Measuring the Power of Learning.®

Research Report
ETS RR–17-11

Toward the Automated Scoring of Written Arguments: Developing an Innovative Approach for Annotation

Yi Song

Paul Deane

Beata Beigman Klebanov

December 2017

ETS Research Report Series

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Toward the Automated Scoring of Written Arguments: Developing an Innovative Approach for Annotation

Yi Song, Paul Deane, & Beata Beigman Klebanov

Educational Testing Service, Princeton, NJ

This project focuses on laying the foundations for automated analysis of argumentation schemes, supporting identification and classification of the arguments being made in a text, for the purpose of scoring the quality of written analyses of arguments. We developed annotation protocols for 20 argument prompts from a college-level test under the framework of the theory of argumentation schemes, which defines reasoning patterns in argumentation. The annotation protocols listed critical questions associated with each argumentation scheme, which makes the argument structure in a text explicit and classifiable. Furthermore, we annotated 200 student essays across four selected argument prompts to test if the annotation protocols can be applied reliably by human annotators. Preliminary results indicate that this method of analyzing argument structure is reliable and promising.

Keywords Argument annotation; argumentation scheme; critical question; assessment

doi:10.1002/ets2.12138

Critically evaluating arguments is an important skill in higher education and the workplace; in fact, recent educational reforms emphasize the importance of tasks that require students to demonstrate sound reasoning and use relevant evidence to support their arguments (Council of Chief State School Officers & National Governors Association, 2010). The concomitant increase in argumentative writing tasks, in both instructional and assessment contexts, has resulted in a high demand for scoring such tasks. Automating the scoring of argumentative writing would help meet this demand. Our work contributes to achieving this goal by providing high-quality human analyses of arguments that can be subsequently used to train automated argument scoring systems.

Essay scoring can be challenging to both human scorers and computers. It has been shown repeatedly that the correlation between essay length and score (either by human or by computer) is high (Attali & Burstein, 2006; Chodorow & Burstein, 2004; Kukich, 2000; Powers, 2005). Although this correlation makes sense, as students tend to put more thought into their elaborations, it could be risky if people presume that longer essays are always of better quality than shorter ones. Another issue that can make scoring argumentative writing tasks problematic is that students with good general writing skills can receive high scores from human (and automated) raters even when their arguments are not convincing. It is essential to recognize that the quality of an argument goes well beyond length, mechanics, grammar, style, vocabulary, and even structure, because arguments in a well-formulated essay may be invalid if based on false premises or may fail to take potential counterarguments into account. The claim we advance in the present work is that we need to capture the structure and content of argument, that is, to identify characteristic argument patterns that make it easier to determine whether an argument is valid and effective.

An analysis of argument structure needs to go beyond identifying basic elements, such as claims (e.g., a thesis sentence), main reasons (e.g., supporting topic sentences), evidence (e.g., elaborating segments), the introduction, and the conclusion. This is because such an analysis groups arguments into general categories, which loses information about content. Our exploration also goes beyond Toulmin's (1958) model of argumentation because Toulmin's model does not take into account the content of individual arguments and instead employs abstract categories that apply equally well to a wide variety of argument types (i.e., claim, grounds, warrant, backing, rebuttal, and qualification). We contend instead that analyzing argument structure should focus on identifying *argumentation schemes* (Walton, 1996), that is, informal reasoning patterns linked to content (how the conclusion is derived from the premises). Because argumentation schemes include schematic content, and take into account the pattern of possible argumentation moves in a larger persuasive text

Corresponding author: Y. Song, E-mail: ysong@ets.org

or dialog, recognizing argumentation schemes can support richer and deeper analyses of argument content and argument quality. In particular, the concept of *critical questions*, introduced by Walton (1996), provides a normative standard for evaluating arguments based on the schemes. Critical questions are the set of questions that are considered relevant moves to the argument of a particular scheme.

In this report, we applied the theory of argumentation schemes to the task of annotating written arguments, as a first step toward our long-term goal of automatically providing rich, content-based scoring and feedback for oral and written arguments. Specifically, we developed annotation protocols for selected writing prompts in an analytical writing task from a college-level test. Human annotators were asked to identify which parts of a student's response raise critical questions relevant to the argument presented in the stimulus text. We then examined how well this approach captured the argument structures present in students' written responses.

It is important to demonstrate that this innovative annotation approach is both valid (as a method for describing argumentation in an essay) and usable (in that annotators can apply it reliably and consistently). Once we establish adequate reliability of the human annotation, we can use human-annotated models to train automated classifiers. To the extent that the resulting models work well, we can identify linguistic cues in the text that enable us to reconstruct the human classification. Where the automated models fail, we identify critical issues to resolve for future research to achieve a stronger model.

Theoretical Framework

There have been critical advances in the study of informal argument in recent years, most notably the development of theories that provide relatively rich schemata for classifying informal arguments, such as the theory of argumentation schemes (Walton, 1996). An argumentation scheme is defined as "a more or less conventionalized way of representing the relation between what is stated in the argument and what is stated in the standpoint" (van Eemeren & Grootendorst, 1992, p. 96). It presents the reasoning pattern, that is, how the conclusion is derived from the premises. Viewed as the "internal structure" of argumentation, it reflects justificatory standards for evaluating the reasonableness of an argument (van Eemeren & Grootendorst, 2004).

Walton and his colleagues (Walton, 1996; Walton, Reed, & Macagno, 2008) identified more than 60 argumentation schemes and developed critical questions associated with each scheme. Consider, for instance, the argumentation scheme *argument from consequences*, which is commonly used when responding to a policy question (deciding whether a particular course of action should be adopted). People argue for (or against) a proposed policy by citing positive (or negative) consequences that would follow if the policy were adopted. For example, someone could argue that governments should ban further construction of nuclear power plants because nuclear power will increase cancer rates in surrounding communities. Argument from consequences is a common method of arguing for a policy, but there are also objections that can be raised, focused on evaluating the consequences. For instance, an opponent could argue that the claimed consequences are not probable or less important than other (undesirable) consequences. These standard objections correspond to what the literature calls *critical questions*. For instance, the argument from consequences scheme is associated with critical questions like the following (Walton, 1996):

- How sure are you that the good (or bad) consequences will actually happen?
- Do you have evidence that these consequences probably will happen?
- Are there potentially opposing consequences that might happen if we implement the policy?

A similar analysis can be applied to practically any argumentation scheme. For instance, *argument from example* is another common form of informal argument. Once again, we may observe that only a few reasonable objections can be raised when someone cites a specific case as support for a general claim. We could question the facts of the case (*question accuracy*), raise doubts about whether the general statement is even relevant to the specific case (*question relevancy*), deny that this case is typical (*question typicality*), or suggest that special circumstances undermine the generalization from this case to other cases (*plead special circumstances*; Walton, 1996). Each of these potential objections corresponds to a critical question in the argument from example scheme. In other words, critical questions provide content-specific standards for evaluating the reasonableness of an argument that generalize across a wide range of arguments that belong to the same argumentation scheme.

Argumentation schemes have been applied in a variety of fields: to support automated detection of arguments (Mochales & Ieven, 2009; Palau & Moens, 2009; Rienks, Heylen, & Van der Weijden, 2005; Verbree, Rienks, & Heylen, 2006), to develop computational representation of arguments (Atkinson, Bench-Capon, & McBurney, 2006; Rahwan, Banihashemi, Reed, Walton, & Abdallah, 2011), to reconstruct the implicit parts of an argument (Feng & Hirst, 2011), to establish a classification system of arguments (Walton & Macango, 2016), and to teach argumentation skills (Nussbaum & Edwards, 2011; Song & Ferretti, 2013). In light of these studies, it appeared reasonable to assume that argumentation schemes would provide a useful framework for analyzing students' written responses in tests that aid in making high-stakes decisions. Perhaps most saliently, one of the most widely accepted college-level tests worldwide has a writing task designed to assess students' critical thinking and analytical writing skills. This analytical writing task thus provides a natural laboratory for evaluating argumentation scheme theory and determining whether it can support more effective methods for describing and scoring the content of student arguments.

In particular, it should be possible to annotate students' argumentative essays based on the theory of argumentation schemes and critical questions. An annotation scheme developed for this analytical writing task would be the first step toward developing automated content analysis (and scoring) of argument essays specifically for the quality of the argumentation they contain and, if successful, might be generalizable to other kinds of argument writing tasks.

Development of Annotation Protocols

The primary goal of this project was to develop annotation protocols that can be reliably used to analyze written arguments, based not on discourse connectives, sentence structure, vocabulary, or style but on substantive meanings and structure of arguments. While the ideal annotation should address specific content, it should also be generalizable, because the ultimate goal is to analyze argument structure in a variety of writing tasks.

The initial exploration focused on identifying argumentation schemes in 20 randomly selected argument prompts in this college-level test. Each prompt was associated with *topic notes* (i.e., a prompt-specific scoring guide that identified valid lines of argument and analysis). Initial efforts to develop annotation categories focused on mapping the content of the topic notes to critical questions and then examining how well the categories worked when applied to student essays. Walton's argumentation schemes provided an appropriate general framework for analyzing arguments in our study, but it could be challenging to apply Walton's schemes to the analysis of students' written responses due to ambiguities in the student responses (e.g., students may use the word "cause" when in fact they mean that a potential consequence could occur after implementing the proposed policy). In addition, many of Walton's schemes shared the same or similar critical questions, which made it difficult to determine which scheme was being applied in any particular case (e.g., a couple questions in the *argument from correlation to cause* and the *argument from cause to effect* are quite close). Duschl (2008) reported a similar challenge in his study of students' written arguments; he addressed the challenge by collapsing some of the categories. To support efficient annotation of argument structure, we adopted a comparable solution: modifying Walton's schemes (or creating new, compound schemes) that defined a simpler set of mutually exclusive categories and associated critical questions.

A close examination of the argument prompts in the test indicated that more than half of the 20 selected prompts dealt with a policy issue (making a decision for or against putting a practice into place to solve some practical problem). The following stimulus illustrates the use of policy issues in this task.

> The following appeared in a memorandum from the new president of the Patriot car manufacturing company.
>
> "In the past, the body styles of Patriot cars have been old-fashioned, and our cars have not sold as well as have our competitors' cars. But now, since many regions in this country report rapid increases in the numbers of newly licensed drivers, we should be able to increase our share of the market by selling cars to this growing population. Thus, we should discontinue our oldest models and concentrate instead on manufacturing sporty cars. We can also improve the success of our marketing campaigns by switching our advertising to the Youth Advertising agency, which has successfully promoted the country's leading soft drink."

Test takers are asked to analyze the reasoning in the argument, consider any assumptions, and discuss how well any evidence that is mentioned supports the conclusion.

This prompt shows a typical pattern in discussing policy issues: (a) a *problem* is stated (e.g., the body style of Patriot cars is old-fashioned), (b) a *plan* is proposed (e.g., discontinuing oldest car models and manufacturing sporty cars and switching to the Youth Advertising agency), and (c) a desirable *goal* will be achieved (e.g., an increase in their market share) if the plan is implemented. This problem – solution structure corresponds in part to Walton's argument from consequences scheme but implicates a number of other, closely related schemes and includes a causal argument (because the proposed plan must be causally efficacious), which corresponds in part to Walton's schemes for argument from correlation to cause and argument from cause to effect. Almost all the policy arguments we examined could be critiqued in very similar ways, corresponding to this complex conceptual structure.

Accordingly, we revised or rearranged some of the critical questions in Walton's theory and created new questions when necessary. For example, challenges to arguments that use a policy scheme fall into the following six categories: (a) problem, (b) goal, (c) plan implementation, (d) plan definition, (e) side effect, and (f) alternative plan. If a student were to write that the president should reevaluate whether this is really a problem, we would mark the response in the "problem" category; when a student questions if there is a better plan to achieve the goal, the response should be categorized in the "alternative plan" category. Similarly, we grouped possible responses into four categories for the causal scheme: challenges to mechanism, challenges to efficacy, challenges to applicability, and intervening factors. Critical questions could be raised for each of these categories. For example, under "causal mechanism," one could ask whether there is a correlation between old-fashioned style and poor sale, whether this is merely a coincidence, or whether other problems (rather than old-fashioned style) could have caused poor sales; under "causal applicability," one could ask if the assumption that young people like sporty cars is true. Each category may cover one or more critical questions relevant to the point of evaluation. Our detailed analysis of the policy scheme and causal scheme of the Patriot Car prompt is presented in Figures 1 and 2, respectively. The general critical questions (can be generalized across prompts based on the same scheme) are presented in bold, and prompt-specific questions are included below the general critical questions.

Some argument prompts use findings from studies and surveys as evidence to support a claim. We developed a scheme named *argument from a sample* to capture this prototype argument structure, which was not specifically addressed in

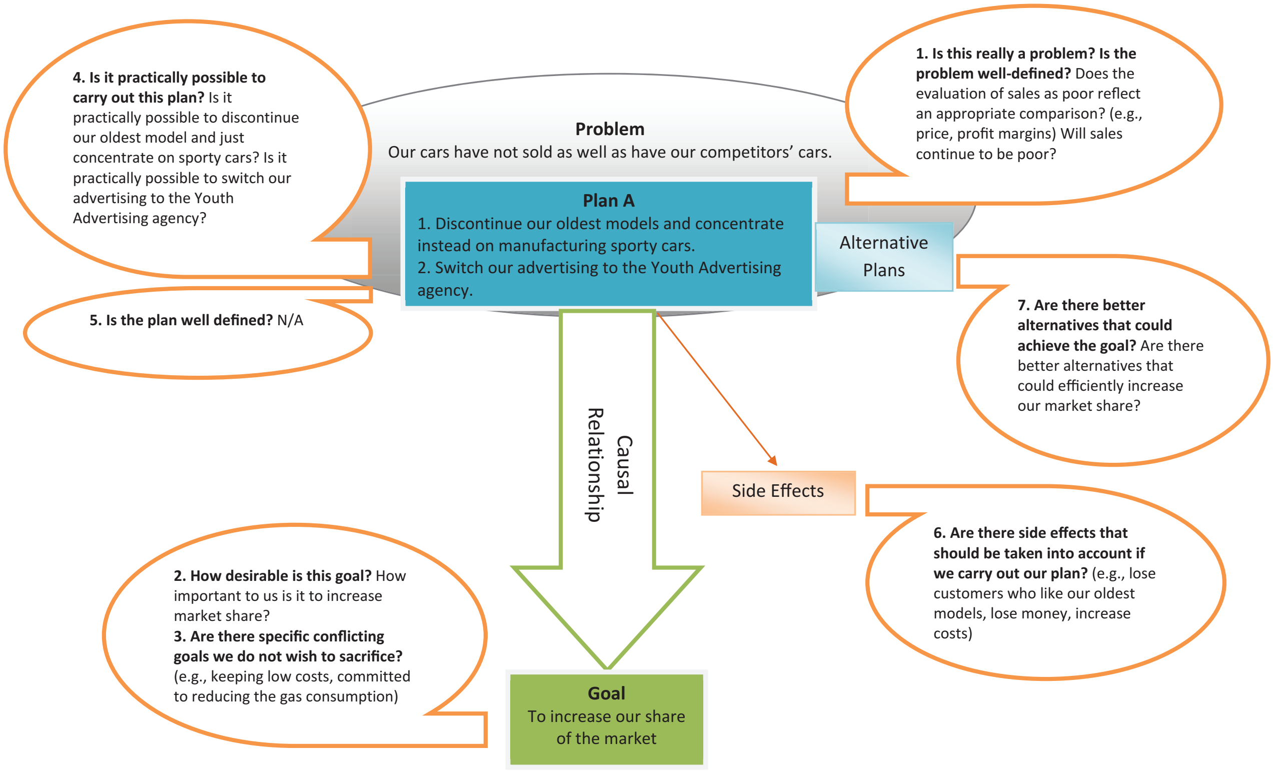

4. Is it practically possible to carry out this plan? Is it practically possible to discontinue our oldest model and just concentrate on sporty cars? Is it practically possible to switch our advertising to the Youth Advertising agency?
Problem
Our cars have not sold as well as have our competitors' cars.
1. Is this really a problem? Is the problem well-defined? Does the evaluation of sales as poor reflect an appropriate comparison? (e.g., price, profit margins) Will sales continue to be poor?
Plan A
1. Discontinue our oldest models and concentrate instead on manufacturing sporty cars.
2. Switch our advertising to the Youth Advertising agency.
Alternative Plans
5. Is the plan well defined? N/A
7. Are there better alternatives that could achieve the goal? Are there better alternatives that could efficiently increase our market share?
Causal Relationship
Side Effects
6. Are there side effects that should be taken into account if we carry out our plan? (e.g., lose customers who like our oldest models, lose money, increase costs)
2. How desirable is this goal? How important to us is it to increase market share?
3. Are there specific conflicting goals we do not wish to sacrifice? (e.g., keeping low costs, committed to reducing the gas consumption)
Goal
To increase our share of the market

Figure 1 Patriot Car policy scheme.

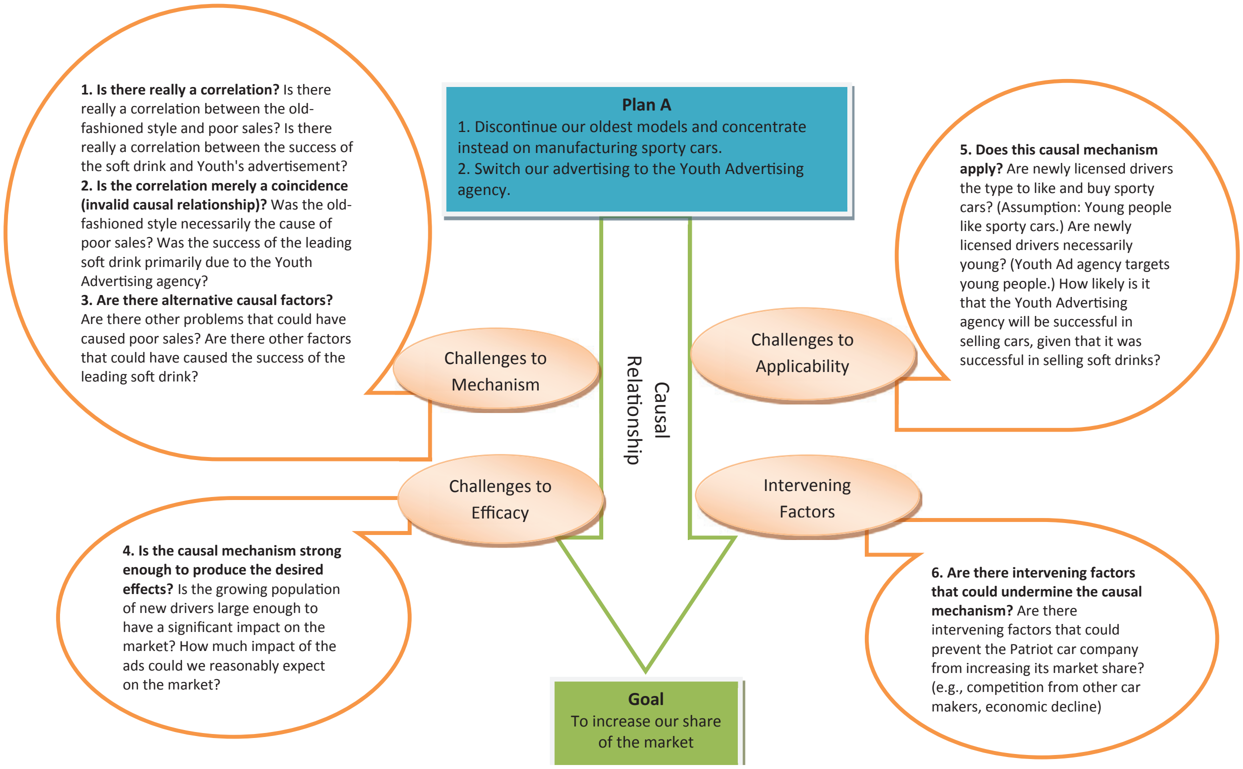

1. Is there really a correlation? Is there really a correlation between the old-fashioned style and poor sales? Is there really a correlation between the success of the soft drink and Youth's advertisement?
2. Is the correlation merely a coincidence (invalid causal relationship)? Was the old-fashioned style necessarily the cause of poor sales? Was the success of the leading soft drink primarily due to the Youth Advertising agency?
3. Are there alternative causal factors? Are there other problems that could have caused poor sales? Are there other factors that could have caused the success of the leading soft drink?
Plan A
1. Discontinue our oldest models and concentrate instead on manufacturing sporty cars.
2. Switch our advertising to the Youth Advertising agency.
5. Does this causal mechanism apply? Are newly licensed drivers the type to like and buy sporty cars? (Assumption: Young people like sporty cars.) Are newly licensed drivers necessarily young? (Youth Ad agency targets young people.) How likely is it that the Youth Advertising agency will be successful in selling cars, given that it was successful in selling soft drinks?
Challenges to Mechanism
Causal Relationship
Challenges to Applicability
Challenges to Efficacy
Intervening Factors
4. Is the causal mechanism strong enough to produce the desired effects? Is the growing population of new drivers large enough to have a significant impact on the market? How much impact of the ads could we reasonably expect on the market?
6. Are there intervening factors that could undermine the causal mechanism? Are there intervening factors that could prevent the Patriot car company from increasing its market share? (e.g., competition from other car makers, economic decline)
Goal
To increase our share of the market

Figure 2 Patriot Car causal scheme.

Walton's work. Evaluations of this type of argument involve examining several aspects of a study or survey, such as the sample size, representativeness of the sample, and validity of the study design. We also created an *argument from comparison* scheme for when an argument involves comparing two parties based on a few criteria. The focus of evaluating arguments from comparison includes examinations of the existing criteria as well as additional criteria to determine which party is better than the other.

As we developed annotation protocols, we had access to *topic notes* for each prompt. Topic notes are scoring guides intended to help scorers determine whether a response provides a valid critique of the argument presented in a specific prompt. For example, the topic note for the Patriot Car prompt includes the following point:

> Patriot may sell fewer cars than do its competitors, but Patriot's cars may be expensive ones with high profit margins and the competitors' cars may be inexpensive ones with low profit margins.

This topic note calls scorers' attention to one possible line of attack: whether existing sales were really as poor as indicated in the stimulus. If existing sales are not actually so bad, there may not be a problem. In other words, this particular line of attack fits the "problem" category in our framework.

We systematically analyzed the topic notes in this fashion, classifying the points they raised using the schemas in Figures 1 and 2. This analysis provided an empirical foundation for the annotation protocols we developed, because the topic notes are developed through review of a broad range of actual student responses. However, our analysis enabled us to fill gaps in the existing topic notes to create an even more comprehensive list of valid responses. For instance, none of the topic notes anticipates an alternative plan for the company, but the policy scheme suggests that alternative plans could be relevant to the argument, and there are, indeed, responses to the Patriot Car prompt in which alternative plans are suggested.

Table 1 lists all 20 argument prompts and identifies the argumentation schemes relevant to each. Altogether, we identified six argumentation schemes. Analyses of these argumentation schemes and their critical questions are presented in Table 2.

Table 1 Distribution of Argumentation Schemes Over 20 Prompts

	Argumentation scheme					
Prompt	Policy	Causal	Sample	Example	Comparison	Position
1				X	X	
2		X				X
3	X	X	X			
4	X	X	X			
5		X	X			
6	X	X				
7	X	X	X			X
8			X	X		
9	X	X				
10	X	X				
11	X	X	X	X		
12		X			X	
13	X	X				
14			X			
15	X	X				
16	X	X				
17		X	X			
18	X	X				
19					X	
20			X			

The causal and policy schemes shown in Figures 1 and 2 were most common in this set of prompts, although argument from a sample was also commonly used. A number of other schemes also played a role, including argument from example, argument from comparison, and argument from a position to know. Note that other prompts may include additional argumentation schemes.

When we developed the annotation protocols for each prompt, we considered not only the topic notes (and the argumentation schemes we had identified for each prompt) but a selection of student essays. We modified and revised our annotation protocols to develop clear-cut categories that would allow us to identify, for any part of a student essay, whether the essay had raised a particular critical question, based on one of the argumentation schemes relevant to that prompt.

To summarize, we analyzed a subset of argument prompts to identify common argumentation schemes and determine whether these schemes can realistically be identified by human annotators. We went through multiple rounds of formulation, application to samples, and reformulation to make the annotation categories clear-cut and easy for annotators to understand, distinguish, and apply. The annotation protocols include argumentation schemes identified in a prompt and associated critical questions that can be raised to challenge the original argument, providing a systematic classification of written responses that are considered as valid moves for the given argumentation schemes. In what follows, we describe the application of these annotation protocols.

Application of the Annotation Protocols

Our next step was to test if the annotation protocols can be used reliably by human annotators to analyze students' responses in the analytical writing task. Given that this was a new approach, we provided intensive training to the annotators, including annotating a small set of essays as practice. In this section, we describe the annotation rules, the tool used, and the annotation procedure, and we report the interannotator agreements.

Annotation Rules

The goal of this annotation work was to identify and classify any text segment (e.g., paragraph, sentence, or clause) that addresses a critical question. Usually, the minimal text segment is at the sentence level, but it could also be at the phrase

Table 2 Annotation Categories That Map to Critical Questions Under the Schemes in the 20 Prompts

Scheme	Category	Critical question
Policy	Problem	Is this really a problem? Is the problem well defined?
	Goal	How desirable is this goal? Are there specific conflicting goals we do not wish to sacrifice?
	Plan implementation	Is it practically possible to carry out this plan?
	Plan definition	Is the plan well defined?
	Side effects	Are there side effects that should be taken into account if we carry out our plan?
	Alternative plan	Are there better alternatives that could achieve the goal?
Causal	Causal mechanism	Is there really a correlation? Is the correlation merely a coincidence (invalid causal relationship)? Are there alternative causal factors?
	Causal efficacy	Is the causal mechanism strong enough to produce the desired effects?
	Applicability	Does this causal mechanism apply?
	Intervening factors	Are there intervening factors that could undermine the causal mechanism?
Sample	Significance	Are the patterns we see in the sample clear-cut enough (and in the right direction) to support the desired inference?
	Representativeness	Is there any reason to think that this sample might not be representative of the group about which we wish to make an inference?
	Stability	Is there any reason to think this pattern will be stable across all the circumstances about which we wish to make an inference?
	Sample size	Is there any reason to think that the sample may not be large enough and reliable enough to support the inference we wish to draw?
	Validity	Is the sample measured in a way that will give valid information on the population attributes about which we wish to make inferences?
	Alternatives	Are there external considerations that could invalidate the claims?
Example	Accuracy	Is the proposition presented by the example in fact true?
	Relevancy	Is there any reason to think that this case might not really be an example of the general statement we are being asked to believe?
	Typicality	Is there any reason to think that this case might not be typical of the kinds of cases that the general statement covers?
	Special circumstances	Are there any special circumstances that might make it dangerous to generalize from this case to other cases?
Comparison	Criterion evaluation	Is this a good criterion?
	Comparison	Do we have evidence for a comparison on this criterion?
	Difference	Are there any differences between the two incidences that could undermine the fairness of comparison?
	Additional criterion	Have we missed any important criterion?
Position	Position	Is the person in a position to know whether it is true?
	Source reliability	Is the person a reliable/credible/honest source?
	Assertion	Did the person assert that it is true?

level when a sentence includes multiple points that correspond to more than one critical question. For each highlighted unit, the annotator will choose a specific topic (the content being addressed), a category (identifying the argumentation scheme and critical question), and, if necessary, a second topic (e.g., in the Patriot Car prompt, car body style and advertising are considered two distinct topics, and if both are addressed in the same sentence, raters are expected to indicate this fact; however, most of the prompts only have one topic).

Some text may not be highlighted during annotation. Nonhighlighted text typically includes generic responses, related to the notion of shell language (Madnani, Heilman, Tetreault, & Chodorow, 2012), restatements of the text in the prompt, statements expressing agreement with the arguments in a prompt, transitional sentences, rhetorical attacks, and any other statements that do not address a critical question. Surface errors (e.g., grammar and spelling) are ignored if they do not prevent people from understanding the meaning in the text.

Following is an example of an annotated excerpt from a student essay.

> Once it hones in on the most desirable type of car, it would not be a bad idea for the company to get a new advertising agency. But it would probably be beneficial to Patriot to research what agencies are serving the most successful car companies, not just the most successful youth advertising because, as stated before, it is probably not just the young people Patriot should focus their attention on, and a company that really knows how to sell soft-drinks does not necessarily really know how to sell cars. [**Advertising Agency: Causal Applicability; Critical Question: Does this causal mechanism apply? How likely is it that the Youth Advertising agency will be successful in selling cars, given that it was successful in selling soft drinks?**]
>
> …
>
> Lastly, if, in order to afford these changes, Patriot must discontinue some number of models, it should first thoroughly examine which ones are selling the least well. It sounds like it might only be a guess that the older models are less popular; it is possible these older models are what made Patriot successful in the first place. [**Body Styles: Causal Mechanism; Critical Question: Is there really a correlation between the old-fashioned style and poor sales?**]

In conclusion, the president of the Patriot car manufacturing company has many flaws in his logic. Without proper research into the target audience and the proper product that would increase their share in the market, the president cannot make the proper decisions to help make the Patriot car manufacturing company successful.

Annotation Tool

All the annotations were carried out using an existing annotation tool, Text Rover, developed and used at Educational Testing Service (ETS). The interface enables annotators to access not only the student essay but also the original writing prompt and the annotation protocol for that prompt. Specifically, the annotation protocol provides information about topics, annotation categories, and critical questions. Its implementation in Text Rover allows annotators to highlight a segment from the text to be annotated and click a button to choose a topic and a category to identify which critical questions were being addressed. Figure 3 provides a screen shot of the Text Rover annotation interface.

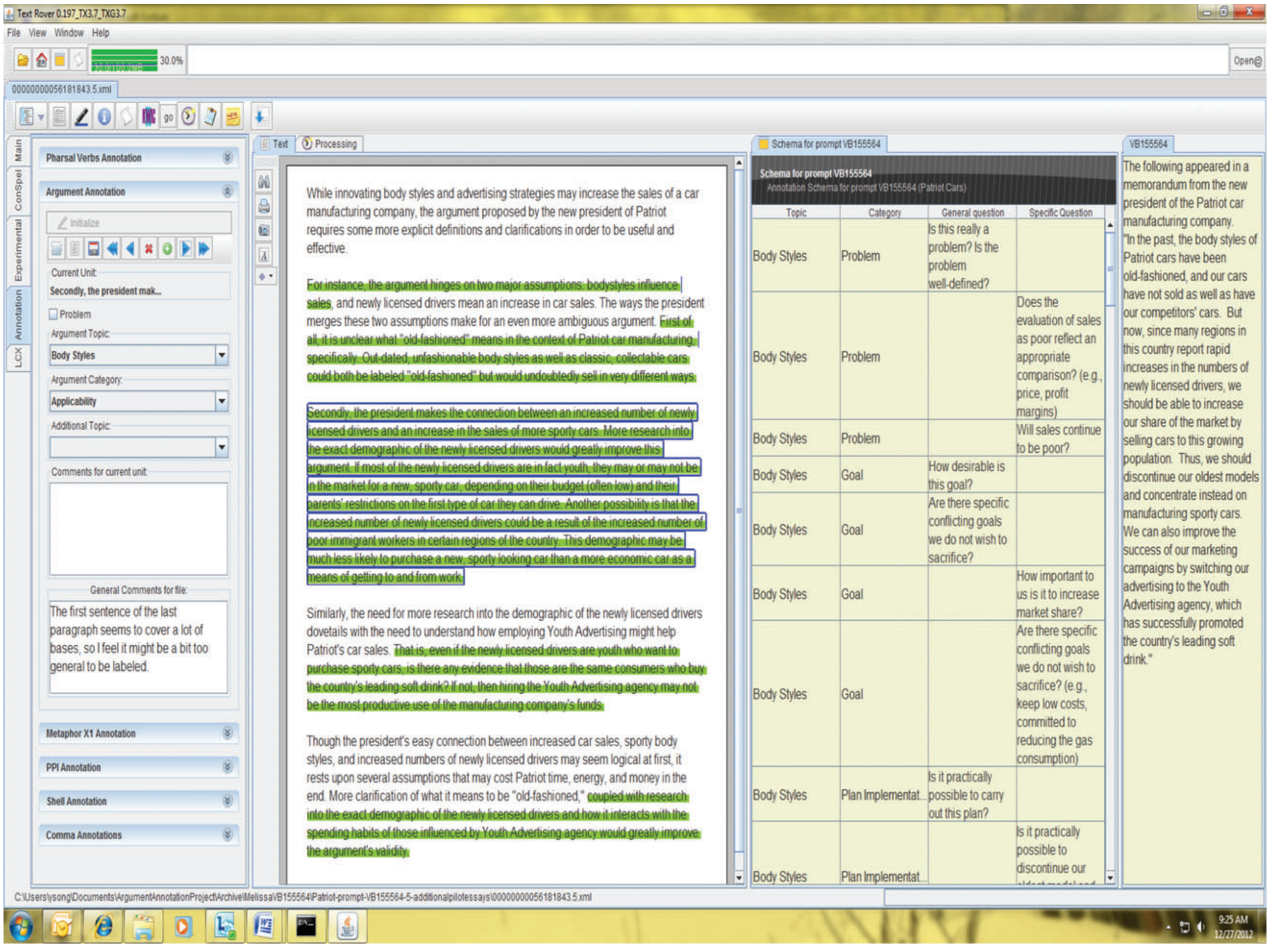
The leftmost column displays the labels assigned by an annotator and associated comments. The center column shows the student response, with highlighting where specific annotations have been provided. The rightmost column presents the prompt and annotation protocol in tabular form. In this interface, critical questions are presented both in general form (i.e., as questions that can be applied to a variety of prompts that employ the same argumentation scheme) and in specific form (i.e., as questions focused on prompt-specific content), grouped by topic and category.

Data and Annotation Procedures

As a first step toward empirical validation of our annotation protocols, we selected four prompts and a set of student responses for annotation. All essays were written by students who took the college-level test between 2008 and 2010. We restricted ourselves to essays that scored 5 or 6 on the 6-point scale, on the grounds that high-scoring essays were likely to contain relatively large numbers of valid critiques and were less likely to have ambiguous arguments or unclear argument structures than essays with low scores. For each of the four prompts, we first randomly selected 25 essays as the practice pool and 50 essays for the main annotation. Our goal was to determine whether we could apply the annotation categories consistently to exemplar essays.

Four research assistants, coming from a background in language-related disciplines in college (with a bachelor's or master's degree), received training on the annotation tool and the target prompts and then annotated the sample essays; two annotators worked on each prompt. Training included an introduction to the theory of argumentation scheme, a clarification of the goal and steps, examples of argument prompts to illustrate argumentation schemes and critical questions, key components of the annotation, and use of the software Text Rover. The first author delivered all the training to both pairs and helped resolve differences of opinion when annotators disagreed during training.

The two pairs started with different prompts, so their training guides were customized toward the assigned prompt. Pair 1 started training on Prompt 14 (see the Pain Medication prompt text in Table 3), whereas Pair 2 started with Prompt 15



00000000056181843.5.xml

Current Unit: Secondly, the president mak...

□ Problem

Argument Topic: Body Styles

Argument Category: Applicability

Additional Topic:

Comments for current unit:

General Comments for file: The first sentence of the last paragraph seems to cover a lot of bases, so I feel it might be a bit too general to be labeled.

While innovating body styles and advertising strategies may increase the sales of a car manufacturing company, the argument proposed by the new president of Patriot requires some more explicit definitions and clarifications in order to be useful and effective.

For instance, the argument hinges on two major assumptions: bodystyles influence sales, and newly licensed drivers mean an increase in car sales. The ways the president merges these two assumptions make for an even more ambiguous argument. First of all, it is unclear what "old-fashioned" means in the context of Patriot car manufacturing, specifically. Out-dated, unfashionable body styles as well as classic, collectable cars could both be labeled "old-fashioned" but would undoubtedly sell in very different ways.

Secondly, the president makes the connection between an increased number of newly licensed drivers and an increase in the sales of more sporty cars. More research into the exact demographic of the newly licensed drivers would greatly improve this argument. If most of the newly licensed drivers are in fact youth, they may or may not be in the market for a new, sporty car, depending on their budget (often low) and their parents' restrictions on the first type of car they can drive. Another possibility is that the increased number of newly licensed drivers could be a result of the increased number of poor immigrant workers in certain regions of the country. This demographic may be much less likely to purchase a new, sporty looking car than a more economic car as a means of getting to and from work.

Similarly, the need for more research into the demographic of the newly licensed drivers dovetails with the need to understand how employing Youth Advertising might help Patriot's car sales. That is, even if the newly licensed drivers are youth who want to purchase sporty cars, is there any evidence that those are the same consumers who buy the country's leading soft drink? If not, then hiring the Youth Advertising agency may not be the most productive use of the manufacturing company's funds.

Though the president's easy connection between increased car sales, sporty body styles, and increased numbers of newly licensed drivers may seem logical at first, it rests upon several assumptions that may cost Patriot time, energy, and money in the end. More clarification of what it means to be "old-fashioned," coupled with research into the exact demographic of the newly licensed drivers and how it interacts with the spending habits of those influenced by Youth Advertising agency would greatly improve the argument's validity.

Schema for prompt VB155564

Annotation Schema for prompt VB155564 (Patriot Cars)

Topic	Category	General question	Specific Question
Body Styles	Problem	Is this really a problem? Is the problem well-defined?	
Body Styles	Problem		Does the evaluation of sales as poor reflect an appropriate comparison? (e.g., price, profit margins)
Body Styles	Problem		Will sales continue to be poor?
Body Styles	Goal	How desirable is this goal?	
Body Styles	Goal	Are there specific conflicting goals we do not wish to sacrifice?	
Body Styles	Goal		How important to us is it to increase market share?
Body Styles	Goal		Are there specific conflicting goals we do not wish to sacrifice? (e.g., keep low costs, committed to reducing the gas consumption)
Body Styles	Plan Implementat...	Is it practically possible to carry out this plan?	
Body Styles	Plan Implementat...		Is it practically possible to discontinue our

VB155564

The following appeared in a memorandum from the new president of the Patriot car manufacturing company.
"In the past, the body styles of Patriot cars have been old-fashioned, and our cars have not sold as well as have our competitors' cars. But now, since many regions in this country report rapid increases in the numbers of newly licensed drivers, we should be able to increase our share of the market by selling cars to this growing population. Thus, we should discontinue our oldest models and concentrate instead on manufacturing sporty cars. We can also improve the success of our marketing campaigns by switching our advertising to the Youth Advertising agency, which has successfully promoted the country's leading soft drink."

C:\Users\ysong\Documents\ArgumentAnnotationProject\Archive\Melissa\VB155564\Patriot-prompt-VB155564-5-additionalpilotessays\00000000056181843.5.xml

9:25 AM
12/27/2012

Figure 3 A screen shot of Text Rover.

(i.e., the Patriot Car prompt). Each pair received a package that contained a detailed analysis of their prompt. It included diagrams presenting an analysis of the relevant argument structures, the topic notes for that prompt, and the annotation protocol, embedded within Text Rover. During training, the research assistants reviewed the annotation package with the lead author and then examined three essays as a group, applying and discussing the annotation labels. Problems and differences of opinions were discussed and resolved immediately. Next, the research assistants annotated training essays independently as practice, and then they reviewed their responses with the lead author until they followed the annotation rules and applied the annotation labels consistently. The number of practice essays varied slightly across prompts, usually between 10 and 15 essays.

After completing training, the research assistants began the actual annotation process. Each pair annotated 50 essays for each prompt assigned to them. To check their agreement, all essays were annotated independently by both annotators. After they had finished all the annotation for a prompt, the annotators discussed problems that they encountered and questions they had with the lead author but were not allowed to change their annotations. The annotators also provided informal suggestions on how to improve the annotation protocols for future use. For instance, they reported cases in which student responses matched one of the general critical questions but were not covered by any of the specific questions developed for that prompt. While the list of content-specific critical questions was not intended to be exhaustive, we added specific questions to the annotation protocols if they appeared frequently in student essays. One pair of annotators also observed that the line between the validity category and alternatives category was not clear, so we rephrased the questions under these two categories to make them more distinct.

After the annotators had completed one prompt, they proceeded to another prompt (Pair 1 to Prompt 3; Pair 2 to Prompt 18). The training, practice, and annotation procedures for the second prompt were exactly the same as for the first

Table 3 Two Sample Prompts

Prompt title	Prompt text
Pain medication (Prompt 14)	A new report suggests that men and women experience pain very differently from one another, and that doctors should consider these differences when prescribing pain medications. When researchers administered the same dosage of kappa opioids — a painkiller — to 28 men and 20 women who were having their wisdom teeth extracted, the women reported feeling much less pain than the men, and the easing of pain lasted considerably longer in women. This research suggests that kappa opioids should be prescribed for women whenever pain medication is required, whereas men should be given other kinds of pain medication. In addition, researchers should reevaluate the effects of all medications on men versus women.
Bone fractures (Prompt 3)	Typically, as people age, their bone mass decreases, making them more vulnerable to bone fractures. A recent study concludes that the most effective way to reduce the risk of fractures in later life is to take twice the recommended dose of vitamin D and calcium daily. The three-year study followed a group of French women in their 80s who were nursing-home residents. The women were given daily supplements of twice the recommended dose of vitamin D and calcium. In addition, the women participated in a light weight-lifting program. After three years, these women showed a much lower rate of hip fractures than is average for their age.

prompt. Each pair practiced on an initial set of training essays and then annotated 50 essays for the second prompt. Once again, all essays were doubly annotated. Finally, to see if there were any additional problems, the two pairs of annotators switched their first prompts (Pair 1 to Prompt 15; Pair 2 to Prompt 14) and annotated a new set of 50 randomly selected essays at score points 5 or 6, using the updated annotation protocols. In total, each pair annotated 150 essays, 50 essays per each of 3 prompts.

Interannotator Agreement

We defined *interannotator agreement* by examining the extent to which annotators not only highlighted the same portions of the targeted essays but also applied the same label to each text segment. We considered two annotators as being in agreement on the classification of a given sentence[1] only when they assigned the same topic and category labels to it or both left it unannotated. Interannotator agreement for each label was measured using the kappa statistic. Unannotated text was treated as an additional category (generic). We also used the kappa statistic to measure how well annotators agreed in identifying sentences as generic in content.

Table 4 shows the kappa value for each category across the three prompts annotated by Pair 1, and Table 5 presents comparable agreement statistics for Pair 2. Because it is not realistic to require annotators to reach a good agreement for rare categories, we excluded particularly rare categories from these tables (i.e., those that were assigned less than one sentence on average per essay). The kappa values for common categories ranged from .54 to .94, with only one exception (i.e., .37 for Advertising Agency: Causal mechanism in Prompt 15 for Pair 1). The majority of kappa values for Pair 1 were greater than .70, and the majority of kappa values for Pair 2 were greater than .60, which suggested that the annotators applied the labels in a relatively consistent way, taking into account the difficulty of this annotation task. The statistics also showed an improvement in the reliability of the Patriot Car prompt (i.e., Prompt 15) after we updated the annotation protocols, but the effect on the Pain Medication prompt (i.e., Prompt 14) was unclear. This result might be due to a pair effect, as Pair 1 seemed to have better agreement in general compared with Pair 2.

Furthermore, we conducted crosstab analyses to examine where disagreements were likely to occur. Table 6 presents an example of such an analysis on Prompt 14, including the number of sentences marked by each annotator for each category (note that rare categories were collapsed into the category labeled "Other"). The results across these prompts consistently show that the most common disagreements between the two annotators in each pair were related to the generic category. In other words, one annotator marked a sentence as containing a critical question that belonged to a particular category, while the other annotator decided that the sentence did not include any critical questions. In Pair 1, one annotator tended to assign the generic category more frequently than the other annotator in all three prompts, but Pair 2 did not show a clear pattern of individual difference in using the generic category. Sometimes, it is just hard to make a decision on whether a sentence contains enough information to be marked as an argument tied to a specific critical question due to its implicit or generic language. Consider the following statement: "The study should be completed again with a different sample of men and women." One annotator evaluated that this sentence addresses the sample size issue of the study, while the other annotator did not think that it clearly refers to sample size as it never says that the original sample size was small and more participants were needed.

Furthermore, crosstab analyses helped us find which categories created confusion. For example, Pair 1 data in Table 6 show that 42 annotations were assigned a label of "stability" by one annotator and a label of "validity" by the other annotator, indicating that these two categories created confusion among the raters and confirming the low kappa values in both categories. We examined the cases where the disagreements occurred and added specific critical questions (relevant to these cases) to the annotation protocols to clarify the criteria for the stability and validity categories. After we modified the annotation protocols, Pair 2 was able to distinguish them (see Table 6, Pair 2).

Discussion

In this report, we presented a new approach to systematically identifying and classifying written arguments. Specifically, we analyzed 20 argument prompts from a worldwide college-level test, identified typical argumentation schemes, listed critical questions associated with each argumentation scheme, and matched the topic notes with appropriate critical questions. Our pilot annotation focused on a small group of high-quality essays (scored 5 or above) across different prompts to examine whether humans (suitably trained) can reliably apply the annotation labels.

The development of our annotation approach was grounded in Walton's (1996) (Walton et al., 2008) theory of argumentation schemes, which classifies various types of arguments based on their reasoning patterns. It is critical to understand the reasoning patterns when analyzing arguments because argumentation is a social activity of reasoning (van Eemeren & Grootendorst, 1992). This method enabled us to analyze argument structure in a way that reflects its reasoning pattern without ignoring or abstracting its content. The analysis focuses on the core of argumentation rather than on content-independent or word-level features (e.g., usage, style, discourse connectives, essay organization, vocabulary).

Although Walton's theory of argumentation schemes is considered the most comprehensive classification of arguments, it is not appropriate to use these schemes directly as annotation protocols for four reasons. First, it is challenging to fully understand the 60 different argumentation schemes sketched out by Walton et al. (2008). The number of schemes is so overwhelming that annotators would not be able to learn all the schemes in a limited training time. Second, the lines between some argumentation schemes are blurry (e.g., argument from consequences, argument from goal), which makes the schemes indistinguishable in some cases. Third, some reasoning patterns in the analytical writing task are not covered by Walton's schemes, and thus there could be a mismatch if we simply relied on his argumentation schemes. Fourth, it is unlikely that all of the 60 argumentation patterns could reasonably apply in the given argumentative setting; indeed, our analysis so far suggests that a subset containing 10 of Walton's schemes is sufficient to cover our data.

Therefore we modified Walton's schemes and created new schemes that best reflect the argument structures in the writing task. Among the six identified argumentation schemes, two schemes were from Walton's work (i.e., argument from example, argument from position to know), two were a combination of Walton's schemes (i.e., policy scheme, causal scheme), and the other two were created by us (i.e., argument from a sample, argument from comparison). Please note that owing to the nature of tests that aid in making high-stakes decisions, the prompts are designed to be comparable in terms of argument types, difficulty level, and structures, and therefore these prompts only elicit a relatively small number of argumentation schemes.

In developing the annotation protocols, we also took into account topic notes written by assessment developers (see the Development of Annotation Protocols section for a description of topic notes). These topic notes were helpful in terms of understanding the purpose of the assessment and confirming that we covered all the valid responses proposed by people who had designed the task. Although topic notes identify the valid arguments, they are prompt dependent, with little room for generalization. For each new prompt, one has to develop a new list of topic notes because the content of the arguments changes. No two prompts share exactly the same topic notes. In contrast, various writing prompts share the same argumentation schemes, which allows us to develop rules that can be generalized across prompts. In addition, although the topic notes for many of the prompts are quite extensive and detailed, topic notes in general are not necessarily comprehensive in their coverage of all possible valid critical questions. Thus it is not appropriate to use topic notes as the only criterion for argument annotation.

Table 4 Agreement of Pair 1

Prompt	Category	Kappa
14	Generic	.65
	Pain medication: Stability	.57
	Pain medication: Sample size	.84
	Pain medication: Validity	.58
	Pain medication: Alternatives	.72
3	Generic	.72
	Bone fractures: Representativeness	.84
	Bone fractures: Stability	.80
	Bone fractures: Sample size	.94
	Bone fractures: Validity	.74
	Bone fractures: Alternatives	.73
15	Generic	.69
	Body styles: Side effects	.71
	Body styles: Causal mechanism	.82
	Body styles: Applicability	.71
	Advertising agency: Causal mechanism	.37
	Advertising agency: Applicability	.75

Table 5 Agreement of Pair 2

Prompt	Category	Kappa
15	Generic	.54
	Body styles: Applicability	.69
	Body styles: Causal mechanism	.57
	Advertising agency: Applicability	.70
18	Generic	.70
	Viewers: Problem	.64
	Viewers: Causal mechanism	.68
	Viewers: Applicability	.65
	Advertising agency: Causal mechanism	.81
14	Generic	.61
	Pain medication: Representativeness	.57
	Pain medication: Stability	.72
	Pain medication: Sample size	.71
	Pain medication: Validity	.63
	Pain medication: Alternatives	.61

Table 6 Crosstab Analysis of the Interannotator Agreement: An Example (Prompt 14)

Pair 1: Annotator 1/Annotator 2	Generic	Other	Alternatives	Sample size	Stability	Validity
Generic	343	4	6	2	27	20
Other	3	10	0	0	0	8
Alternatives	28	8	121	0	1	9
Sample size	15	1	0	58	0	0
Stability	56	0	3	0	113	42
Validity	14	1	11	2	3	92

Pair 2: Annotator 1/Annotator 2	Generic	Other	Alternatives	Representativeness	Sample size	Stability	Validity
Generic	238	0	39	10	24	50	35
Other	10	18	2	0	0	0	10
Alternatives	7	0	139	21	4	3	11
Representativeness	1	0	3	37	8	0	1
Sample size	5	0	0	2	61	0	1
Stability	20	0	12	0	1	173	8
Validity	4	0	29	4	0	5	122

Note. As the two pairs annotated different sets of essays, the categories identified were not exactly the same.

We have to note that this kind of annotation work is labor intensive. Asking critical questions is a skill that some people do not possess. Many of those who do possess it have built it up implicitly, by practicing and reading arguments, until they have a variety of implicit argumentation schemes in their heads. Thus people need sufficient training time to make this tacit knowledge more explicit and apply the protocols consistently. In annotation, they must not only identify meaningful chunks of textual information in the essays but also assign the right label for the selected text. Despite these complexities, it is a worthwhile investigation, because developing a systematic classification of argument structures not only plays a critical role in this project but also has a potential contribution to other assessments on argumentation skills (e.g., essay tasks in state tests).

Our results indicate a moderate to strong degree of agreement between annotators ($\kappa = .70$ for one pair of annotators, $\kappa = .60$ for the second pair). However, the way we applied the kappa statistic was relatively stringent, because some annotators may have agreed that the same category applied but applied that category to a slightly larger or smaller portion of the student essay. Although we hope to improve performance in future studies (e.g., discussing more practice essays to resolve issues that could lead to disagreement), we believe that the level of annotator agreement we obtained shows that this method captures the argument structure in the student essays we analyzed and can be applied by annotators in a relatively consistent manner.

This work has a number of potential applications. First, it provides a systematic approach to analyzing the argument content of critical essays of the sort used in the writing task. As a result, it could provide a more efficient method of developing new prompts, focused on sampling particular argumentation schemes and providing scope for critiques based on specific critical questions. It could also provide a more efficient method of developing topic notes to support argument essay scoring.

Second, and more generally, the argument structures we have identified can be applied much more broadly, to a variety of argumentation tasks, and could support development of assessments and instructional approaches for younger students, not just for this postsecondary population. The writing task analyzed in this report is designed just for students who are going to apply to graduate schools, whereas younger students might benefit from learning to critically examine an argument, using critical questions as a criterion. Formative assessments focused on this skill may help promote students' reasoning skills.

Third, it provides a method for precisely indicating what parts of a student essay provide critical content relevant to the analytical writing task. This kind of detailed annotation is a prerequisite to any form of automated argument analysis and, especially, feedback.

In our ongoing work, we have been extending human annotations to a large pool of student essays (both high-quality and low-quality essays) to support the development of an automated scoring system. Extending the annotation method to essays across the full score range poses additional challenges, because essays that achieve low scores are likely to have problems in the area of relevancy, coherence, and logic, which will create difficulties in interpreting and identifying the arguments.

Successful annotation is the first step toward building an automated scoring system for evaluating written arguments. Preliminary results of the natural language processing work aimed at automatically detecting argumentation-relevant parts in students' essays have been reported separately (see Song, Heilman, Beigman Klebanov, & Deane, 2014).

Overall, our results suggest that the conceptual elements of Walton's approach to informal argumentation can fruitfully be extended to support developing and scoring assessments focused on critical reasoning. Although many challenges remain to be addressed, we have been able to provide a detailed analysis of how student essays address the problem of criticizing someone else's arguments, using argumentation schemes and critical questions.

Acknowledgments

We thank our annotators, Patrick D. Houghton, Melissa Lopez, Matthew Mulholland, and Laura McCulla, for applying this new approach. We would also like to thank Michael Flor for developing and refining Text Rover to support this project, Diane Napolitano for sampling student essays and compiling the data files, and Binod Gyawali for assisting with the data analyses.

Note

[1] Although annotators could mark clauses, these cases were rare. Most of the time, whole sentences were marked. Therefore, we treated all annotations as being for whole sentences. We also treated multisentence annotations as separate sentence-level instances.

References

Atkinson, K., Bench-Capon, T., & McBurney, P. (2006). Computational representation of practical argument. *Synthese*, *152*, 157–206.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v. 2 *Journal of Technology, Learning, and Assessment*, *4*(3), 1–31. Retrieved from http://www.jtla.org

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (Research Report No. RR-04-04). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2004.tb01931.x

Council of Chief State School Officers & National Governors Association. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: Author. Retrieved from http://www.corestandards.org/ela-literacy

Duschl, R. A. (2008). Quality argumentation and epistemic criteria. In S. Erduran & M. P. Jimenez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom based research* (pp. 159–175). Dordrecht, Netherlands: Springer.

Feng, V. W., & Hirst, G. (2011, June). *Classifying arguments by scheme*. Paper presented at the 49th annual meeting of the Association for Computational Linguistics, Portland, OR.

Kukich, K. (2000). Beyond automated essay scoring. *IEEE Intelligent Systems*, *15*(5), 22–27.

Madnani, N., Heilman, M., Tetreault, J., & Chodorow, M. (2012). Identifying high level organizational elements in argumentative discourse. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 20–28). Montreal, QC: Association for Computational Linguistics.

Mochales, R., & Ieven, A. (2009, June). *Creating an argumentation corpus: Do theories apply to real arguments? A case study on the legal argumentation of the ECHR*. Paper presented at the 12th international conference on Artificial Intelligence and Law, Barcelona, Spain.

Nussbaum, E. M., & Edwards, O. V. (2011). Critical questions and argument stratagems: A framework for enhancing and analyzing students' reasoning practices. *Journal of the Learning Sciences*, *20*, 443–488.

Palau, R. M., & Moens, M. F. (2009). Automatic argument detection and its role in law and the semantic web. *Proceedings of the 2009 Conference on Law, Ontologies and the Semantic Web* (pp. 62–70). Amsterdam, Netherlands: IOS Press.

Powers, D. E. (2005). *"Wordiness": A selective review of its influence, and suggestions for investigating its relevance in tests requiring extended written responses* (Research Memorandum No. RM-04-08). Princeton, NJ: Educational Testing Service.

Rahwan, I., Banihashemi, B., Reed, C., Walton, D., & Abdallah, S. (2011). Representing and classifying arguments on the semantic web. *Knowledge Engineering Review*, *26*(4), 487–511.

Rienks, R., Heylen, D., & Van der Weijden, E. (2005). *Argument diagramming of meeting conversations*. Retrieved from http://www.idiap.ch/workshop/icmi05/pdf/MMMP_paper24.pdf

Song, Y., & Ferretti, R. P. (2013). Teaching critical questions about argumentation through the revising process: Effects of strategy instruction on college students' argumentative essays. *Reading and Writing: An Interdisciplinary Journal*, *26*, 67–90.

Song, Y., Heilman, M., Beigman Klebanov, B., & Deane, P. (2014). Applying argumentation schemes for essay scoring. *Proceedings of the First Workshop on Argumentation Mining* (pp. 69–78). Baltimore, MD: Association for Computational Linguistics.

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

van Eemeren, F. H., & Grootendorst, R. (1992). *Argumentation, communication, and fallacies: A pragma-dialectical perspective*. Mahwah, NJ: Lawrence Erlbaum.

van Eemeren, F. H., & Grootendorst, R. (2004). *A systematic theory of argumentation: A pragma-dialectical approach*. Cambridge, England: Cambridge University Press.

Verbree, D., Rienks, H., & Heylen, D. (2006). First steps towards the automatic construction of argument-diagrams from real discussions. *Proceedings of the 2006 Conference on Computational Models of Argument* (pp. 183–194). Amsterdam, Netherlands: IOS Press.

Walton, D. (1996). *Argumentation schemes for presumptive reasoning*. Mahwah, NJ: Lawrence Erlbaum.

Walton, D., & Macango, F. (2016). A classification system for argumentation schemes. *Argument and Computation*, *6*, 219–245.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. New York, NY: Cambridge University Press.

Suggested citation:

Song, Y., Deane, P., & Beigman Klebanov, B. (2017). *Toward the automated scoring of written arguments: Developing an innovative approach for annotation* (Research Report No. RR-17-11). Princeton, NJ: Educational Testing Service. https://dx.doi.org/10.1002/ets2.12138

Action Editor: Donald Powers

Reviewers: Jill Burstein and Brian Riordan