Exploring the Effect of a Scaffolding Design on Students' Argument Critique Skills

YI SONG

Educational Testing Service
660 Rosedale Rd, Princeton, NJ
USA
ysong@ets.org

YIGAL ATTALI

Educational Testing Service
660 Rosedale Rd, Princeton, NJ
USA
yattali@ets.org

SZU-FU CHAO

Educational Testing Service
660 Rosedale Rd, Princeton, NJ
USA
schao@ets.org

Abstract: In this project, we examined the impact of scaffolded tasks on middle school students' argument critique skills. The study results showed a small positive impact of the scaffolding on student performance on one controversial issue, but not the other, indicating that student skills of writing critiques could be affected by the topic and argument content. Additionally, students from low-SES families did not perform as well as their peers. Student performance on the critique tasks had moderate or strong correlations with students' state reading and writing test scores. Implications of the scaffolding and critique task design are discussed.

Résumé: Dans ce projet, nous avons examiné l'impact des tâches échafaudées sur les capacités des élèves de septième et huitième année de critiquer des arguments. Les résultats de l'étude ont montré un léger impact positif de l'échafaudage sur les performances des élèves sur une question controversée, mais pas sur l'autre, indiquant que les compétences des élèves en matière de rédaction de critiques pourraient être affectées par le sujet et le contenu de l'argumentation. De plus, les élèves issus de familles à faible statut socio-économique n'ont pas été aussi performants que leurs pairs. Les performances des élèves sur les tâches de critique avaient des corrélations modérées ou fortes avec les résultats des tests de lecture et d'écriture des élèves. Les implications de l'échafaudage et de la conception des tâches de critique sont discutées.

Keywords: Argument critique, assessment, fallacy, reasoning, scaffolding

1. Introduction

To become open-minded and critical readers and listeners, students should learn not only to understand what an author or speaker is saying, but also learn to question the assumptions, premises, and reasoning in the text or speech to determine whether the claim is true and well supported. In the U.S. educational system, the skills of making logical arguments and using relevant evidence are greatly emphasized. For example, students are expected to "Delineate and evaluate the argument and specific claims in a text, assessing whether the reasoning is sound and the evidence is relevant and sufficient; recognize when irrelevant evidence is introduced."[1] Evaluating arguments is a challenging skill because students must recognize reasoning flaws and identify specific points in a text that are vulnerable to objections and counterarguments. This goes beyond keeping track of which reason supports which point.

Argumentative writing studies show that students often fail to include relevant evidence, consider alternative perspectives, or critically evaluate others' arguments (e.g., Ferretti, MacArthur and Dowdy 2000; National Center for Educational Statistics 2012; Nussbaum and Kardash 2005; Nussbaum and Edwards 2011; Song and Ferretti 2013). Students often presume that information in a given argument is true or valid, rather than asking questions that would reveal reasoning flaws (e.g., Song and Ferretti 2013). When students have prior misconceptions, they are likely to hold onto them and ignore counter-evidence (Kuhn, Cheney and Weinstock 2000). When writing, many students draw a quick conclusion based on their personal experience and remain insensitive to the limitations of these examples. Not surprisingly, 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 (e.g., Ferretti, Lewis and Andrews-Weckerly 2009). Ferretti and his colleagues pointed out the im-

[1] http://www.corestandards.org/ELA-Literacy/RI/8/

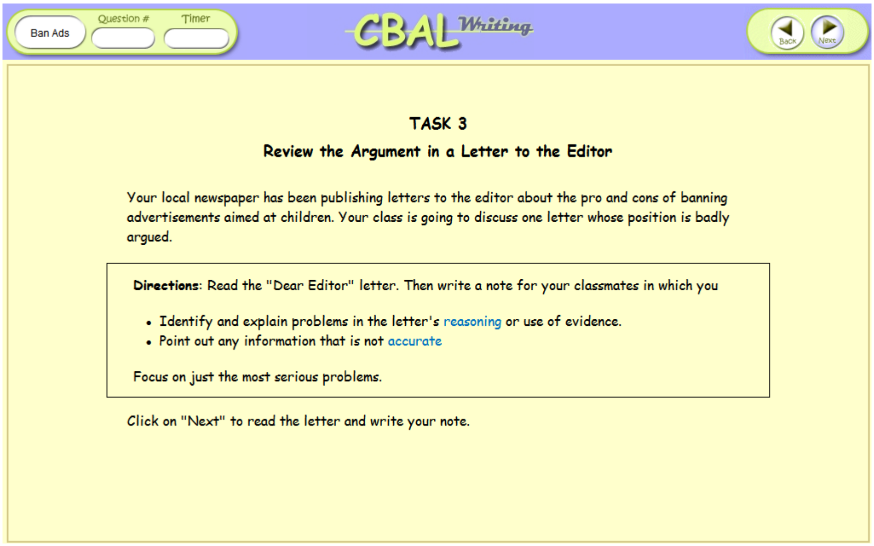
portance of argumentation schemes (Walton 1996) in analyzing and evaluating arguments.

Recognition of typical reasoning flaws is normally taught during middle and high school in the U.S.; therefore, informed by curriculum standards and learning sciences research, our goal is to examine how well students can apply these skills and, in their own words, explain reasoning errors in people's arguments. Meanwhile, we also aim to facilitate the development of middle school students' argument critique skills of informal logic (Walton 2016) by designing scaffolded tasks. In the following sections, we will present the research background, describe the study method, report the study results, and discuss implications for the assessment and instruction of the skill required for evaluating arguments.

2. Theoretical background and related research

To succeed in college and career settings, students must learn to comprehend, critique, and construct reasoned arguments. Understanding logical fallacies is an essential skill in critiquing arguments and constructing plausible arguments. Informal fallacies refer to arguments that are "psychologically persuasive but logically incorrect; that do as a matter of fact persuade but, given certain argumentative standards, shouldn't" (Copi and Burgess-Jackson 1996, p. 97). For instance, people sometimes provide evidence for their claims, but the evidence could be irrelevant or insufficient. Arguers may jump to a conclusion too quickly, while not adequately supporting the conclusion with the premises, which is called *hasty generalization* (Walton 1999). Such arguments are often involved with generalizations from one case to a large population, or from a stereotype group to a specific case. For example, it is probably a hasty generalization if someone claims that TV brings families together based on an observation that his neighbor's family watches TV together every night. Jumping to a conclusion can also be found in *post hoc* arguments, in which people draw a causal conclusion between two events based on the observed correlation between the two events (Walton 2009). Although causal conclusions are often based on correlations, *post hoc* occurs when the arguer overlooks some evidence that ought to be taken into account before reaching the conclusion. There are a host of other informal fallacies (e.g., ad hominem, straw man, slippery slope, and red herring) that make arguments unpersuasive, and students' ability to identify informal fallacies is influenced by their familiarity with norms of argumentation (Neuman, Weinstock, and Glasner 2006; Weinstock, Neuman and Tabak 2004). To detect whether an argument is misused or fallacious, it is important to understand the underlying forms of argumentation, also called *argumentation schemes* (Walton 1996). These schemes represent the relationship between what is stated in the claim and its supporting justificatory structure (Walton 1996; Walton, Reed, and Macagno 2008). Consider a common argumentation scheme, *argument from example,* which is often used in essays when students try to support their arguments with examples from their everyday experiences or background knowledge. Arguments from examples can be evaluated by asking 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? These questions raise critical concerns about the argument strategy, which, if not addressed, could elicit strong counterarguments. For instance, if someone provides an atypical case, then one could undermine the argument by pointing out the limitation of its generalizability. Thus, asking scheme-relevant critical questions helps us differentiate fallacious arguments from reasonable arguments by encouraging recognition of unwarranted assumptions in people's reasoning (Walton 1996; Walton et al. 2008).

The current study builds upon our previous work examining student performance in an argument critique task (Song, Deane and Fowles 2017). In this task, students first read a letter that consists of arguments against banning advertisements aimed at children (see Figure 1). The letter is written to include common fallacious arguments in informal logic, such as hasty generalization, circular argument, post hoc, etc. Students are asked to identify and explain these reasoning problems in their written critiques. We found that the majority of eighth graders (83%) did not write high-quality argument critiques (i.e., ones that identified and clearly explained major problems in the reasoning in the letter). Students' responses showed several characteristic difficulties: (1) being off-task; (2) failing to identify fallacious arguments; (3) having difficulty explaining problems; and (4) not connecting their criticisms with specific parts of the text.



TASK 3

Review the Argument in a Letter to the Editor

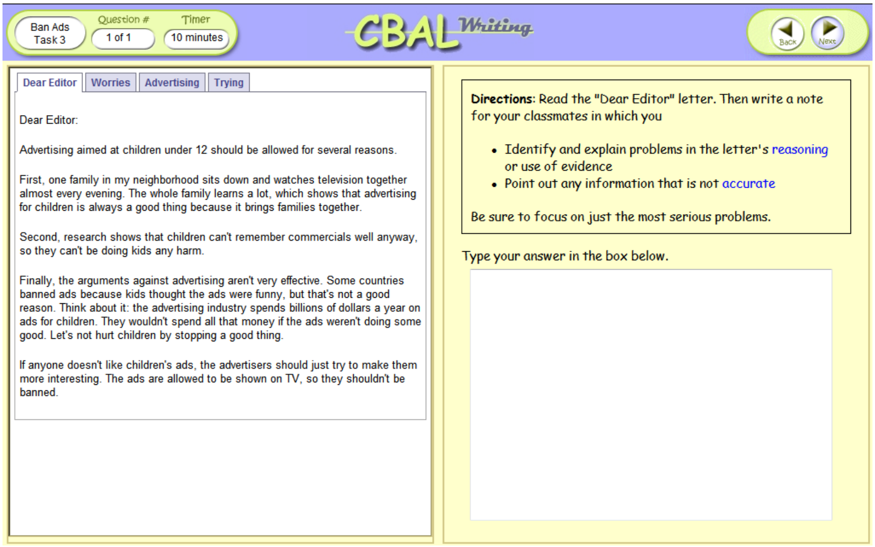
Your local newspaper has been publishing letters to the editor about the pro and cons of banning advertisements aimed at children. Your class is going to discuss one letter whose position is badly argued.

Directions: Read the "Dear Editor" letter. Then write a note for your classmates in which you

- Identify and explain problems in the letter's reasoning or use of evidence.
- Point out any information that is not accurate

Focus on just the most serious problems.

Click on "Next" to read the letter and write your note.



Dear Editor:

Advertising aimed at children under 12 should be allowed for several reasons.

First, one family in my neighborhood sits down and watches television together almost every evening. The whole family learns a lot, which shows that advertising for children is always a good thing because it brings families together.

Second, research shows that children can't remember commercials well anyway, so they can't be doing kids any harm.

Finally, the arguments against advertising aren't very effective. Some countries banned ads because kids thought the ads were funny, but that's not a good reason. Think about it: the advertising industry spends billions of dollars a year on ads for children. They wouldn't spend all that money if the ads weren't doing some good. Let's not hurt children by stopping a good thing.

If anyone doesn't like children's ads, the advertisers should just try to make them more interesting. The ads are allowed to be shown on TV, so they shouldn't be banned.

Directions: Read the "Dear Editor" letter. Then write a note for your classmates in which you

- Identify and explain problems in the letter's reasoning or use of evidence
- Point out any information that is not accurate

Be sure to focus on just the most serious problems.

Type your answer in the box below.

Figure 1. Ban Ads Argument Critique Task (Song et al. 2017, p. 6)

Therefore, we incorporated scaffolding components into the assessment within a Vygotskian framework (Vygotsky 1978), breaking this complex task into easier, more "doable" steps. Specifically, we designed a "lead-in" task as light scaffolding to the written critique task and conducted a new study to explore the effect of the scaffolding. The lead-in task targets the skill of identifying common reasoning flaws, which is supposed to be critical for writing an argument critique. Such scaffolding may provide the support and structure necessary for students to learn how to write a critique or complete the task successfully because they see a model of a successful critique (of different arguments on the same topic) before they write their own.

In addition, this project serves the goal of identifying strategies that can help underserved learners do better in school. There is a persistent achievement gap between socio-economic groups in the U.S., with lower-income students performing worse on average than higher-income students from the same social group (e.g., Leu et al. 2015). Thus, there is reason for examining the subgroup performance and exploring the effectiveness of the scaffolding for the low-income group. Our research questions were:

1. Does the scaffolding design help improve student performance on the written critique tasks?
2. Is there an achievement gap between students from low-income and high-income families?
3. Does the scaffolding design measure students' argument critique skills reliably?

3. Method

Participants

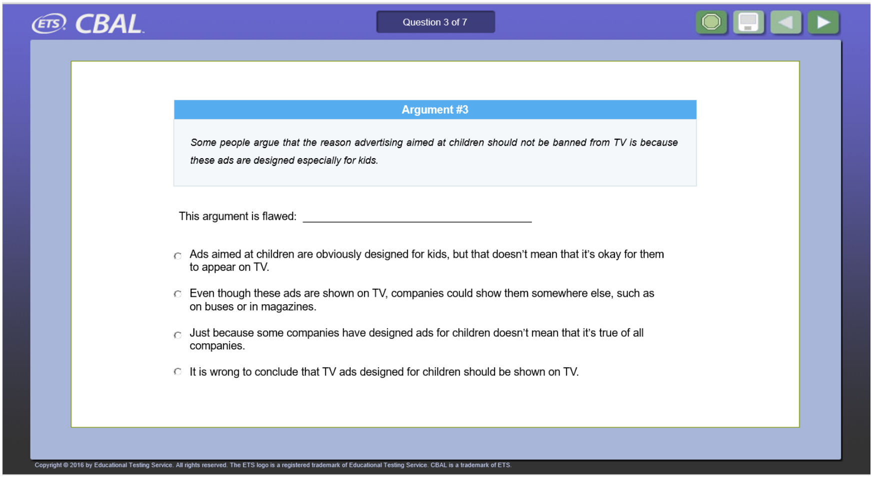
We contacted teachers who previously participated in prior research projects and recruited those who expressed an interest in using the argument critique tasks with their students. Participants were sampled from three middle schools in two U.S. states. A total of 472 students from grade 7 (n = 231) and grade 8 (n = 241) were included. The sample represented a diverse group: Caucasian 39%; Hispanic 28%; Asian 18%; African American 13%; Other 1%; Unreported 1%. Most students (79%) were from low-income families, as they were qualified for receiving free or reduced-price lunch.

Instruments

Argument Critique Tasks

We used two existing argument critique tasks designed from the same blueprint: Ban Ads and Cash for Grades. Both were scenario-based assessments, which contain a structured series of tasks within an overarching scenario context that provides a purpose for reading and writing from source texts (Sabatini, O'Reilly and Deane 2013). The Ban Ads scenario raises the issue of whether the United States should ban advertising to children under the age of 12, and the Cash for Grades scenario asks whether students should be rewarded with money for getting good grades. The original critique tasks ask students to evaluate the arguments presented in a letter to the editor. In doing so, students must write a critique of the argument (i.e., identify and explain problems in the reasoning or use of evidence and point out any inaccurate information).

To create the scaffolded versions of the critique task, we added a “lead-in” task for the Ban Ads and Cash for Grades scenarios. The lead-in task consists of seven multiple-choice items as part of the scenario. These items assess whether students can identify common reasoning errors and choose appropriate words or phrases in written critiques about the issue under discussion (Ban Ads or Cash for Grades). Specifically, items 1–5 pose arguments with various fallacies (e.g., post hoc, hasty generalization), and students need to select the option that correctly explains the reasoning error in a given argument. Items 6 and 7 focus on the critique writing skill, in which students are supposed to select appropriate words or phrases to help formulate a brief and meaningful critique. See Figure 2 for a couple examples of multiple-choice items in Ban Ads.



Argument #3

Some people argue that the reason advertising aimed at children should not be banned from TV is because these ads are designed especially for kids.

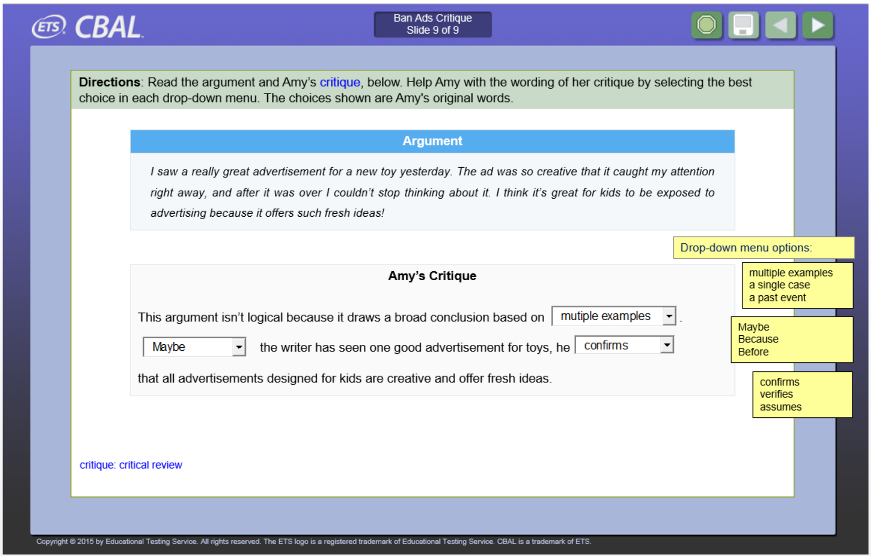
This argument is flawed: ________________________________

○ Ads aimed at children are obviously designed for kids, but that doesn’t mean that it’s okay for them to appear on TV.

○ Even though these ads are shown on TV, companies could show them somewhere else, such as on buses or in magazines.

○ Just because some companies have designed ads for children doesn’t mean that it’s true of all companies.

○ It is wrong to conclude that TV ads designed for children should be shown on TV.



Directions: Read the argument and Amy's critique, below. Help Amy with the wording of her critique by selecting the best choice in each drop-down menu. The choices shown are Amy's original words.

Argument

I saw a really great advertisement for a new toy yesterday. The ad was so creative that it caught my attention right away, and after it was over I couldn't stop thinking about it. I think it's great for kids to be exposed to advertising because it offers such fresh ideas!

Drop-down menu options:

Amy's Critique

This argument isn't logical because it draws a broad conclusion based on [mutiple examples ▾].

[Maybe ▾] the writer has seen one good advertisement for toys, he [confirms ▾] that all advertisements designed for kids are creative and offer fresh ideas.

multiple examples
a single case
a past event

Maybe
Because
Before

confirms
verifies
assumes

critique: critical review

Copyright © 2015 by Educational Testing Service. All rights reserved. The ETS logo is a registered trademark of Educational Testing Service. CBAL is a trademark of ETS.

Figure 2. Sample multiple-choice items in the Ban Ads "leading-in" task

Two types of forms were then created based on the tasks: one includes both the lead-in task and the original writing task; the other has the original writing task only. Moreover, the lead-in task has two versions: one provides feedback that tells students the correct answers (e.g., Option A is the correct answer), and the other does not provide any feedback. Therefore, there were three conditions reflecting three different lead-in settings: lead-in only, lead-in plus feedback, and no lead-in (control).

State reading and writing test

Students' reading and writing scores on relevant state standardized tests were provided by their teachers. Given the different states, one school took Partnership for Assessment of Readiness for College and Careers (PARCC)[2] reading and writing assessments, and the other two schools took ACT Aspire assessments.[3] These tests were intended to measure students' ELA reading and writing

[2] https://parcc-assessment.org/
[3] https://www.actaspire.org/

abilities, evaluating how well the students are meeting academic expectations.

Procedure

A two-phase study (one class period) was administered with three conditions: control (no lead-in task), scaffolded (lead-in task only), or scaffolded with feedback (lead-in task plus feedback). In other words, students in the control condition worked on the original written critique tasks in both phases, while students in the experimental conditions worked on the scaffolded lead-in task and the original written critique task in Phase I and only the original written critique task in Phase II. Each individual student received a participant ID prior to the study, and we randomly assigned the IDs to the three conditions.

Although Ban Ads and Cash for Grades were designed from the same blueprint, the two topics could pose a varying level of difficulty to students due to factors such as the content, the types of arguments in the letter, the given arguments in the tasks, and the alignment between the lead-in task and the written critique task. Therefore, we randomly assigned the order of Ban Ads and Cash for Grades across the phases, such that some students completed Ban Ads in Phase I and Cash for Grades in Phase II, while other students completed them in the opposite order.

Scoring

Responses to the lead-in tasks were automatically scored by the computer. Each lead-in task consists of seven multiple-choice items. One point was assigned to each question, so students could earn up to 7 points for the lead-in task. The overall quality of each critique was rated on a scale from 0 to 4 by human raters. In scoring students' written critiques, the following aspects were considered: (a) whether students identified and clearly explained most of the major problems in the letter's reasoning and (b) whether students expressed ideas in an appropriate tone for the class. See Figure 3 for the scoring rubric.

“BAN ADS” Argument Critique Task Scoring Rubric	
4 **Excellent**	An “Excellent” Critique • Identifies and clearly explains most of the major problems in the letter’s reasoning, • Points out inaccurate information, AND • Expresses ideas in a clearly appropriate tone for the class
3 **Adequate**	An “Adequate” Critique • Identifies and explains some of the major problems in the letter’s reasoning and inaccurate information, OR • Identifies only one major problem in the letter’s reasoning but explains it extremely well, AND • Expresses ideas in a generally appropriate tone for the class
2 **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: **a)** Explains the problem(s) poorly, if at all **b)** Misinterprets parts of the letter or includes irrelevant information **c)** Misinterprets an important part of the task **d)** Expresses ideas in a somewhat inappropriate tone for the class
1 **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: **a)** Is very confusing

	b) Seriously distorts the letter or includes mostly irrelevant information **c)** Seriously distorts the writing task **d)** Expresses ideas in a highly inappropriate tone for the class
0 **No Credit**	A response receives “No Credit” for any one of the following reasons: **a)** Identifies no problems in the letter’s reasoning and/or use of inaccurate information **b)** Not long enough for critical-thinking skills to be judged **c)** Not written in English **d)** Off topic **e)** Blank **f)** Only random key strokes

Figure 3. Scoring Rubric (Song et al. 2017, p. 7)

Candidates voluntarily participated in this project as scorers on Amazon Turk. These scoring candidates first studied the materials online themselves. The materials included the critique tasks, scoring rubrics, topic notes, and anchor responses. After they reviewed these materials, they scored a practice set of student responses and then participated in a scoring qualification test. A total of 90 Amazon Turk participants passed the qualification test for a topic (by reaching 80% of the agreement with the pre-assigned scores in a test set of student responses) and were assigned to score student responses. Half of the raters scored Ban Ads, and the other half scored Cash for Grades. We randomly assigned student responses to these raters. Each response received at least eight scores. Given the design, we conducted a generalizability analysis (Webb, Shavelson, and Haertel 2006) for each topic, focusing on the 20 responses with scores from all 45 raters to assess rater reliability. Overall, the observed generalizability coefficients were greater than .90, and the more raters that were included, the greater the generalizability coefficient. Therefore, the average of all ratings was considered the final score for the response.

Data analysis

Nonparametric methods were applied to all analyses relating to the writing scores because of inconsistent patterns in score distributions (right skewed in most cases). The multiple-choice scores generally followed a normal distribution. Table 1 presents the average performance on the lead-in task and written critique task in Phase I and the written critique task in Phase II. Regardless of condition and topic, students did not achieve high scores in the lead-in task (below the 3.5, the mid-point of 7 points), and their performances on the written critique tasks were low in general.

Table 1. Mean scores of the Phase I and Phase II tasks across conditions

Task Order (Condition)	***N***	***Phase I Lead-In (SD)***	***Phase I Critique (SD)***	***Phase II Critique (SD)***
Ban – Cash (Control)	81	n/a	1.38 (1.18)	.94 (1.01)
Ban – Cash (S no FB)	81	3.22 (1.88)	2.00 (1.34)	1.10 (1.09)
Ban – Cash (S & FB)	76	3.43 (1.91)	1.88 (1.37)	1.07 (1.17)
Cash – Ban (Control)	80	n/a	1.27 (1.08)	1.44 (1.24)
Cash – Ban (S no FB)	78	2.83 (1.66)	1.35 (1.15)	1.31 (1.23)
Cash – Ban (S & FB)	76	3.14 (1.61)	1.45 (1.22)	1.46 (1.33)

Note. Ban = Ban Ads; Cash = Cash for Grades; S = scaffolded condition that has a lead-in task; FB = feedback; n/a = not available.

To answer RQ1 regarding the impact of the scaffolding design, we first ran a 2 (topic order) X 3 (condition) factorial ANOVA on ranks (Wobbrock, Findlater, Gergle and Higgins 2011) to examine the potential for main effects or an interaction effect on the scores of the written critique tasks in both phases. Then, we applied the Kruskal-Wallis test to first compare the writing scores from Phase I for each topic between conditions, then to compare the writing scores regardless of phases for each topic between conditions. For RQ2 about the students' performance from low-SES families, we used the Mann-Whitney test to compare writing scores between the low-SES and high-SES groups in each phase for each condition. For RQ3, concerning the reliability of the scaffolded task design, we examined Spearman's rank correlations between Phase I scores (i.e., the critique writing scores and the combined sum of the lead-in and critique writing scores) and the state test scores for each topic.

Across all statistical analyses, the significance level α was set at 0.05. Also, by conducing power analysis with the assumption of a statistical power of 0.9, a minimum sample size of 324 was needed for a 2X3 factorial ANOVA analysis. Our sample was therefore sufficient for the analyses we planned to conduct.

4. Results

RQ1. Does the scaffolding design help improve student performance on the written critique task?

A factorial ANOVA on ranks was conducted to compare the main effects of order of task topic and condition and the interaction effect between them on Phase I and Phase II written critique task scores. The topic order included two levels (Ban Ads - Cash for Grades, and Cash for Grades - Ban Ads), and the condition consisted of three levels (control, scaffolded, scaffolded with feedback). Results showed significant main effects of topic order for written critique scores: $F(1, 466) = 10.28$, $p = .001$ for Phase I; and $F(1, 466) = 7.56$, $p = .006$ for Phase II. A significant main effect of scaffolding version only appeared for the Phase I written critique score: $F(2, 466) = 3.24$, $p = .040$. No significant interaction effect was found in either phase. In addition, for students who started with Ban Ads, Wilcoxon Signed Rank tests showed that they had significantly higher scores for Ban Ads than Cash for Grades (*all p's* < .001) in all the conditions. On the other hand, for students who started with Cash for Grades, they performed similarly on both topics in all three conditions. These results indicated that the task topics had an effect. Therefore, we proceeded to analyze the data by examining the scaffolding effect on each topic.

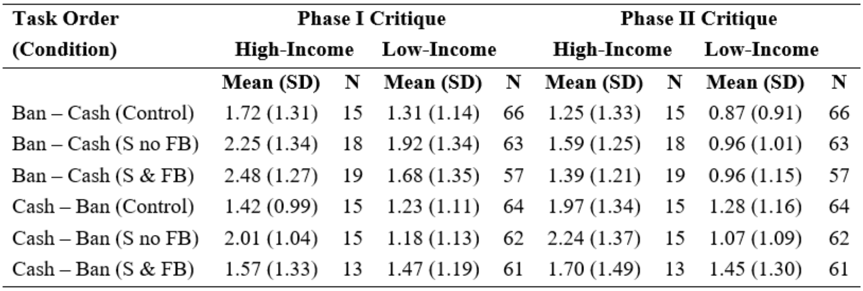
To compare the scores of the Phase I written critique for each topic among the conditions, a Kruskal-Wallis test showed that there was a statistically significant difference in Phase I Ban Ads scores among the three conditions ($\chi^2(2) = 8.54$, $p = .014$). Post-hoc pairwise comparisons with Type I error controlled across the tests using the Bonferroni approach showed that students in the scaffolded only condition performed significantly better than students in the control condition (*Cohen's d* = 0.45), but there was no significant difference between the two scaffolded conditions. As for the Phase I Cash for Grades scores, we did not find any significant difference among the three conditions [$\chi^2(2) = 0.46$, $p = .797$].

Next, we further compared the scores of the written critiques on each topic across the conditions regardless of the phases (e.g., three groups took Ban Ads in Phase I and three groups took Ban Ads in Phase II). A Kruskal-Wallis test showed a significant difference among the conditions for Ban Ads only [$\chi^2(5) = 18.61$, $p = .002$]. As previously indicated, the post-hoc pairwise comparisons indicated that this difference was driven by the significant difference between the scaffolding only and the control conditions. These results suggested that students critiquing the Ban Ads arguments may benefit from the scaffolding design, but the feedback was not helpful.

RQ2. Is there an achievement gap between students from low-income and high-income families?

Table 2 shows the average writing performance scores in each phase for every condition from the low and high SES groups, which were defined based on whether the students were receiving reduced-price/free lunch or not. The low-income groups performed worse on the written critique tasks than the high-income groups in both phases under the same condition. Mann-Whitney tests revealed significant differences in the scaffolded only condition (Phase II Cash for Grades: *U* = 394.5, *p* = .050; Phase I Cash for Grades: *U* = 276.5, *p* = .015; Phase II Ban Ads: *U* = 245, *p* = .005), and the scaffolded with feedback condition (Phase I Ban Ads: *U* = 359, *p* = .028). The corresponding effect sizes (Cohen's d) were 0.45, 0.58, 0.68, and 0.52, further suggesting that the differences between the low and high SES groups were not trivial.

Table 2. Written critique scores of low-SES and high-SES groups



Task Order	Phase I Critique				Phase II Critique			
(Condition)	High-Income		Low-Income		High-Income		Low-Income	
	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N
Ban – Cash (Control)	1.72 (1.31)	15	1.31 (1.14)	66	1.25 (1.33)	15	0.87 (0.91)	66
Ban – Cash (S no FB)	2.25 (1.34)	18	1.92 (1.34)	63	1.59 (1.25)	18	0.96 (1.01)	63
Ban – Cash (S & FB)	2.48 (1.27)	19	1.68 (1.35)	57	1.39 (1.21)	19	0.96 (1.15)	57
Cash – Ban (Control)	1.42 (0.99)	15	1.23 (1.11)	64	1.97 (1.34)	15	1.28 (1.16)	64
Cash – Ban (S no FB)	2.01 (1.04)	15	1.18 (1.13)	62	2.24 (1.37)	15	1.07 (1.09)	62
Cash – Ban (S & FB)	1.57 (1.33)	13	1.47 (1.19)	61	1.70 (1.49)	13	1.45 (1.30)	61

Note. Ban = Ban Ads; Cash = Cash for Grades; S = scaffolded condition that has a lead-in task; FB = feedback.

RQ3. Does the scaffolding design measure students' argument critique skills reliably?

To answer this question, we focused on Phase I performance because no scaffolding was involved in Phase II. Given that feedback did not have any effect on student performance, we combined the two scaffolded conditions. Participants with PARCC scores were used to examine the correlation between the Phase I task and the state test, given that the majority of our sample took this test (Table 3).

Table 3. Correlations between the critique task performance (Phase I) and state test scores

Task (Condition)	***Measure***	***N***	***State Reading***	***State Writing***
Ban (Control)	CR	44	.61	.55
Ban (Scaffolded)	CR	96	.64	.63
Ban (Scaffolded)	MC + CR	96	.71	.67
Cash (Control)	CR	49	.68	.70
Cash (Scaffolded)	CR	93	.73	.65
Cash (Scaffolded)	MC + CR	93	.68	.54

Note. Ban = Ban Ads; Cash = Cash for Grades; CR = constructed response; MC = multiple choice.

The following analysis included 282 students for whom PARCC test scores were available. We correlated the Phase I scores with the state test scores for control and non-control conditions on each topic. Spearman’s rank correlations ranged from .61 to .73 between Phase I scores and PARCC reading scores, while Phase I scores’ correlations with PARCC writing ranged from .54 to .70 (*all p’s* < .01). The results showed that Phase I scores have higher correlations with the state reading scores than the state writing scores, except for the Cash for Grades control condition. In addition, correlations between the scores from scaffolding design (non-control) and the state scores tended to be higher than those without (control condition), except for the correlations between Cash for Grades and state writing.

5. Conclusion

In this study, we examined the effect of scaffolding on student performance on argument critique tasks. The study results showed a small positive impact of scaffolding on one topic, but not the other, revealing that student skills of writing critiques could be affected by topic and argument content. Additionally, students from low-SES families did not perform as well as their peers. We also found that student performance on the critique tasks had moderate or strong correlations with the state reading and writing test scores. It is not surprising that many students did not produce high-quality critiques because these kinds of tasks might have been new to them. Middle school students may be just beginning to learn some common reasoning errors, while other students may not be introduced to logical fallacies or evidence-based justification until high school. In our study, some students appeared to have difficulty explaining the reasoning problems, even when they had a rough sense that something was wrong in the given arguments. For example, a student wrote: "I am here to explain some mistakes that have been made while writing this letter to the editor. Such as, in your first reason you don't really know if advertising is always a good thing. Also, it may or may not bring families together." This student pointed out that advertisements may not bring families together as it is claimed in the letter but did not provide an explanation that the claim is overgeneralizing from a single example.

One implication from our study is that scaffolded lead-in tasks and feedback themselves may not be strong enough to support the development of students' skills in critically evaluating arguments, especially for students from low-income families. Each lead-in task only consisted of seven multiple-choice items on reasoning flaws, and the feedback simply informed the students of the correct answers without providing an explanation of the rationale. Prior research has shown that the effects of feedback on student learning may vary due to individual differences, task characteristics, and feedback type (Shute 2008). Students may benefit more from receiving feedback that provides an explicit explanation of the reasoning flaws in the given arguments and shows the strategy for identifying such flaws.

In addition, the meta-analysis study on academic interventions for students with low SES status conducted by Dietrichson, Bog, Filges, and Jorgensen (2017) shows that tutoring and feedback and process monitoring have relatively robust average effect sizes on elementary and middle school students' academic achievements when teachers are involved in the learning process. There has been rich evidence of the persistent achievement gap between low SES students and high SES students (e.g., Kim and Quinn 2013; Sirin 2005; White 1982). Low SES students may require intensive instructional support to catch up. Therefore, students from low-income families and low-achieving students might benefit more from direct instruction as to 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 et al. 2004). Typically, this higher-order, argumentative thinking skill does not develop naturally (Kinsler 1990). Teaching students effective strategies could help improve their ability to evaluate the strengths and weaknesses of a given argument, as indicated in existing studies (Nussbaum and Edwards 2011; Song and Ferretti 2013). In future work, we could design scaffolded tasks that require students to use some strategies (e.g., asking critical questions) as well as provide feedback that enables students to learn these strategies.

Another implication from our study is that students' reactions to different topics and content may vary. Students in general performed better on Ban Ads than Cash for Grades, even though these two task sets were designed from the same blueprint. While prior research that involved these forms did not reveal significant differences in student performance (Deane et al. 2019; van Rijn, Graf and Deane 2014), it is important to point out that the types of fallacies in these two written critique tasks are quite different. The Ban Ads critique consists of several different types of reasoning errors, such as hasty generalization, begging the question, post hoc, false assumption, and contradictory information, but the Cash for Grades critique focuses on the post hoc fallacy, which involves jumping to a quick causal conclusion based on the correlation of two factors. Given that the scaffolding design in the current study is aimed at covering a variety of common reasoning errors, it is aligned better with the Ban Ads written critique task than the Cash for Grades written critique task. In Cash for Grades, students are expected to demonstrate a thorough analysis of various factors that could contribute to a situation (e.g., social and personal reasons, as well as school factors that might lead to better academic achievement). This appears to be challenging to the participating students. While many inductive arguments to causal conclusions are established on correlations, the fallacy of post hoc occurs when people overlook evidence that ought to be taken into account and make a quick causal conclusion. Walton (2009) presented three critical questions that match the argumentation scheme for arguing from correlation to causation, which can help people detect the post hoc fallacy: (1) Is there really a correlation between A and B? (2) Is there any reason to think that the correlation is any more than a coincidence? (3) Could there be some third factor C that is causing both A and B? (p. 216) Future research could explore how to effectively support students' learning of argumentation schemes and critical questions and examine the impact on argument critique skills.

We also found that many students expressed that awarding students with cash for good academic performance is a great idea, which could result in acknowledging the author's arguments rather than objectively critiquing them. Critical thinking research has shown that prior belief biases typically lead to accepting fallacious arguments and invalid conclusions in a variety of critical thinking and evaluation tasks (West, Toplak and Stanovich 2008). People tend to accept belief-consistent arguments more readily than belief-inconsistent arguments (Klaczynski et al. 1997; Stanovich and West 2007).

The study has some limitations. First, data was collected from a convenience sample and so cannot be considered representative of the U.S. middle school population as a whole. The participating schools were involved in other prior research projects and were selected on a first-come, first-served basis. Second, we did not collect information about participants' motivation, which probably impacted student responses. We can infer from student responses and the timing data that some did not take the test seriously.

Roughly 5% of the students[4] wrote something irrelevant to the task or simply copied some text from the task (e.g., "The school should have a program where kids meet other kids and make friends"). These students spent less than one minute completing their written responses. Motivation is always an issue for low-stakes assessments. Third, we did not measure students' background knowledge or beliefs on the two topics used in the test, so we could not draw any conclusion regarding how those aspects affected their argument critique performance. Only two topics were used in this study, and student performance may vary if different topics were presented.

In sum, the results of this study suggest the need to increase effort to design effective scaffolding for students who are developing skills associated with identifying fallacious arguments. Students need to learn to detect fallacies that abound in everyday argumentation, which often contribute to unfounded beliefs and misconceptions. In future work, we will continue our investigation of how to optimize the scaffolding and assessment design to support the development of students' argument critique skills, such as designing explanatory feedback and including activities that promote skills of asking critical questions.

References

CCSS. 2018. Students who are college and career ready in reading, writing, speaking, listening, & language. Retrieved March 21, 2020, from http://www.corestandards.org/ELA-Literacy/introduction

Copi, Irving M., and Keith Burgess-Jackson. 1996. *Informal logic*. Englewood Cliffs, NJ: Prentice Hall.

Deane, Paul, Yi Song, Peter van Rijn, Tenaha O'Reilly, Randy E. Bennett, Mary Fowles, John Sabatini and Mo Zhang. 2019. The case for scenario-based assessment of written argumentation. *Reading and Writing* 32: 1575–1606.

[4] We calculated the mean scores of each task for each condition after removing these students and found that the mean scores were similar to those presented in Table 1. So, we decided to include all the participants in our analyses.

Dietrichson, Jens, Martin Bog, Trine Filges and Anne-Marie K. Jorgensen. 2017. Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research* 87(2): 243–282.

Ferretti, Ralph P., Willan E. Lewis and Andrews-Weckerly, Scott. 2009. Do goals affect the structure of students' argumentative writing strategies? *Journal of Educational Psychology* 101: 577–589.

Ferretti, Ralph P., Charles A. MacArthur and Nancy S. Dowdy. 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(4): 694–702.

Kim, James. S. and David M. Quinn. 2013. The effects of summer reading on low-income children's literacy achievement from kindergarten to grade 8: A meta-analysis of classroom and home interventions. *Review of Educational Research* 83(3): 386–431. doi:10.3102/0034654313483906

Kinsler, Kimberly. 1990. Structured peer collaboration: Teaching essay revision to college students needing writing remediation. *Cognition and Instruction* 7(4): 303–321.

Klaczynski, Paul A., David H. Gordon and James Fauth. 1997. Goal-oriented critical reasoning and individual differences in critical reasoning biases. *Journal of Educational Psychology* 89(3): 470–485.

Kuhn, Deanna. 2009. The importance of learning about knowing: Creating a foundation for development of intellectual values. *Child Development Perspectives* 89(2): 112–117.

Kuhn, Deanna, Richard Cheney and Michael Weinstock. 2000. The development of epistemological understanding. *Cognitive Development* 15(3): 309–328.

Leu, Donald J., Elena Forzani, Chris Rhoads, Cheryl Maykel, Clint Kennedy and Nicole Timbrell. 2015. The new literacies of online research and comprehension: Rethinking the reading achievement gap. *Reading Research Quarterly* 50(1): 37–59.

National Center for Education Statistics. 2012. *The Nation's Report Card: Writing 2011* (NCES 2012-470). Institute for Education Sciences, U.S. Department of Education, Washington, D.C.

Neuman, Yair, Michael P. Weinstock and Amnon Glasner. 2006. The effect of contextual factors on the judgement of informal reasoning fallacies. *The Quarterly Journal of Experimental Psychology* 59(2): 411–425.

Nussbaum, Michael E. and Ordene V. Edwards. 2011. Critical questions and argument stratagems: A framework for enhancing and analyzing students' reasoning practices. *The Journal of the Learning Sciences* 20(3): 443–488.

Nussbaum, Michael E. and CarolAnne M. Kardash. 2005. The effects of goal instructions and text on the generation of counterarguments during writing. *Journal of Educational Psychology* 97(2): 157–169.

Rijn, Peter van, Edith A. Graf and Paul Deane. 2014. Empirical recovery of argumentation learning progressions in scenario-based assessments of English language arts. *Spanish Journal of Educational Psychology (Psicologia Educativa)* 20: 109–115.

Sabatini, John, Tenaha O'Reilly and Paul Deane. 2013. Preliminary reading literacy assessment framework: Foundation and rationale for assessment and system design (Research Report 13-30). Princeton, NJ: Educational Testing Service.

Shute, Valerie J. 2008. Focus on formative feedback. *Review of Educational Research* 78(1): 153–189.

Sirin, Selcuk R. 2005. Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3): 417–453. doi:10.3102/00346543075003417

Song, Yi. and Ralph Ferretti. 2013. Teaching critical questions about argumentation through the revising process: effects of strategy instruction on college students' argumentative essays. Special Issue: *Reading and Writing: An Interdisciplinary Journal, 26*(1), 67–90.

Song, Yi**.,** Paul Deane, and Mary E. Fowles. 2017. Examining students' ability to critique arguments and exploring assessment and instructional implications. (Research Report No. RR-17-16). Princeton, NJ: Educational Testing Service.

Stanovich, Keith E. and Richard F. West. 2007. Natural myside bias is independent of cognitive ability. *Thinking & Reasoning* 13(3): 225-247. doi:10.1080/13546780600780796

Vygotsky, Lev S. 1978. *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Walton, Douglas. 1996. *Argumentation schemes for presumptive reasoning.* Mahwah, NJ: Lawrence Erlbaum.

Walton, Douglas. 1999. Rethinking the Fallacy of Hasty Generalization. *Argumentation* 13(2): 161–182.

Walton, Douglas. 2009. Jumping to a conclusion: Fallacies and standards of proof. *Informal Logic* 29(2): 215–243.

Walton, Douglas, Chris Reed and Fabrizio Macagno. 2008. *Argumentation schemes*. New York, NY: Cambridge University Press.

Webb, N. M., R. J. Shavelson and E. H. Haertel. 2006. Reliability coefficients and generalizability theory. *Handbook of Statistics* 26(4): 81–124.

Weinstock, Michael, Yair Neuman and Iris Tabak. 2004. Missing the point or missing the norms? Epistemological norms as predictors of students' ability to identify fallacious arguments. *Contemporary Educational Psychology* 29(1): 77–94.

West, Richard F., Maggie E. Toplak and Keith E. Stanovich. 2008. Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology* 100(4): 930–941. doi:10.1037/a0012842.

White, Karl R. 1982. The relation between socioeconomic status and academic achievement. *Psychological Bulletin* 91(3): 461–481. doi:10.1037/0033-2909.91.3.461

Wobbrock, Jacob O., Leah Findlater, Darren Gergle and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '11). Vancouver, British Columbia (May 7–12, 2011). New York: ACM Press, pp. 143–146. Retrieved March 7, 2020, from http://faculty.washington.edu/wobbrock/pubs/chi-11.06.pdf