high inter-rater reliability (agreement = 94%) for the textbase scoring between the judgments of the two independent raters blind to condition. Overall, accuracy of predicate recall was high for both the verbatim \( (M = .92) \) and the textbase scoring \( (M = .94) \) as shown in Table 3.

The analysis was conducted on the proportion of correct recall. There was a significant effect of predicate Genre in that event predicates \( (M = .987, S = .05) \) were recalled better than claim predicates \( (M = .844, S = .13) \) for the verbatim criterion, \( t(27) = 7.75, p < .001 \). The same pattern was found for the textbase criterion: Event predicates \( (M = .989, S = .04) \) were recalled better than claim predicates \( (M = .888, S = .12) \), \( t(27) = 5.77, p < .001 \). Recall for each predicate group was compared to perfect recall. The proportion of event predicates correctly recalled was not significantly different from 100% for either the verbatim criterion, \( t(27) = 1.44, p = .161 \); or the textbase criterion, \( t(27) = 1.54, p = .134 \). The proportion of argument predicates correctly recalled, however, was significantly different from 1 for both the verbatim criterion, \( t(27) = 6.52, p < .001 \); and the textbase criterion, \( t(27) = 4.99, p < .001 \). These results suggest that participants did not have general memory or motivation problems nor was there a problem with recalling verbs in general. An analysis of the effect of predicate type for argument claims found that policy predicate claims were recalled better than value claims for both the verbatim criterion, \( t(27) = 2.46, p < .025 \); and the textbase criterion, \( t(27) = 3.95, p < .001 \). This replicates the results from Experiment 1; participants have difficulty recalling argument claims and value claims are particularly difficult.

**TABLE 3**

<table>
<thead>
<tr>
<th>Predicate Type</th>
<th>General M</th>
<th>General SD</th>
<th>Specific M</th>
<th>Specific SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event predicates</td>
<td>.987 .04</td>
<td>.991 .05</td>
<td>.987 .04</td>
<td>.987 .07</td>
</tr>
<tr>
<td>Argument predicates</td>
<td>.946 .09</td>
<td>.830 .18</td>
<td>.888 .12</td>
<td>.799 .19</td>
</tr>
</tbody>
</table>
An analysis of the errors (shown in Table 2) shows that policy items are most frequently recalled as a different policy claim (92%), whereas value arguments are mis-recalled in many different forms especially in a general “good/bad” gist form. It is interesting to note that when recall is this immediate and the items are interspersed with explanations, participants generally do not erroneously recall value claims as policy claims.

The results of this experiment are somewhat remarkable. Our participants were only 84% accurate at immediately recalling argument predicates of a claim they had just read with no intervening task or reason presented. In contrast, they were 99% accurate when a phrase had a narrative form. These results may be due to either a poorly encoded initial representation as a result of heuristic or superficial processing or to a problem with retrieving the verbatim representation. According to Voss et al. (1993), upon the reading of a claim, other beliefs and knowledge become activated. These other propositions may influence the situation model or gist representation of the claim. Because of the end-of-sentence wrap-up (Just & Carpenter, 1980) that occurs at the end of the claim, it may be that this representation is accessed at recall rather than the surface code despite the immediacy of the recall. This retrieval of a gist representation during evaluation would be consistent with Reyna and Brainerd’s (1995) findings. These experiments do not rule out that a verbatim representation is available, just that it is not consulted for recall.

The results of Experiment 2 show that the poor recall of claim predicates that we observed in Experiment 1 was not due to general memory problems or the motivation level of our participants; nor was it due to a general difficulty recalling predicates. Instead, these results further point to a specific problem with the representation of claim predicates. Experiment 3 examines whether this problem might be a result of poor general reading ability and whether this lack of precision has implications for one’s ability to detect argument flaws.

**EXPERIMENT 3**

Experiment 3 addresses whether reading skill plays a role in the precise representation of arguments. Hannon and Daneman (2004) recently found that reading skill is a factor in whether readers rely on a shallow or non-verbatim representation when reading texts containing low-level anomalies. Less-skilled readers are more likely to engage in shallow reading than skilled readers. For instance, Barton and Sanford (1993) gave participants short questionnaires that included an item containing a logical inconsistency as in 12a and 12b (one does not bury survivors):

12a. When an aircraft crashes, where should the survivors be buried?
12b. When a bicycle accident occurs, where should the survivors be buried?
For item 12a, there is high semantic overlap between the context (air crashes) and the main action (burying), which should smoothly satisfy the readers’ need for global coherence. Item 12b, in contrast, presents a context (bike accidents) that does not generally imply death and this low semantic overlap should reduce global coherence. Barton and Sanford found that 67% of readers did not detect the conflict of burying survivors in sentence 12a, whereas only 20% failed to detect the conflict in sentence 12b. They argued that when global coherence is satisfied, readers conduct a more shallow analysis of the sentence. This shallow processing leads to the failure to detect the anomaly. Hannon and Danemen have since replicated this result and found a stronger incidence of shallow processing for poor readers than for good readers. Thus, reading ability may influence the degree to which readers can maintain the verbatim representation or initially acquire such a representation (i.e., lax initial encoding). Experiment 3 tests whether poor claim predicate recall occurs for both skilled and less-skilled readers.

The second purpose of Experiment 3 is to examine whether those who are able to recall predicates are also better skilled at distinguishing warranted from unwarranted arguments. The poor recall for the claim’s predicate found in the first two experiments is troublesome because the predicate determines the set of reasons that can support a given claim. If readers do not have a good representation of the predicate when reading the reason, then they would likely have difficulty determining whether the claim is warranted with respect to a given reason. For example, the same claim may be warranted by one reason (as in Argument 13a) but unwarranted (flawed) when supported by another reason (as in Argument 13b):

13a. **Warranted**: Recycling is beneficial because recycling helps to protect the environment.
13b. **Unwarranted**: Recycling is beneficial because most companies are now equipped for recycling.

Thus, a reader who does not accurately represent the specific predicate while attending to the claims will not be able to distinguish the difference between Arguments 13a and 13b.

In Experiment 3, participants were given a set of arguments and asked to recall their claims as in Experiments 1 and 2. Then they were given a different set of warranted and unwarranted arguments and asked to judge whether they were flawed or acceptable. It was expected that participants who were able to recall the specific predicate would be better at distinguishing flawed arguments from acceptable arguments.

**Method**

**Participants.** Eighty-seven undergraduates at Northern Illinois University participated in this experiment and received course credit for participating. All were native English speakers.
Materials and design. A single list of the 24 claim–reason arguments from Experiment 1 was used to so that all participants’ performance could be directly compared. One half of the items had policy claims, and one half had value claims; however, items were not crossed in this experiment.

A flawed-argument judgment task was created to test participants’ ability to distinguish warranted arguments from unwarranted arguments. Twenty-four target arguments were created: 12 unwarranted arguments where the reason did not support the claim and 12 warranted arguments where a reason was presented that could support the specific predicate of the claim. There were also 12 foil items, circular arguments, and arguments with untrue reasons to ensure that participants used the flawed side of the scale. None of the themes overlapped with the themes from the recall test. The test was presented to participants in a booklet with items in a single random order.

Reading skill was determined using the Nelson–Denny test (Brown, Fishco, & Hanna, 1993). Most of the 38 items on this test tap comprehension at the textbase level (Magliano, Millis, Ozuru, & McNamara, 2003). A median split on the Nelson–Denny test of reading comprehension was used to group these participants into 36 less-skilled readers (raw score of below 28) and 44 skilled readers (raw score of above 28). Participants with scores at the median were excluded from the analyses (n = 7).

Procedure. Participants read the complete arguments in a claim–reason order as in Experiment 1. Upon pressing the spacebar, the argument was removed and participants were prompted to recall the claim in “as close to the exact words” as they could. Finally, they were asked to make a forced-choice agreement judgment (i.e., “Do you agree with the argument you just read? Y/N”). After completing all the argument recall trials, participants were given the flawed-argument judgment task. In this task, participants were given 36 new unrelated arguments in a short booklet. They were asked to “mark each argument that is logically flawed regardless of whether you agree with it or not. You want to consider only the structure of the argument not how convincing it is given everything you know.” Finally, they were given 20 min to complete the Nelson–Denny test of comprehension.

Results and Discussion

The average textbase and verbatim recall is shown in the bottom of Table 1 for each Claim Element (predicate and theme), each Claim Type (policy and value), and for each Reading Skill level (less-skilled and skilled). A mixed ANOVA with Claim Element and Claim Type as within-subjects factors and Reading Skill as a between-subject factor found a significant effect of Claim Element for both the verbatim scoring, \( F(1, 78) = 154.00, MSE = .019, p < .001 \); and the textbase scoring, \( F(1, 78) = 183.86, MSE = .013, p < .001 \). Participants recalled themes of claims (\( M = .878 \) and \( M = .977 \) for verbatim and textbase, respectively) better than predicates
There was a significant main effect of Claim Type for both the verbatim scoring, $F(1, 78) = 27.82$, $MSE = .009$, $p < .001$; and the textbase scoring $F(1, 78) = 68.18$, $MSE = .007$, $p < .001$. Participants recalled elements of policy claims ($M = .811$ and $M = .930$ for verbatim and textbase, respectively) better than value claims ($M = .755$ and $M = .852$ for verbatim and textbase, respectively). There was also a significant effect of Reading Skill for both the verbatim scoring, $F(1, 78) = 14.76$, $MSE = .047$, $p < .001$; and the textbase scoring, $F(1, 78) = 17.69$, $MSE = .021$, $p < .001$. Overall, Skilled Readers recalled elements of the claim ($M = .830$ and $M = .926$ for verbatim and textbase, respectively) better than less-skilled readers ($M = .736$ and $M = .857$ for verbatim and textbase, respectively). There was a significant Claim Element $\times$ Claim Type interaction for both verbatim scoring, $F(1, 78) = 41.83$, $MSE = .009$, $p < .001$; and the textbase scoring, $F(1, 78) = 106.44$, $MSE = .006$, $p < .001$. The difference between predicate and theme recall was larger for the value claims. The critical comparison, however, is whether there exists a difference in recall accuracy for claim predicates and themes. Bonferroni post-hoc tests ($p < .05$) on both the verbatim and textbase analyses revealed that themes were recalled better than predicates for both policy and value claims.

Finally, there was a significant Claim Element $\times$ Reading Skill interaction (for both verbatim scoring, $F(1, 78) = 5.40$, $MSE = .019$, $p < .05$; and the textbase scoring, $F(1, 78) = 8.07$, $MSE = .013$, $p < .01$). The difference between predicate and theme recall was smaller for skilled readers than less-skilled readers. There are two relevant simple effects that can be conducted on this Claim Element $\times$ Reading Skill interaction. The first method is to compare the skill difference for each element type. Bonferroni post-hoc tests ($p < .05$) on both the verbatim and textbase analyses revealed that skilled readers recalled predicates ($M = .752$ and $M = .858$ for verbatim and textbase, respectively) significantly better than less-skilled readers ($M = .623$ and $M = .753$). Skilled readers also recall themes ($M = .907$ and $M = .994$ for verbatim and textbase, respectively) better than less-skilled readers ($M = .850$ and $M = .961$). Thus, better readers have better memory for claim predicates and themes. The second method is to compare the element effect for each skill type to determine whether the theme–predicate difference occurs with both skilled and less-skilled readers. Bonferroni post-hoc tests ($p < .05$) revealed that themes ($M = .850$ and $M = .961$ for verbatim and textbase, respectively) were recalled better than predicates for less-skilled ($M = .623$ and $M = .753$). The same pattern was found for skilled readers: Themes ($M = .907$ and $M = .994$ for verbatim and textbase, respectively) were recalled better than predicates ($M = .752$ and $M = .858$). Thus, skill in reading comprehension increases one’s general recall of the text and reduced the predicate–theme recall difference, but it does not eliminate the relatively poorer recall of claim predicates. Because this difference occurs for both skilled and less-skilled readers, the effect is most likely not entirely a result of superficial processing but more a reliance on a gist level of representation.
The pattern of recall errors for both skilled and less-skilled readers is shown in Table 2. Again, policy items were generally recalled as a different policy claim (85%). The value arguments, when recalled incorrectly, were recalled as a different value predicate or in a general “good/bad” gist form. As in the errors for Experiments 1 and 2, the policy form is a salient form, whereas the value form is less salient.

Experiment 3 was also conducted to determine whether the accuracy of predicate memory is related to one’s skill at detecting flawed arguments. A tertiary split on the predicate recall data identified those who were skilled at recalling the predicate (n = 28, M = .851, SD = .06 and n = 28, M = .914, SD = .042 for verbatim and textbase, respectively) from those who were less skilled in predicate recall (n = 29 M = .499, SD = .094 and n = 24 M = .627, SD = .081 for verbatim and textbase, respectively). Table 4 shows the average proportion judgment accuracy for each Argument type (warranted vs unwarranted) and each predicate-recall skill (less-skilled and skilled) for both the Textbase and Verbatim scoring method. A mixed analysis of covariance with Argument type as a within-subjects factor and Nelson–Denny as a covariate to control for any effects on evaluation judgments due to reading skill alone. This analysis revealed a significant effect of Predicate-recall Skill for the verbatim criterion, $F(1, 54) = 4.74, MSE = .051, p < .05$; but only a marginal effect for the textbase criterion, $F(1, 49) = 2.93, MSE = .052, p < .02$.

We also conducted a regression analysis using the verbatim recall data to assess the impact of recall skill on flawed judgment detection. The criterion variable was the judgment score averaged across good and unwarranted judgments for each participant. We entered the Nelson–Denny score in the first step to control for reading skill. The average recall scores for the predicate were then entered in the second step. The regression equation significantly predicted flawed judgment performance, $F(2, 78) = 6.17, MSE = .023, p < .01, R^2 = .137$. Predicate recall scores accounted for a significant change in $R^2$ after controlling for Nelson–Denny, $F(1, 78) = 6.01, p < .02, R^2 = .066$. The $B$ for Nelson–Denny in the full model was .145 ($t = 1.45, p = .216$), and the $B$ for predicate recall in the final model was .284 ($t = 2.45, p < .02$). Thus, predicate recall ability significantly predicted flawed judgment performance over and above the effect of reading skill.

<table>
<thead>
<tr>
<th></th>
<th>Textbase</th>
<th></th>
<th>Verbatim</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Warranted</td>
<td>Unwarranted</td>
<td>Warranted</td>
<td>Unwarranted</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Less-skilled recallers</td>
<td>.573</td>
<td>.18</td>
<td>.562</td>
<td>.17</td>
</tr>
<tr>
<td>Skilled recallers</td>
<td>.719</td>
<td>.20</td>
<td>.691</td>
<td>.22</td>
</tr>
<tr>
<td>Skilled recallers</td>
<td>.721</td>
<td>.21</td>
<td>.691</td>
<td>.22</td>
</tr>
</tbody>
</table>
=.093. As predicted, those skilled in recalling the predicate in its verbatim form were more accurate at judging warranted arguments as “good” and unwarranted arguments as “flawed.” The covariate also approached significance, $F(1, 54) = 3.21$, $MSE = .051$, $p = .079$. No other effects were significant (all $F$s < 1). Thus, recall ability predicted evaluation skill controlling for reading ability. There were no differences in judgment performance for the foil items between the different recall skill levels (verbatim: $F < 1$, less-skilled $M = .771$, skilled $M = .784$; textbase: $F < 1$, less-skilled $M = .754$, skilled $M = .787$).

The results of this experiment replicate the relatively poor recall of claim predicates found in Experiments 1 and 2 and show that it occurs with both skilled and less-skilled readers, although the difference is less for skilled readers. Experiment 3 also shows that precise predicate recall matters for argument evaluation. Participants who made more errors recalling claim predicates also made more errors rejecting unwarranted arguments.

As noted in the discussion of Experiment 1, it is possible that whether a person agrees with an argument affects their memory for that argument. To investigate this possibility, we created an agreement variable using the agreement judgments provided by the participants. The items with which each participant agreed were coded as agree items and the items with which each participant disagreed were coded as disagree items. Table 5 presents the means for all of the conditions. We conducted a 2 Agreement (agree vs. disagree) x 2 Claim Element (predicate vs. theme) x 2 Claim Type (policy vs. value) x 2 Reading Skill (skilled vs. less-skilled) ANOVA with Agreement as a new within-subjects factor on the mean proportion recall for each subject. We conducted this analysis on the verbatim recall data, and

<table>
<thead>
<tr>
<th>Participant Agreement Decision</th>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Predicate</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Theme</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reading skill</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
</tr>
<tr>
<td>MS</td>
<td>DM</td>
<td>S</td>
</tr>
<tr>
<td>Less-skilled readers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Policy</td>
<td>.687</td>
<td>.26</td>
</tr>
<tr>
<td>Value</td>
<td>.508</td>
<td>.23</td>
</tr>
<tr>
<td>Skilled readers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Policy</td>
<td>.816</td>
<td>.18</td>
</tr>
<tr>
<td>Value</td>
<td>.666</td>
<td>.20</td>
</tr>
<tr>
<td>TABLE 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average Proportion Recall as a Function of Claim Element (Predicate vs. Theme), Claim Type (Policy vs. Value), Reading Skill (Less-Skilled vs. Skilled Readers), and Participant Agreement</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
we only report the effects of Agreement. There was a significant Claim Element \times Agreement interaction, \(F(1, 78) = 8.01, MSE = .029, p < .01\). The predicate–theme difference was larger for the agree items than disagree items; however, Bonferroni post-hoc tests (\(p < .05\)) revealed that recall was better for themes than predicates for both the agree (\(M = .669\) and \(M = .894\) for predicates and themes, respectively) and disagree items (\(M = .711\) and \(M = .859\) for predicates and themes, respectively). There was a significant Claim Type \times Agreement interaction, \(F(1, 78) = 8.48, MSE = .029, p < .01\). Bonferroni post-hoc tests (\(p < .05\)) revealed that recall was better for policy (\(M = .825\)) than value (\(M = .739\)) claims for agree items, but there was no difference between policy (\(M = .789\)) and value (\(M = .782\)) claims for disagree items. From both of these interactions it appears that agreeing with an argument has a larger impact on memory than disagreeing with an argument—that is, differences between claim elements and claim types are exacerbated by agreement rather than disagreement. Finally, there was a significant three-way interaction among Agreement, Claim Type, and Reading Skill: \(F(1, 78) = 4.07, MSE = .029, p < .05\). To understand this interaction, a 2 Agreement (Agree vs. Disagree) \times 2 Claim Type (Policy vs. Value) ANOVA was conducted separately for each Reading Skill group. The ANOVAs revealed that the Claim Type \times Agreement interaction was only significant for the less-skilled group: \(F(1, 35) = 9.29, MSE = .035, p < .01 (F < 1 \text{ for the skilled group})\). Bonferroni post-hoc tests (\(p < .05\)) for the less-skilled group means revealed the same pattern as reported earlier—that is, recall was better for policy (\(M = .780\)) than value (\(M = .684\)) claims for agree items, but there was no difference between policy (\(M = .708\)) and value (\(M = .746\)) for disagree items. This suggests that the effect of Agreement has a stronger impact on the Claim Type memory of less-skilled readers. These results are consistent with those of Experiment 1. Although participants’ opinion did affect the magnitude of the predicate–theme recall difference, it did not eliminate it. Therefore, it is not the case that participants rely on a gist representation only when they read attitude consistent claims.

**GENERAL DISCUSSION**

These three experiments show that readers do not always rely on a precise representation of an argument’s claim when reading persuasive text. By the time our participants had finished reading a simple argument claim followed by a reason, they were only 63% accurate at recalling the precise predicate compared to 85% accurate at recalling the precise theme they had just read. Even under more immediate recall conditions (reason–claim order and immediate recall), participants still recalled the predicate at a lower rate (\(M = .80\)) compared to the theme (\(M = .89\)). This predicate–theme difference was also true for both less-skilled readers (\(M = .62 \text{ vs. } M = .85\)) and skilled readers (\(M = .75 \text{ vs. } M = .91\)) and when participants were constructing a representation to
make an agreement judgment—a task that should require a careful analysis of the claim’s predicate. It is somewhat disturbing to realize that people might make such judgments based on only a gist representation of the argument.

The experiments ruled out several possible alternative explanations for the relatively poor predicate recall. The difficulty with predicate recall was not due to participants’ poor general memory or lack of motivation. Participants’ verbatim recall was much better for argument themes (M = .87) and narrative predicates (M = .98) than for argument predicates (M = .67). Thus, our participants could precisely recall the presented information and did so for some content but did not for claim predicates. The imprecise predicate representation was also not due completely to making an agreement judgment. There was a 22% predicate–theme difference for recall following Claim–Reason–Judgment items (Experiment 1) compared to a 19% difference for recall following Claim–Reason items (Experiment 3). The imprecise predicate representation was also not completely due to the delay in reading the reason because a similar pattern of inferior recall for predicates was found for immediate recall. Finally, the difficulty with predicate recall was not due to only items they agreed with or a general difficulty with negative predicates.

Experiment 3 suggests that relying on a gist representation may be responsible for at least some of the difficulty students have with evaluating written arguments. During the flawed-judgment evaluation task, participants had the arguments available so judgments were not a result of poor memory for the argument. Rather, it appears to be a problem of lack of attention to the precise wording of the claim when making this judgment. When evaluation requires attention to the precise predicate of an argument’s claim, students are fairly likely to have either mistaken it immediately upon reading it or to have converted it to something more general or consistent with their model of the sentence’s meaning. Such a conversion makes them less likely to notice that a reason is not warranted or that a series of reasons have differing logical relations to the claim.

Not everyone is affected by the recall problems we observed. Many of our participants accurately recalled the predicates. Likewise, the graduate students that Britt and Kurby (2005) observed, who did not show the same evaluation problems as undergraduates, presumably did not have these recall problems or they could not have performed so well on the evaluation task. Why are some readers more likely to mis-represent the claim of an argument than others? One possibility is that the ability to maintain the verbatim representation after end-of-sentence wrap-up is related to one’s working memory span (Just & Carpenter, 1992). Working memory span is a well-known correlate of performance on a number of reading, reasoning and aptitude tests (Copeland & Radvansky, 2004; MacDonald, Just, & Carpenter, 1992; Verschueren, Schaecken, & D’Ydewalle, 2005). It is likely to be fairly normally distributed in a sample of experiment participants so we might reasonably expect to sample some high span individuals who could easily maintain the verbatim representation and make no recall errors and some low span individuals who
would immediately lose the verbatim representation and have to rely solely on gist. It is also likely that graduate students, who are selected from the undergraduate population based on Graduate Record Exam scores and other aptitude and achievement measures, will have higher average working memory spans than undergraduates, thus accounting for the Britt and Kurby results.

Familiarity with the argument genre may also influence one’s ability to represent argument claims precisely. By the time the average student reaches college, he or she will have read countless narrative texts and expository textbooks but relatively little persuasive prose and even fewer arguments. Thus, most undergraduate readers are unlikely to have acquired genre-specific skills from the reading experiences they have received prior to college. If this were the case, it would be consistent with Perkins’s (1985) and Kuhn’s (1991) conclusion that students’ failings may be a result of lack of practice, exposure, and explicit instruction in the skill of argumentation. This increased exposure should help students more fully develop an argument-specific schema and increase the probability that it will become activated when appropriate.

Britt and Larson (2003) speculated that argument schemata could explain the phrase order effects they observed. They found that arguments were read faster and recalled better when the arguments were presented in a standard order (claim preceding the reason). They proposed that readers construct an argument schema when they encounter a claim and this schema guides the interpretation and importance of all other elements in the argument. It may be the case that some students’ lack of familiarity with the genre results in a failure to develop an argument schema that could lead to a reliance on a gist representation during claim processing. These results are consistent with recent change-detection findings. For example, Sturt, Sanford, Stewart, and Dawydiak (2004) found that participants were more likely to detect a change to a word that was focused in the discourse than one not in discourse focus. They argued that readers often settle for a “good-enough” representation, and that features that guide coherence may dictate when a more detailed representation is constructed. In the case of arguments, coherence is guided by an activated argument schema and this should require a reader to encode and maintain a detailed representation of the claim predicate. Because the predicate of an argument determines the set of reasons that can be used to support a claim, the claim predicate should be in discourse focus and should remain in focus until the reason has been connected to the claim. Thus, it would be predicted that skilled reasoners should differ from less-skilled reasoners in the reliance on a gist representation during argument processing.

In addition to the general argument schema, specific knowledge of predicate meanings may also play a role. The semantic and pragmatic constraints of a predicate limit the potential reasons that can support a claim. We refer to one’s knowledge of these constraints as a logical predicate schema. For instance, the logical predicate schema for “should ban” would include at least three elements: The exis-
tence of a problem, the ability to regulate it, and the relative quality of the proposed solution. When arguing about whether something should be banned, reasons can be presented that either support the existence of a problem (e.g., accidents from driving while talking on a cell phone) or against the existence of a problem (e.g., problem is irresponsible driving not cell phones) or that its benefits outweigh its harms (e.g., cell phone access critical in an emergency). One might also argue whether the governing body has the legal or practical ability to regulate the behavior (e.g., constitutional powers, police resources, and fairness of regulation). Finally, one could consider the quality of the proposed solution, either in terms of its potential effectiveness or its foreseeable repercussions (e.g., it will stress already burdened police resources). Readers can use their detailed knowledge of claim predicate schemas to guide their analysis of whether a reason supports a claim. We believe these predicate schemas also guide reasoners to systematically find weaknesses in an argument, to verify that an argument is warranted and to construct more complex and elaborated arguments. Participants less familiar with specific argument predicates may not have a well-developed logical predicate schema and this may result in a reliance on an over-general gist representation rather than a verbatim or textbase representation of the predicate.

A final individual difference characteristic that may impact one’s ability to represent argument information precisely is the meta-cognitive awareness that precision matters for argument comprehension. We recently examined whether people’s evaluation of arguments could be affected by training them to pay closer attention to the precise wording of the claim. Britt and Kurby (2005) manipulated the presence of a tutorial that trained participants to correctly identify claim predicates and argument flaws followed by practice with immediate feedback. Tutor and non-tutor participants did not differ in their accuracy in judging well-structured arguments, but they did differ in their accuracy in detecting flawed arguments. The non-tutor participants correctly rejected 59% of the unwarranted arguments, whereas the participants given training on identifying predicates rejected 76% of the unwarranted arguments. Thus, drawing students’ attention to the predicate of claims did lead to more accurate argument evaluation. Logically, these training recipients must also have been able to remember the predicates at the time of the task. Thus, it is conceivable that people may be able to exert additional effort or engage in some other strategy that can improve their ability to recall claim predicates.

In addition to its implications for general comprehension of argumentative texts such as newspaper editorials, this finding of poor verbatim recall of claim predicates also has implications for surveys and questionnaires. Pollsters and attitude researchers are very attuned to the precise wording of the statements they ask people in opinion surveys (e.g., “Would you say that going to war with Iraq was a ‘war of choice’ or a ‘war of necessity’?”), and with good reason. The results of these experiments, however, suggest that some participants may quickly
lose the precise wording of a question or convert it to something different. These results lead us to question what people are really agreeing to in polls and attitude questionnaires. Future research will need to be done to understand the extent of this problem and the possible repercussions of this imprecision on polling and survey data.

ACKNOWLEDGMENTS

This research was supported by a grant from the United States Department of Education’s Cognition and Student Learning Research Program (R305H050133). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of this organization.

REFERENCES


Brainerd, C. J., & Reyna, V. F. (1998). When things that were never experienced are easier to “remember” than things that were. *Psychological Science, 9*, 484–489.


APPENDIX

1. **Policy.** Universities should encourage students to travel the world because it will help students gain deeply meaningful knowledge.

2. **Value.** It is a valuable experience for students to travel the world because it will help students gain deeply meaningful knowledge.

3. **Policy.** We should stop spending billions of dollars fighting a war on drugs because we do not go after those who launder the money.

4. **Value.** We’ve wasted billions of dollars fighting a war on drugs because we do not go after those who launder the money.

5. **Policy.** Universities should randomly test athletes for steroids because they provide unfair muscle growth to their users.

6. **Value.** It is ethical for universities to randomly test athletes for steroids because they provide unfair muscle growth to their users.

7. **Policy.** Libraries shouldn’t be allowed to use filters on their internet software because they could screen out critical information.

8. **Value.** It is wrong for libraries to use filters on their internet software because they could screen out critical information.

9. **Policy.** Small towns should prohibit large chain stores from relocating near them because they often eliminate local businesses in small towns.

10. **Value.** It is a serious threat to small towns when large chain stores relocate near them because they often eliminate local businesses in small towns.

11. **Policy.** We should require forest fires to burn in a controlled way because dead underbrush puts the whole area at risk.
6. **Value.** We need to have forest fires burn in a controlled way because dead underbrush puts the whole area at risk.

7. **Policy.** Vitamins should be taken more frequently by Americans because they are necessary for healthy development and living.
7. **Value.** Vitamins are not harmful because they are necessary for healthy development and living.

8. **Policy.** Car insurance companies should charge higher rates to those under the age of 25 because younger drivers represent a greater risk of accidents.
8. **Value.** Car insurance companies are justified in charging higher rates to those under the age of 25 because younger drivers represent a greater risk of accidents.
9. **Policy.** Experimentation on animals should continue because it has led to crucial medical advances.
9. **Value.** Experimentation on animals is a useful research activity because it has led to crucial medical advances.

10. **Policy.** Hunting for sport should be permitted because otherwise some animal species would overpopulate the earth.
10. **Value.** Hunting for sport is good for society because otherwise some animal species would overpopulate the earth.

11. **Policy.** The state highway speed limit should remain 65 miles per hour because it balances safety with convenience.
11. **Value.** The state highway speed limit of 65 miles per hour is best because it balances safety with convenience.

12. **Policy.** River boat gambling should be legalized because it provides a huge source of revenue from tourism and taxes.
12. **Value.** Legalizing river boat gambling would benefit cities because it provides a huge source of revenue from tourism and taxes.

13. **Policy.** The U.S. should intervene in other countries’ affairs because local events can catastrophically impact the entire world.
13. **Value.** The U.S. is right to intervene in other countries’ affairs because local events can catastrophically impact the entire world.

14. **Policy.** Employers should provide their workers with annual vacations because people need time to spend with their families.
14. **Value.** Employers are morally responsible to provide workers with annual vacations because people need time to spend with their families.
15. **Policy.** People should drive SUVs because they are safer in an accident.
15. **Value.** It is wiser to drive SUVs because they are safer in an accident.

16. **Policy.** Companies should not be permitted to use sweatshops because they often mistreat workers in poor countries.
16. **Value.** It is improper for companies to use sweatshops because they often mistreat workers in poor countries.

17. **Policy.** People should limit their garden to plants native to the region because this will most clearly reflect the natural beauty of the area.
17. **Value.** It is preferable to garden with plants native to the region because this will most clearly reflect the natural beauty of the area.

18. **Policy.** We should ensure that children with violent temperaments be given treatment at a young age because it is easier to cure such behavior in younger children.
18. **Value.** It is important that children with violent temperaments be given treatment at a young age because it is easier to cure such behavior in younger children.

19. **Policy.** Pornography should be prohibited in our society because it promotes violence toward women.
19. **Value.** Pornography is harmful to our society because it promotes violence toward women.

20. **Policy.** Cloning should not be allowed in our society because it will be used for large scale organ harvesting.
20. **Value.** Cloning will be detrimental to society because it will be used for large scale organ harvesting.

21. **Policy.** Attending graduate school should be encouraged for most students because it enables them to find a better job.
21. **Value.** Attending graduate school is worthwhile for most students because it enables them to find a better job.

22. **Policy.** Sex education should be taught in public high-schools because it will improve students’ decisions at a critical time in their lives.
22. **Value.** Teaching sex education is appropriate in public high-schools because it will improve students’ decisions at a critical time in their lives.

23. **Policy.** Natural foods should not be preferred over processed foods because many times they are actually higher in sodium and fat.
23. *Value*. Natural foods are not always healthier than processed foods because many times they are actually higher in sodium and fat.

24. *Policy*. Recycling should be federally mandated because it helps to protect the environment.

24. *Value*. Recycling is very beneficial because it helps to protect the environment.