Measuring the Power of Learning.®

Research Report
ETS RR–17-16

# Examining Students' Ability to Critique Arguments and Exploring the Implications for Assessment and Instruction

Yi Song

Paul Deane

Mary Fowles

August 2017

# ETS Research Report Series



Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

# Examining Students' Ability to Critique Arguments and Exploring the Implications for Assessment and Instruction

Yi Song, Paul Deane, & Mary Fowles

Educational Testing Service, Princeton, NJ

This study assessed 1,706 eighth-grade students' reasoning abilities through a critique task in an integrated argumentative reading and writing test. Results indicate that the majority of students did not detect fallacious arguments or clearly explain problems in the arguments. They encountered various challenges in critiquing an argument: (a) being off-task, (b) failing to identify fallacious arguments, (c) having difficulty explaining specific problems in reasoning, and (d) not connecting their criticisms with particular parts of the text being critiqued. Task design and instructional implications are discussed.

**Keywords** Argument critique; reasoning; fallacy; assessment

doi:10.1002/ets2.12166

Thoughtful, well-reasoned debate and argumentation skills are the cornerstone of a lively democracy. With the publication and multistate endorsement of the Common Core State Standards (CCSS), American schools are revamping their curricula to address the increased emphasis on argumentative reading and writing (Council of Chief State School Officers & National Governors Association, 2010). Yet argumentation is a complex set of skills that can be challenging to teach, and studies show that many students are ill prepared for the expectations of college and career (Graff, 2003; Hillocks, 2002; Nation's Center for Education Statistics, 2008). To help students achieve college and career readiness, we have been exploring how best to support teachers as they guide students to higher levels of presenting and critiquing arguments by providing assessment tasks that unpack and then scaffold this complex practice. The work described in this report, part of the *CBAL*® learning and assessment tool research initiative, is intended to measure students' ongoing progress in argumentation skills while providing informative feedback for instruction and worthwhile learning experiences for students (Bennett, 2010).

The design and development of CBAL argumentative reading and writing assessments are based on a cognitive competency model that incorporates curriculum standards with the findings of learning sciences research. This competency model specifies what students should know and be able to do at various stages as they progress toward college and career mastery of argumentation skills and is therefore stated in the form of a learning progression for argumentation skills (Bennett, Deane, & van Rijn, 2016; Deane & Song, 2014, 2015; Van Rijn, Graf, & Deane, 2014).

This publication reports the results of a study focusing on one component of argumentation: critique. In the following sections, we present the research background and rationale, describe the study method and results, and discuss implications for assessment and instruction.

## Background and Rationale

In written arguments, people should clearly express their own opinions on controversial issues, justify their standpoints by providing reasons and evidence, and critically respond to alternative standpoints (Ferretti, MacArthur, & Dowdy, 2000; van Eemeren, Grootendorst, & Henkemans, 1996). Unfortunately, students often fail to include critical argumentative elements or present them clearly, to provide sufficient evidence to support their arguments, or to respond to alternative viewpoints (Ferretti, Lewis, & Andrews-Weckerly, 2009; Ferretti et al., 2000). When given opportunities to revise their essays, students tend to make changes irrelevant to the meaning of the text, unconnected to genre considerations, and as a result, these revisions often do not help improve the quality of writing (Scardamalia & Bereiter, 1986). Although research

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

has indicated that the ability to revise develops over time (Fitzgerald & Markman, 1987), many college students are still unable to improve their arguments through revision (Kinsler, 1990). Song and Lavigne (2006) found that feedback from both college students and eighth-grade students focused on surface-level errors, such as spelling, grammar, punctuation, and capitalization, instead of suggesting how to improve the arguments in terms of their claims, reasons, and supporting evidence.

Another common problem is that students often presume that information in an argument is true rather than asking questions that would reveal potential fallacies. For example, students may use quotations from books, arguments heard on TV programs, or opinions posted on websites without evaluating whether the information is adequately supported by evidence. Students might make hasty generalizations based on their personal experience, insensitive to the limitations of these examples. Students who are unable to distinguish reasonable from fallacious arguments are unlikely to make appropriate decisions about whether they should accept or reject an argument. Several studies (Ferretti et al., 2009; Nussbaum & Edwards, 2011; Song & Ferretti, 2013) have suggested that students could benefit from learning how to evaluate arguments against appropriate criteria.

To detect whether an argument is misused or fallacious, it is important to understand the underlying argumentation strategies, also called argumentation schemes (Walton, 1996). These schemes represent the relationship between what is stated in the claim and its supporting justificatory structure (van Eemeren & Grootendorst, 1992; van Eemeren, Grootendorst, & Henkemans, 2002; Walton, 1996; Walton, Reed, & Macagno, 2008). Take, for example, the argument from expert opinion scheme. The writer states that some purported authority has supported a particular claim. An opponent could respond in several ways—by asking for evidence that the so-called authority really is an expert, by asking for proof that the expert's statements really do support the claim, or by questioning whether the person's expertise actually matters. Each of these responses corresponds to a critical question in the argument from expert opinion scheme (Walton, 1996). Consider another argumentation scheme, argument from example. When someone offers an example to support a claim, the opponent might ask the following questions (Walton, 1996): (a) Is the example true? (b) Is it a relevant example of the general statement we are being asked to believe? (c) Is this example typical of the kinds of cases that the general statement covers? (d) Are there any special circumstances that could undermine the generalization from this case to other cases? If not well answered, these critical questions could elicit strong counterarguments. Thus, asking scheme-relevant critical questions helps us differentiate fallacious arguments from reasonable arguments by encouraging consideration of unwarranted assumptions in people's reasoning (van Eemeren & Grootendorst, 1992; van Eemeren et al., 2002; Walton, 1996; Walton et al., 2008).

In light of the argumentation scheme theory, Nussbaum and Edwards (2011) taught seventh graders a simplified version of critical questions selected from Walton (1996) in combination with integrative and refutation stratagems (specific argument types expressed in discourse). They found that students in the experimental group made more arguments that integrated both sides of an issue, demonstrating considerations for comparing values of the two sides and developing solutions that address these practical concerns. They concluded that applying the evaluative criteria, informed by critical questions in argumentative discourse, leads to better argumentation.

Another intervention study using critical questions focused on college students. Song and Ferretti (2013) taught 30 college students two commonly used argumentation schemes (i.e., argument from consequences and argument from example) and critical questions associated with these schemes. Compared to students in the contrasting conditions, those who learned critical questions wrote essays that were of higher quality and included more counterarguments, alternative standpoints, and rebuttals. These findings indicate that strategy instruction that includes critical standards of argumentation increases college students' sensitivity to alternative perspectives and improves their ability to critique arguments.

Whereas these prior studies used critical questions to prompt students to justify their own arguments in essay tasks, we were interested in examining middle school students' ability to critique other people's arguments. Recognition of typical reasoning flaws (e.g., hasty generalization, post hoc, and bandwagon) is normally taught during middle and high school; therefore, informed by curriculum standards and learning sciences research, we took a further step to examine how well students can apply these skills and, in their own words, explain reasoning errors in people's arguments. Critiquing arguments is a complex skill. It requires students to critically examine the argument—including reasoning flaws, unwarranted claims, and assumptions—and then to clearly explain how these features have weakened the argument. A high-quality argument critique should also demonstrate thoughtful considerations of critical questions.

In this report, we use the data from a critique task in a summative assessment to answer the following research question: How well do eighth graders critique arguments, and in particular, what major challenges do they encounter in undertaking a critique task? An answer to this question can potentially provide information that can be used to improve instruction in critiquing arguments.

## Method

### Participants

Participants in this study were students from a volunteer sample of 35 schools in 20 states. A total of 1,706 eighth-grade students were included in the analysis. Based on the available demographic data, roughly equal numbers of male and female students participated. The majority of students (61%) were Caucasian; other groups included Hispanic (15%), African American (11%), and Asian (8%). More than one third of the students were from low-income families, receiving free or reduced-price lunch. According to self-reported data, roughly 4% of the students were either current or former English learners. Tables 1 and 2 present the demographic information for this sample.

### Instruments

This study drew upon a CBAL scenario-based Grade 8 test of argumentation skills. Students engage in a sequence of steps in which they research the issue of whether the United States should ban advertising to children under the age of 12 and, in the culminating task, write essays to present and support their positions on this issue. The Ban Ads test consists of four tasks. In the first task, students become informed about the issue by reading and summarizing both pro and con articles about banning ads aimed at children. The second task requires students to organize a list of statements according to which position (pro or con) each statement supports and then to identify the relationship between claims and evidence (i.e., does the evidence support a claim, weaken it, or is it irrelevant?). In the third task (see Figure 1), students are asked to review a letter to the editor. In doing so, they have to write a critique of its argument (i.e., identify and explain problems in the reasoning or use of evidence and point out any inaccurate information). Finally, students write an essay for their local newspaper to state and support their opinion on this issue. The analysis in this report focuses on Task 3, the critique.

In the critique task, students are asked to evaluate the arguments in a letter to the editor. This text has four major reasoning flaws. First, the writer overgeneralizes an observation about a single family. The fact that one family watches TV together and learns a lot does not strongly support the conclusion that advertising for children is always a good thing. The family may "learn a lot" from the programs they watch, not the ads, and in any case, a sample of one is not strong support. Second, the writer reasons unrealistically that companies spend billions of dollars on children's advertisements, and therefore the ads are good for kids. Presumably, companies spend such money because they want to sell their products. Third, the writer misrepresents arguments against advertising aimed at children. People want to ban ads because they are harmful to children, not because they are uninteresting. Finally, the writer provides what is essentially a circular argument (that children's advertisements should not be banned because they are allowed to be shown on TV). The letter also contains two pieces of inaccurate information, which students should be able to recognize by examining the source texts they have already read earlier in the assessment (and which are available to them when they work on the critique task). The writer claims that children cannot remember commercials, but the source texts reported conflicting results from research. In addition, it is not true that some countries banned ads because children thought that the ads were funny.

The argument critique task was designed to be the most challenging part of this eighth-grade assessment. We included this task in the pilot test for two reasons: to discriminate at the upper end of the score distribution and to collect a wide range of samples so that we could identify ways to provide additional support for low-performing students.

In scoring students' critiques, the following aspects were considered: (a) whether students identified and clearly explained most of the major problems in the letter's reasoning, (b) whether students pointed out inaccurate information,[1] and (c) whether students expressed ideas in an appropriate tone for the class. The overall quality of each critique was rated on a scale from 0 to 4. Five research assistants received training and practice sessions in using the scoring rubric (see Figure 2) and were provided with anchor responses and topic notes that listed the valid attack points. Four of these

**Table 1** Test Sample Distribution by Location Demographics

| | Ban Ads | |
|---|---|---|
| Demographics | *N* | % |
| State | | |
| AL | 120 | 7 |
| AR | 15 | 1 |
| AZ | 17 | 1 |
| CA | 181 | 11 |
| CO | 48 | 3 |
| GA | 63 | 4 |
| KY | 56 | 3 |
| MA | 27 | 2 |
| MI | 147 | 9 |
| MN | 52 | 3 |
| MS | 41 | 2 |
| NJ | 161 | 9 |
| NY | 118 | 7 |
| OH | 138 | 8 |
| PA | 43 | 3 |
| SC | 134 | 8 |
| SD | 42 | 3 |
| TN | 32 | 2 |
| TX | 250 | 15 |
| WI | 21 | 1 |
| Region | | |
| East | 349 | 21 |
| Midwest | 471 | 28 |
| South | 640 | 38 |
| West | 246 | 14 |
| Locale | | |
| Rural | 730 | 43 |
| Suburban | 561 | 33 |
| Urban | 415 | 24 |

*Note.* Owing to rounding, percentages may not sum to 100.

raters independently scored 25% of the total number of the critiques. A fifth rater scored 20% of the total number of the critiques to check the interrater agreement. Their exact agreement (same scores) was 67%, and agreement within 1 point (adjacent scores) was 96%. The quadratically weighted kappa coefficient was .84. The scores provided by the first four raters were used in this analysis.

To further understand problems in students' critiques, we randomly selected 100 students' responses, in proportions determined by the overall score distribution. Specifically, in this small sample, 35 responses were scored a 0 (no credit), 20 responses received the score of 1 point, 28 responses had a score of 2 points, 13 responses had received the score of 3 points, and 4 responses earned the full credit of 4 points. We analyzed the content of these responses qualitatively in an effort to understand what argument-evaluation skills students can already perform and what challenges they still face.

## Results

The mean score of students' critiques was 1.29 ($SD = 1.18$) on a 4-point scale. Roughly 16% of the students achieved 3 points or above, and more than one third of the students received a 0 (see Table 3 for the score distribution). The results indicate that the majority of eighth graders in this sample may not yet be able to provide a successful critique of the letter's argument, as illustrated in Table 3. We used weighted scores when calculating correlations among tasks (Task 1, summary score weight = 1; Task 2, analysis score weight = 1; Task 3, critique score weight = 2; Task 4, essay score weight = 3; see Fu & Wise, 2012, for details). The results show that the critique task score is moderately correlated[2] with all the other subtasks ($r = .54, .52, .56$ with Task 1, Task 2, and Task 4, respectively), $p < .001$ for all these cases, and that the critique task score is strongly correlated with the Ban Ads total score ($r = .76$), $p < .001$. These results indicate that poor performance on the

**Table 2** Student Demographics

| | Ban Ads | |
|---|---|---|
| Demographics | *N* | % |
| Gender | | |
| M | 817 | 48 |
| F | 802 | 47 |
| Unreported | 87 | 5 |
| Race | | |
| African American | 193 | 11 |
| Asian/Pacific Islander | 132 | 8 |
| Hispanic | 260 | 15 |
| Native American | 2 | 0 |
| White | 1,032 | 61 |
| Unreported | 87 | 5 |
| Low socioeconomic status | | |
| N | 882 | 52 |
| Y | 605 | 36 |
| Unreported | 219 | 13 |
| English language learner (ELL) | | |
| Current ELL (Y) | 35 | 2 |
| Former ELL (N) | 31 | 2 |
| English Proficient (N) | 1,335 | 78 |
| Unreported | 305 | 18 |

*Note.* Owing to rounding, percentages may not sum to 100.

**Table 3** Critique Task Raw Score Frequency and Percentage

| Score | Frequency | % |
|---|---|---|
| 0 | 608 | 36 |
| 1 | 338 | 20 |
| 2 | 483 | 28 |
| 3 | 215 | 13 |
| 4 | 62 | 4 |
| Total | 1, 706 | 100 |

*Note.* Owing to rounding, percentages may not sum to 100.

**Table 4** Ban Ads Task Mean Scores After Score Weights Were Applied

| | Maximum possible score | *N* | Mean | *SD* | Correlation with Ban Ads total score (*r*) |
|---|---|---|---|---|---|
| Task 1: Summary | 11 | 1,669 | 3.85 | 1.89 | .70 |
| Task 2: Analysis | 16 | 1,702 | 11.43 | 2.92 | .73 |
| Task 3: Critique | 8 | 1,706 | 2.58 | 2.36 | .76 |
| Task 4: Essay | 30 | 1,475 | 14.30 | 6.15 | .91 |

*Note. r* values for all these tasks are significant at the .001 level.

critique task was characteristic of low-performing students overall. Table 4 presents the mean scores (after score weights are applied) of the Ban Ads tasks and the correlation between each task score and the total score.

We also examined the amount of time students spent on the critique task. On average, students spent 5.32 minutes ($SD = 2.52$) responding to the task, but the time range varied dramatically, from 6 seconds to 10 minutes. The time spent on the task was moderately correlated with the score ($r = .44$).

In the selected sample of 100 students,[3] 35 students received no credit. Among them, 17 students reacted to the critique task as if they were asked to express their own opinion on the Ban Ads issue. For example, one student wrote, "People shouldn't take advantage of children under the age of twelve." Another student responded, "I think they should keep ads. It should be the parents to teach their children about things they sell on TV." Some of these responses were

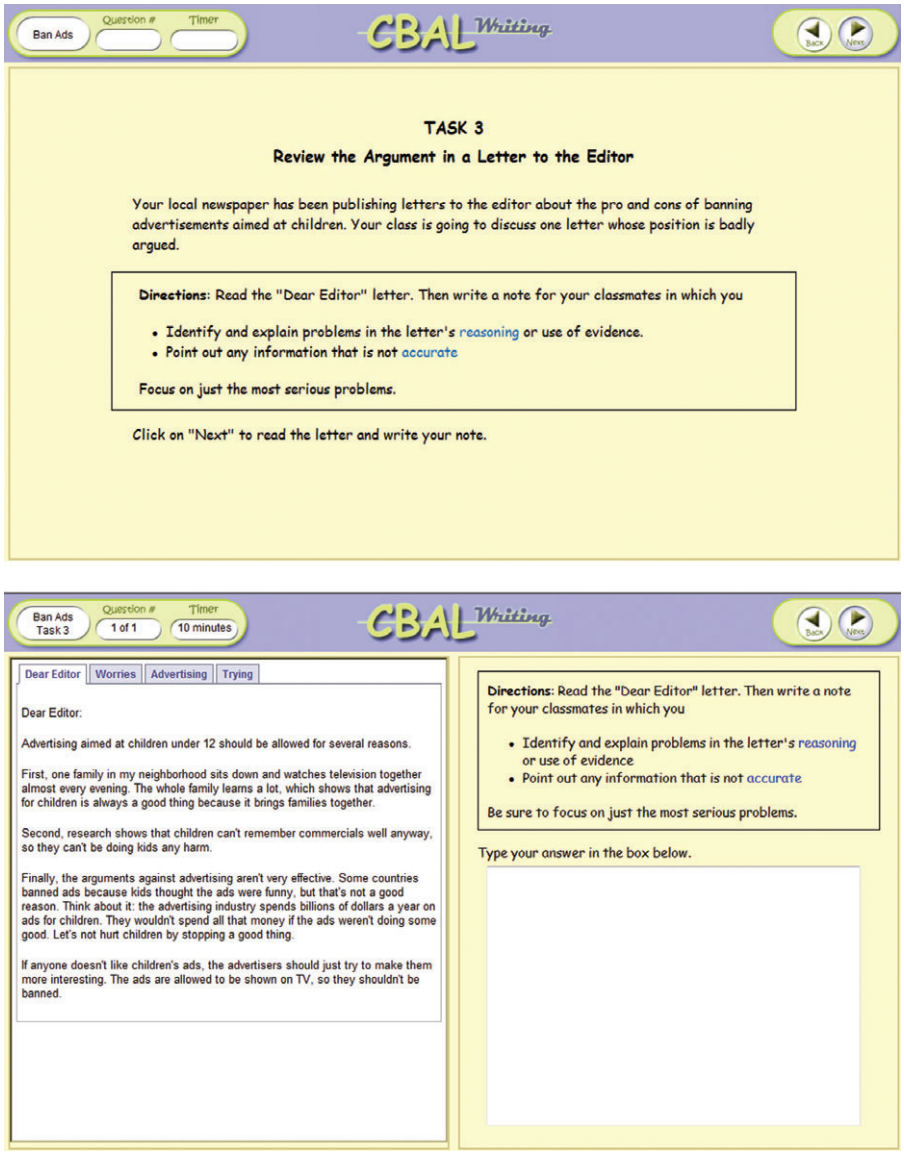


**Figure 1** Ban Ads Task 3—Review the argument in a letter to the editor.

minimal; others were more elaborated because students added reasons to support their own opinions on this issue, for instance,

> I support the summary because children at the age of 12 will fully understand the purpose of advertising. They will know the difference of good and bad choices. Their parents will also help. Ads were made to give someone a good thing to use.

This student agreed with the letter's arguments and offered three reasons to support that position. However, the critique task required students to evaluate the logical soundness of the arguments, not to explain their own positions on the issue.

Another six students received no credit because they responded to the critique task by summarizing or simply copying the major arguments in the letter. For example, one of them wrote,

> Children under 12 should be allowed to watch ads because one family does and is fine. Also most kids can't remember them. The advertisement industries spend millions of dollars and wouldn't spend that money if it wasn't doing some good.

| "BAN ADS" TASK 3<br>Critique a "Dear Editor" Letter for Your Class | |
|---|---|
| **4**<br>**Excellent** | An "Excellent" Critique<br>• Identifies and clearly explains most of the major problems in the letter's reasoning,<br>• Points out inaccurate information, AND<br>• Expresses ideas in a clearly appropriate tone for the class |
| **3**<br>**Adequate** | An "Adequate" Critique<br>• Identifies and explains some of the major problems in the letter's reasoning and inaccurate information, OR<br>• Identifies only one major problem in the letter's reasoning but explains it extremely well, AND<br>• Expresses ideas in a generally appropriate tone for the class |
| **2**<br>**Limited** | A "Limited" critique *identifies at least one major problem* in the letter's reasoning and/or use of inaccurate information but is limited in one or more of the following ways:<br>• Explains the problem(s) poorly, if at all<br>• Misinterprets parts of the letter or includes irrelevant information<br>• Misinterprets an important part of the task<br>• Expresses ideas in a somewhat inappropriate tone for the class |
| **1**<br>**Minimal** | A "Minimal" response *identifies or implies a problem* in the letter's reasoning and/or use of accurate information but displays one or more of the following problems:<br>• Is very confusing<br>• Seriously distorts the letter or includes mostly irrelevant information<br>• Seriously distorts the writing task<br>• Expresses ideas in a highly inappropriate tone for the class |
| **0**<br>**No Credit** | A response receives "No Credit" for any one of the following reasons:<br>• Identifies no problems in the letter's reasoning and/or use of inaccurate information<br>• Not long enough for critical-thinking skills to be judged<br>• Not written in English<br>• Off topic<br>• Blank<br>• Only random key strokes |

**Figure 2** Scoring rubric.

Another four students received no credit because their responses expressed pure agreement, as in this example: "I do not see any reasoning flaws in the letter. The information in the letter is accurate."

A few students who received no credit attempted to evaluate the letter, but their attempt was unsuccessful. One reason for this failure is that they only gave a brief and generic response. Examples include the following: "The writer was not that accurate with the letter" and "The arguments against advertising aren't very effective." These statements were too general to be meaningful. The students did not tell which part in the letter was not accurate or which arguments are not effective. Other students who received no credit focused on surface features of the letter, for example, "One problem is one paragraph it only has 2 sentences for the paragraph." This student evaluated the letter based on the number of sentences in a paragraph, and that was the only criterion used.

The rest of the no-credit responses merely copied the original text. Overall, most no-credit responses reflected a misunderstanding of the task, suggesting that many of these eighth-grade students were unfamiliar with the idea of critiquing someone else's argument and thus failed to establish appropriate goals for the critique task.

Students who earned 1 point on this task were able to identify at least one piece of inaccurate information or one reasoning flaw, but they did not explain the problems clearly. A majority of the students who received 1 point correctly identified the inaccurate statement that children cannot remember commercials, but their responses were very short and provided no further explanation, whereas students with higher scores usually referred back to the articles that provide the counterevidence. Students at this performance level often included irrelevant information, distorted the information in the letter, and/or displayed a little understanding of the goal of the critique task. For example, one student wrote,

> Advertising aimed at children under 12 should be allowed for several reasons. Those people think that the children that are 12 so be able to watch the TV advertisers because they think by the age of 12 they will understand it. One that is not accurate is some countries banned ads because kids thought the ads were funny, but that's not a good reason.

This student identified the inaccurate information about why some countries banned children's advertisements, but she did not explain why this was not a good reason. Furthermore, the first two sentences had nothing to do

with the arguments in the letter, suggesting that the student still thought the primary task was to support her own opinion.

> At score 1, student responses were often generic, for example,
> The letter is not accurate and the reasoning is bad. Like the author of the letter is using a whole lot of his own observation and not using evidence. Even when he is using evidence he tended to twist it around.

This student did not specify where the author used his own observation rather than evidence or how he twisted the evidence around. It is interesting that this student understood the importance of evidence in support of arguments but failed to provide any himself.

The challenge of making specific connections to the letter was also evident in the response that met the criteria for a score of 2 or 3. They sometimes included a relatively generic statement, such as, "Another flaw in the article is that the author made statements that were not very relevant to the topic and did not have proper conclusions drawn from then." In this excerpt, the student did not specify which statements in the article were irrelevant to the topic, nor why they were irrelevant.

Of the 100 students in the selected sample, only 30 explicitly identified at least one reasoning flaw. For example, one student challenged the author's flaw of overgeneralizing from a single example: "The letter suggests that commercials bring his neighbor's family together, but just because it brings them closer doesn't necessarily mean that it brings every family closer together." Another way to challenge the example was to focus on its accuracy or relevancy, as in this student response: "Sometimes advertising for children isn't always a good thing. Just because a whole family is watching television doesn't mean the kids are paying attention to the commercials. I believe that advertisements should be banned for children under twelve." This student identified the unwarranted assumption that the family is watching commercials, not other television programs. Both of the preceding responses correctly identified a specific reasoning flaw, yet they focused on only a single argument in the letter. The most successful critiques (earning the top score of 4) identified multiple problems, explained the problems, and provided textual evidence to help pinpoint where the problems occurred, as in this response:

> In the first paragraph, you were getting off topic. The whole paragraph is based around "watching television brings families together" which is not relevant to the topic at hand. Secondly, the support in the first paragraph is weak. Just assuming that all families that watch TV learn lots of things is not relevant to all families. In your second paragraph, you said that children can't remember commercials well anyway, which is not correct. According to the American Psychological Association, children can recall the message and want the product. I would have checked your sources carefully. Lastly, your last paragraph is also incorrect as well. Most countries don't ban ads because the children thought they were funny, they banned them because they thought they were misleading and harmful to a child's development.

In summary, the overwhelming majority of eighth-grade students (83%) in this sample failed to demonstrate a clear understanding of the critique task or control of the necessary supporting skills required for a score of 3 or 4. A close examination of 100 students' responses revealed several characteristic patterns of difficulty: (a) being off-task; (b) failing to identify fallacious arguments; (c) having difficulty explaining specific problems in reasoning; and (d) not connecting their criticisms with particular parts of the text being critiqued, resulting in highly generic responses that could have been written in response to almost any stimulus.

## Discussion

The ability to read, write, and analyze arguments is an essential academic skill across many subject areas and a focus in tests that help in making high-stakes decisions (Graff, 2003). For instance, the CCSS English language arts writing standards emphasize writing logical arguments in which students should demonstrate sound reasoning and use relevant evidence (Council of Chief State School Officers & National Governors Association, 2010). Despite the importance of argumentation, argumentation skill is not well developed by middle school. In our data, consistent with the existing literature, approximately 17% of the eighth-grade students wrote an adequate critique. Many of them did not seem to understand the requirements for writing such a critique. They simply offered reasons to support their own opinions on

whether advertisements that target children should be banned, summarized the letter, or gave only generic responses without using any textual evidence. Other students were able to identify inaccurate information, but very few eighth-grade students in our sample demonstrated the ability to detect fallacious arguments and clearly explain problems in reasoning. One possible explanation is that middle school students may be just beginning to learn some common reasoning errors, while other students may not be introduced to fallacious arguments or evidence-based justification until high school, so writing a critique might be a completely new task to some students.

One way to enhance students' skills in critiquing arguments is to incorporate scaffolding components into the assessment, thus breaking this complex task into easier, more "doable" steps. This approach identifies which component skills in argumentation students have already mastered and which they are still learning. For example, the scaffolding could help reveal whether students understand the task requirement, know how to apply effective strategies for analyzing arguments, and can then write a coherent critique of the arguments. Furthermore, students might perform the task better if they saw a model of a successful critique (of a different text on the same topic) before they write their own. Such scaffolding may provide the support and structure necessary for students to learn how to write a critique or complete the task successfully.

For this reason, in our ongoing research, we are designing seven new selected-response (SR) items as a scaffolded lead-in activity. In the context of writing a critique, one of the key skills is to identify reasoning errors in people's arguments. Therefore, the new SR items focus on different types of reasoning errors. Each SR item presents a common reasoning flaw: Items 1 – 5 require students to select an option that correctly explains the reasoning error in a given argument, and Items 6 and 7 focus on framing a case and ask students to select appropriate connective words in a brief critique (see Figure 3 for sample items). The SR items provide examples that model how to critique arguments with various reasoning flaws. By guiding students through these prerequisite skills, the SR items prepare students to write their own critiques, while assessing skills not easily measured in an open-ended critique task.

In addition to scaffolded lead-in tasks, middle school students might benefit from direct instruction in how to detect various reasoning flaws and unwarranted assumptions because they are still in the process of developing an epistemological understanding of argumentation and may not yet be able to assume the perspective of an objective evaluator (Kuhn, 2009; Weinstock, Neuman, & Tabak, 2004). Typically, this high-order, argumentative thinking skill does not develop naturally (Kinsler, 1990; Song & Ferretti, 2013). We believe that teaching students effective strategies would significantly improve their ability to evaluate the strengths and weaknesses in an argument, as indicated in existing studies (Nussbaum & Edwards, 2011; Song & Ferretti, 2013). A strategy that incorporates critical questions could be helpful in that the questions would function as criteria for evaluating arguments.

Here is an illustration of how a strategy of asking critical questions can help identify the weaknesses in the arguments and generate potential counterarguments in the CBAL critique task. The author of the letter offers an example (i.e., "one family in my neighborhood sits down and watches television together almost every evening") to support the claim that advertising brings families together. After identifying the argumentation scheme (i.e., argument from example), we could ask critical questions associated with this scheme (Walton, 1996). For instance, is this a relevant example? (Is the example of family watching TV together relevant to the argument that advertising for children is a good thing?) This example might not be relevant to the argument if the family is watching a TV series or a sports program instead of commercials. Another question we could ask: Is this a typical example? (Do other families watch TV together every evening?) This example might not be a typical case because not every family is watching TV every evening. Each family may have its own activities after dinner, such as taking a walk, playing board games, or reading some books. It is dangerous to make a generalization based on a single family's activity.

Future research should explore how to effectively support students' learning of argumentation schemes and critical questions. One approach taken by Nussbaum and Edwards (2011) was to give middle school students a few modified critical questions that were easy to understand, free of difficult logical terms. In addition, most of the questions were broad or general and therefore could be asked in many different circumstances, as they do not require knowledge of argumentation schemes (e.g., *What is the likelihood? How important? Why? So what? How do you know?*). Although Nussbaum and Edwards's approach helped reduce the complexity of learning about critical questions, it is important to recognize that knowledge about one argumentation scheme may not transfer to another scheme, especially when each scheme has its own set of critical questions. Therefore the goal of our future research is to answer this question: If students are taught commonly used argumentation schemes, will it improve their ability to evaluate an argument by focusing on its reasoning?

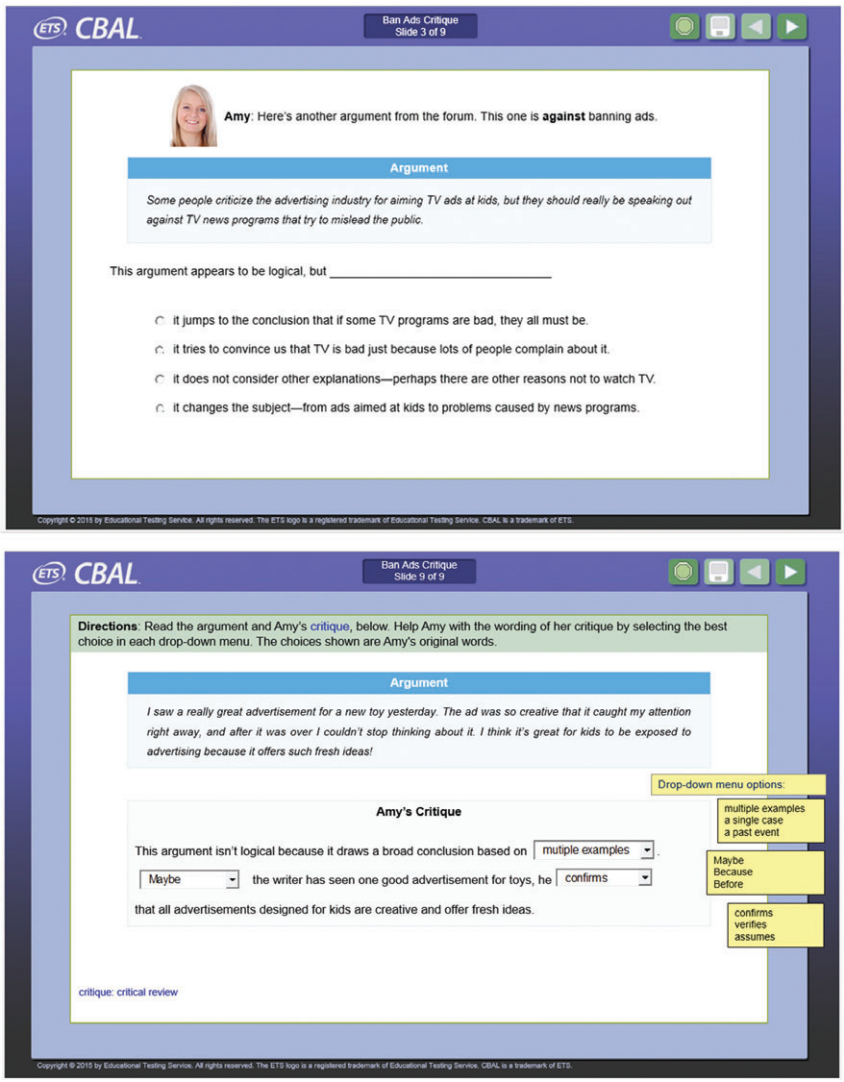


**Figure 3** Examples of new SR critique items for Ban Ads.

Note that this study has a number of limitations that should be taken into account when interpreting the results. First, although we had a relatively large sample, it was collected from a convenience sample and so cannot be considered representative of the U.S. middle school population as a whole. Second, the scoring rubric was defined by qualitative judgment based upon internal research at Educational Testing Service on argument-evaluation skills and not by a conventional standard-setting process; thus, there is no direct link between the score rubrics and expectations about performance on eighth-grade standards (either in a particular state or on the CCSS). Third, we did not examine how student motivation impacted student responses. As this was a low-stakes test, some students might not have taken it seriously; in fact, the brief time some students spent on the task may suggest that they did not try to respond but simply clicked through the questions. Fourth, this was the only task in the Ban Ads test that assesses the ability to perform critiques, so it is entirely possible that students who performed poorly on this task might have done better on a more extensive sequence of argument critique tasks. Designing and pilot testing a variety of tasks that target the critique skills should help us better understand the challenges students are facing. This point is reinforced by the large percentage of students who wrote off-task responses, suggesting that we might have elicited more successful critiques by carefully structuring the task to help students understand task goals and expectations. Each of these limitations in the study suggests lines of research that may yield more effective ways to design assessments focused on students' ability to critique flawed arguments.

Overall, however, the weak performance students displayed on the written critique task is consistent with developmental expectations. Only the top-performing students were able to meet task expectations. Students who performed less well on other argument tasks were the most likely to be off-task or to put in a minimal performance that failed to address the logical flaws in the argument. We hope in future work to examine the extent to which this characteristic pattern of performance can be modified by appropriate scaffolding and directions or by targeted interventions designed to raise students' argumentation skills.

## Acknowledgments

This research was funded by the CBAL research initiative at Educational Testing Service. Thanks to Randy E. Bennett and Tenaha O'Reilly for their helpful suggestions and edits on an earlier version of this report. We are grateful to Ayleen Gontz for her editorial assistance.

## Notes

1 Students can click on the tabs to see articles that provide research-based information about advertisements for children.
2 Strength of correlation as suggested in Evans (1996): <.20, very weak; .20 – . 39, weak; .40 – .59, moderate; .60 – .79, strong; >.79, very strong.
3 Note that the spelling and some grammatical errors in the quoted responses have been corrected given that these aspects were not taken into account in scoring, but the content has been left unchanged.

## References

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*, 70 – 91.

Bennett, R. E., Deane, P., & van Rijn, P. (2016). From cognitive-domain theory to assessment practice. *Educational Psychologist, 51*, 1 – 26.

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*. Retrieved from http://www.corestandards.org/the-standards/ELA-Literacy

Deane, P., & Song, Y. (2014). A case study in principled assessment design: Designing assessments to measure and support the development of argumentative reading and writing skills. *Spanish Journal of Educational Psychology, 20*, 99 – 108.

Deane, P., & Song, Y. (2015). *The key practice, "discuss and debate ideas": Conceptual framework, literature review, and provisional learning progressions for argumentation* (Research Report No. RR-15-33). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12079

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences.* Pacific Grove, CA: Brooks/Cole.

Ferretti, R. P., Lewis, W. E., & Andrews-Weckerly, S. (2009). Do goals affect the structure of students' argumentative writing strategies? *Journal of Educational Psychology, 101*, 577 – 589.

Ferretti, R. P., MacArthur, C. A., & Dowdy, N. S. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology, 93*, 694 – 702.

Fitzgerald, J., & Markman, L. (1987). Teaching children about revision in writing. *Cognition and Instruction, 4*, 3 – 24.

Fu, J., & Wise, M. (2012). *Statistical report of 2011 CBAL™ multistate administration of reading and writing tests* (Research Report No. RR-12-24). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02306.x

Graff, G. (2003). *Clueless in academe: How schooling obscures the life of the mind.* New Haven, CT: Yale University Press.

Hillocks, G., Jr. (2002). *The testing trap: How state writing assessments control learning*. New York, NY: Teachers College Press.

Kinsler, K. (1990). Structured peer collaboration: Teaching essay revision to college students needing writing remediation. *Cognition and Instruction, 7*, 303 – 321.

Kuhn, D. (2009). The importance of learning about knowing: Creating a foundation for development of intellectual values. *Child Development Perspectives, 89*, 112 – 117.

Nation's Center for Education Statistics. (2008). *The nation's report card: Writing 2007* (U.S. Department of Education Publication No. NCES 2008-468). Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2007/2008468.pdf

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

Scardamalia, M., & Bereiter, C. (1986). Research on written composition. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 778–803). New York, NY: Macmillan.

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., & Lavigne, N. C. (2006, April). *Effect of feedback from reviewers at different levels on 8th-grade students' revision and persuasive writing*. Poster presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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., & Henkemans, F. S. (1996). *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Mahwah, NJ: Lawrence Erlbaum.

van Eemeren, F. H., Grootendorst, R., & Henkemans, F. S. (2002). *Argumentation: Analysis, evaluation, presentation.* Mahwah, NJ: Lawrence Erlbaum.

van Rijn, P. W., Graf, E. A., & Deane, P. (2014). Empirical recovery of argumentation learning progressions in scenario-based assessments of English language arts. *Spanish Journal of Educational Psychology, 20*, 109–115.

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

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

Weinstock, M., Neuman, Y., & Tabak, I. (2004). Missing the point or missing the norms? Epistemological norms as predictors of students' ability to identify fallacious arguments. *Contemporary Educational Psychology, 29*, 77–94.

## Suggested citation:

Song, Y., Deane, P., & Fowles, M. (2017). *Examining students' ability to critique arguments and exploring the implications for assessment and instruction* (Research Report RR-17-16). Princeton, NJ: Educational Testing Service https://doi.org/10.1002/ets2.12166

**Action Editor:** Heather Buzick

**Reviewers:** Randy Bennett and Tenaha O'Reilly