Effects of Critical Discussions on Middle School Students’ Written Historical Arguments

Daniel R. Wissinger
Indiana University of Pennsylvania

Susan De La Paz
University of Maryland

In this experimental study, 151 middle school students explored 3 historical controversies, first reading and discussing primary source documents in groups, then writing arguments on their own. Students were either randomly assigned to an experimental condition, using argumentative schemes and critical questions as guides during discussions, or to a comparison condition in which a traditional set of questions was used to guide discussions. Students in both conditions read the same historical controversies and used the same text structure heuristic to better compare reading and writing outcomes after students participated in discussion. The findings after instruction indicate comparable reading comprehension and comparable composing skill on general writing measures across conditions. Importantly, the findings also indicate disciplinary benefits for students in the experimental condition in terms of their ability to learn historical content and regarding the quality of students’ historical reasoning in their written arguments. Argument schemes and critical questions appeared particularly helpful in facilitating students’ substantiation of claims and development of rebuttals.

Keywords: historical reasoning, discussion, written argumentation, middle school

Supplemental materials: http://dx.doi.org/10.1037/edu0000043.supp

Argumentation is an important genre and a key feature of successful writing by academics across disciplines (Lea & Street, 1998). Students’ argumentation may be viewed as a milestone of conceptual development (e.g., Kuhn, 2005), and as evidence of critical thinking (Halpern, 1998). Writing arguments has been shown to promote better audience awareness and syntactic complexity than writing narratives and description (Crowhurst & Piche, 1979), and helps advanced students understand content (Wiley & Voss, 1999). In addition, argumentation has become a central feature of the Common Core State Standards, which now integrates reading and writing into English language arts, history and social studies, science and technical subjects. Students are expected to become proficient in composing “logical arguments based on substantive claims, sound reasoning, and relevant evidence” (CCSSI, 2010).

Stevens, Wineburg, Herrenkohl, and Bell (2005) contend that the nature of effective argumentation differs across disciplines because the epistemological criteria for causal explanations differ.

This idea has been confirmed by Carter, Ferzli, and Wiebe (2007) in studies in biology, Brown (2007) in mathematics, and Montesano, Sano (2010) and De La Paz et al. (2014) in history. Students therefore must learn to differentiate between everyday argumentation and the argumentation of the disciplines (Moje, 2008; Shanahan & Shanahan, 2008). Unfortunately, data from the National Assessment of Educational Progress (NAEP) shows that the 74% of Grade 8 and 73% of Grade 12 writers who scored Basic or below were unable to provide adequate support for their positions when crafting basic arguments (U.S. Department of Education, 2011). Given that students struggle to make general arguments and the unique demands that accompany writing in the disciplines, it is likely that they will face even greater challenges when asked to write arguments in the context of history.

Theoretical Framework

Historical Writing Requires Domain-Specific Reasoning

History requires taking an investigative stance toward the past and an understanding of the norms of knowledge construction and communication in the discipline (Bain, 2006). It demands an understanding of historical knowledge, where historical knowledge comes from, and how it is developed (Nokes, 2013). These historical ways of thinking are evident in experts’ reading. Wineburg (1991) observed that when exploring events in the past, historians simultaneously account for the influence of time, place, and cultural and linguistic norms that existed in the time period in question on the motivations of the authors. Historians situate authors and events in the context of contemporary events, peers,
and ideas, recognizing that their writings highlight the relationships between contiguous events (cf. Mink, 1987). History then is “filled with inferences, decisions about significance, interpretations, inclusions and omissions, generally accepted facts, and even speculations” (Nokes, 2013, p. 55).

Writing in history requires the ability to construct an argument from multiple, sometimes conflicting, sources of evidence. To do so, writers must compare and evaluate documents for the presence and merits of competing claims and evidence (Britt, Rouet, Georgi, & Perfetti, 1994; Kuhn, Weinstock, & Flaton, 1994) in order to make the case for a particular interpretation (Mink, 1987). Further, although the structure of argumentation in history is comparable to argumentation in other disciplines, the nature of evidence and the connection between evidence and claim are discipline-specific.

Argumentation Models

Toulmin’s (1958) argumentation framework is perhaps the most widely used model in education and outlines key aspects of argumentative writing including claim, data, warrant, and counterargument—components that can be used to support conclusions on issues in most disciplines. The central element is the claim; a statement that is advanced to support a conclusion. Claims are established through chains of reasoning that connect the claim and other forms of evidence. In Toulmin’s framework, the acceptability of one’s argument improves when the author assembles evidence to support the claim, and details the conditions under which he or she holds the claim to be true. Though more commonly used as a guide for constructing arguments, Toulmin’s model has also been widely used to examine argumentative structures (van Eemeren, Grootendorst, & Henkemans, 1996). However, one important limitation is that it does not establish domain-specific standards for the quality of argumentation elements (Nussbaum, 2011). Another limitation is that it fails to consider both sides of a controversial issue (Jonassen & Kim, 2010; Litto, 2000).

Walton (1996) provides an alternate model that suggests argumentation is a goal-directed and interactive activity in which proponents advance their opinions through the provision or disproving of assumptions. When arguments are presumptive, Walton perceives reasoning as tentative and open to challenge. The burden of proof is therefore shifted to the opposition in a dialogue. Specifically, counterarguments are equally important as the original argument. Walton and his colleagues (Walton, Reed, & Macagno, 2008) define presumptive argumentation schemes, or conventionalized ways of representing the relationship between a standpoint and its supporting justificatory structure, and have catalogued dozens of argument structures with wide-ranging purposes. Based on these qualities and other pedagogical advantages, we believe their model is relevant for the field of history and has several advantages over Toulmin’s model when working with adolescents.

First, Walton and colleagues dialogue theory offers a comprehensive framework for evaluating historical arguments because it provides several argumentation schemes that are useful for determining the defensibility of arguments in the context of history. For example, the cause-to-effect argumentation scheme, which involves making an explicit claim that some event caused another outcome, is commonly used by historians (von Ranke, 2010) and frequently appears in social studies textbooks. The expert opinion, rule, and consequence argumentation schemes are also consistent with practices used by historians to investigate the past (Gaddis, 2002). These schemes, or common forms of argument (Walton, 1996), along with their attached critical questions can be applied to specific sets of circumstances and used to list ways of challenging arguments created using the scheme (Gordon & Walton, 2003). If a critical question is not answered satisfactorily, the proposition advanced in the scheme may be weakened or defeated. Thus, schemes and critical questions might function as heuristics, guiding students in their evaluation of information in history classrooms.

Second, because Walton and his colleagues “recognize that arguments unfold in the dialectical interchange between two or more parties” (Nussbaum, 2011, p. 4), the framework can be integrated into traditional instructional activities such as discussions, and may serve as a cognitive tool for scaffolding the development of historical arguments. Facilitating discussions with Walton et al.’s (2008) framework may teach students to present argumentation schemes, use critical questions to evaluate evidence, and to pose alternative claims on controversies. Prior work also suggests that discussion can serve as a bridge between everyday and disciplinary arguments if it is used to critically evaluate evidence (Felton & Herko, 2004; Reisman, 2012; Reznitskaya et al., 2001).

Recent work has demonstrated the utility of Walton’s framework for teaching students to write sophisticated, domain-general arguments. Nussbaum and Edwards (2011) taught middle school students critical questions for several argument schemes that were relevant to classroom discussions of current events (e.g., argument from consequences, and from an established rule). Students who had access to this form of instruction developed more integrative arguments and better refutation skills than peers in a comparison group. Moreover, Song and Ferretti (2013) found that college students who learned to ask and answer critical questions about the argument from consequences and argument from example schemes while writing about controversial topics wrote papers that were of higher quality, with more counterarguments, alternate standpoints, and rebuttals than students who learned to use argumentation schemes without the accompanying critical questions, or than students who were in a control group who learned neither. These results demonstrate the benefits of Walton’s approach for writing everyday arguments. However, it remains an open question as to whether the use of Walton et al.’s argumentation schemes and critical questions in promoting deliberative discussions about historical controversies could facilitate students’ abilities to write better historical arguments.

Finally, prior descriptive work by De La Paz, Ferretti, Wissinger, Yee, and MacArthur (2012) demonstrated that eighth- and 11th-grade writers used a wide range of argumentation...
schemes in response to specific historical controversies (such as who was the most promising leader for African Americans during the Progressive Era and causes of the Mexican American war), and that good and poor writers’ essays could be reliably differentiated by several of the argumentation schemes that were employed. Good writers routinely used more argumentation strategies than poor writers, and used three strategies more often (argument from example, argument from consequences, and argument from expert opinion), not only to warrant their standpoints about topics but also to frame their use of evidence. Although adolescents used these schemes to frame their use of evidence in response to historical topics, the field lacks information on the potential for social studies teachers to use such elements as cognitive scaffolds in an intervention. Therefore, these results led us to examine in the present study whether explicitly teaching argumentation schemes could help students to engage in more sophisticated historical reasoning and compose historical arguments.

Present Aims

The purpose of this intervention study was to evaluate whether middle school students could learn to use specific heuristics proposed by Walton and his colleagues during discussion to learn history content and write historical arguments. We initially suggested that participating teachers use four argument schemes and critical questions: (a) rule, (b) expert opinion, (c) consequences, and (d) cause and effect. Teachers chose the argument from expert opinion, and argument from consequences because they aligned best with the historical controversies during instruction, learning capabilities of students, and the proposed intervention. They eliminated the cause and effect scheme because they believed argument from consequences aligned best with investigation topics. Depending on how students responded to the historical question, potential consequences of each historical event could be very different, which made it easier for students to consider as well as to rebut the other-side argument. The argument from rule scheme was dropped because teachers believed two schemes were developmentally appropriate for a 3-week intervention.

We used an experimental design to compare students’ reading comprehension, their learning, and their ability to write arguments in history after receiving instruction that varied by the focus of discussion (disciplinary vs. generic). Several procedures were in effect over the 15-day instructional period to control for potential differences that were not the focus of the current investigation. Specifically, teachers shared background information on the historical events; the teacher or class read documents out loud when they were first introduced; students discussed the materials; and, students learned about text structure to ensure that they were aware of the central elements of historical writing.

Hypotheses

We predicted that some learning outcomes would be the same after instruction because, as noted above, the two conditions shared some elements. First, because students in both groups read and discussed historical documents, we anticipated that they would perform equally well at posttest on a generic reading comprehension measure. Our comprehension measure therefore served as a nonequivalent dependent measure (Shadish, Cook, & Campbell, 2002). Second, we predicted that students in both groups would demonstrate similar gains on two generic writing outcomes (overall persuasiveness and length of essays), because students wrote the same number of essays, and received instruction on the inclusion of text structure elements that were targeted for instruction.

In contrast, we expected students in the experimental condition to demonstrate greater benefits on two measures. We expected that they would learn more historical content and we expected they would have better performance on a disciplinary writing outcome (a global measure of historical thinking, or the extent to which they demonstrated sophisticated substantiation, perspective recognition, contextualization, and rebuttal in their papers) after participating in disciplinary discussions. We anticipated these outcomes because the experimental instruction was intended to promote domain-specific thinking whereas the comparison condition was designed to promote general understanding.

Method

Participants and Setting

In all, 151 students from a suburban district in the northeastern United States participated in the study. The sample included 99 sixth- and 52 seventh-graders ranging from 11 to 14 years old who (a) returned consent forms approving participation in the investigation and granting permission to access academic records, (b) completed the Gates MacGinitie Reading Test [GMAT-4; MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2002], and the essay composition subtest of the Wechsler Individual Achievement Test [WIAT-III; Psychological Corporation, 2009] before starting the investigation, and (c) participated in all pretest and posttest assessments. There were 76 students (36 male, 40 female) who were randomly assigned to the experimental condition, and 75 students (39 male, 36 female) who were randomly assigned to the comparison condition. The initial pool included 167 students, but 16 students were excluded from all analyses (3 were absent during testing, 3 were absent 1 week or more during the study and 11 did not provide consent).

The students were all proficient speakers of English. Fifty-three percent of the participants qualified for free or reduced-price lunch. Students’ ethnicities were reported as: 87% Caucasian, 7% African American, 4% Hispanic, and 2% Asian. Twenty students (12%) received special education services and had Individualized Education Plans (IEPs). Among the students with disabilities, 18 were identified with specific learning disabilities (SLD) in reading, math, and/or writing, with two students receiving services for emotional and behavioral disorder (EBD). None of the students had visual, auditory, or physical disabilities that impeded their progress in the general education curriculum.

Students came from six classrooms (4 sixth- and 2 seventh-grade) located in a Title I middle school. The six classroom teachers of record, two retired elementary school teachers, and the first author (a special education teacher in the district) delivered instruction. Five of the participating teachers were female, and the remaining four teachers were male. The four sixth-grade teachers had taught an average of 15 years (range = 3–29), and the two seventh-grade teachers had taught an average of 16 years (range = 6–26). The remaining three teachers had an average of 28 years of teaching experience (range = 9–35).
Design

We used a pretest–posttest control group design, forming two conditions by random assignment. After randomly assigning to condition, students were randomly assigned a second time within conditions to form small groups of six to eight students, resulting in two experimental and control groups (a total of four groups) in each classroom. The second randomization allowed us to reduce student-to-teacher ratios for discussions. We believed this would foster greater participation in dialogue, and enhance learning outcomes in both conditions. Thus, there were a total of 12 experimental and 12 comparison groups in six classrooms. Table 1 presents demographic information with respect to participants’ grade, gender, ethnicity, special education classification, eligibility of free and reduced meals, and academic characteristics.

Table 1
Demographic and Academic Characteristic by Experimental and Control Condition Prior to Instruction

<table>
<thead>
<tr>
<th>Variable</th>
<th>Experimental</th>
<th>Comparison</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade</td>
<td>76</td>
<td>75</td>
<td>.953</td>
</tr>
<tr>
<td>Sixth</td>
<td>50</td>
<td>49</td>
<td>.569</td>
</tr>
<tr>
<td>Seventh</td>
<td>26</td>
<td>26</td>
<td>.380</td>
</tr>
<tr>
<td>Gender</td>
<td>Female</td>
<td>36</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>40</td>
<td>36</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Caucasian</td>
<td>67</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Students w/IEPs</td>
<td>Yes</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>64</td>
<td>66</td>
</tr>
<tr>
<td>Eligible for free-reduced lunch</td>
<td>Yes</td>
<td>39</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>37</td>
<td>34</td>
</tr>
<tr>
<td>Academic characteristics</td>
<td>Writing</td>
<td>97.29 (7.17)</td>
<td>97.89 (6.77)</td>
</tr>
<tr>
<td></td>
<td>Students w/IEPs</td>
<td>89.58 (7.75)</td>
<td>92.67 (11.09)</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
<td>45.14 (3.05)</td>
<td>46.27 (3.28)</td>
</tr>
<tr>
<td></td>
<td>Students w/IEPs’</td>
<td>39.64 (3.45)</td>
<td>40.32 (2.47)</td>
</tr>
<tr>
<td></td>
<td>PSSA Reading (Overall)</td>
<td>1374.03 (192.64)</td>
<td>1402.40 (190.62)</td>
</tr>
<tr>
<td></td>
<td>Students w/IEPs’</td>
<td>1196.58 (251.10)</td>
<td>1284.44 (330.98)</td>
</tr>
<tr>
<td>Social studies</td>
<td>Grades (1st quarter)</td>
<td>2.92 (0.46)</td>
<td>3.01 (0.23)</td>
</tr>
</tbody>
</table>

Note. The p values for demographic characteristics are based on chi-square goodness of fit models. The p values for academic characteristics are based on analysis of variance (ANOVA) models. Mean scores are listed with standard deviations placed in parenthesis. Demographic. N = the number of students in the experimental and comparison condition; Students w/IEPs = the number of students identified with disabilities and their non-disabled peers; Eligible for free-reduced lunch = the number of students who met the eligibility guidelines for free and reduced lunch prices. Academic. WIAT–III = Wechsler Individual Achievement Test–3rd edition essay composition subtest scaled score; GMRT-4 = Gates-MacGinitie Reading Test–4th edition raw score; PSSA reading = Pennsylvania State System of Assessments reading scaled score; Social studies grades were based on a 4-point scale where 4.0 = A, 3.0 = B, 2.0 = C, and 1.0 = D.
Table 2
Summary of Historical Questions, Source Documents and Characteristics for Each Historical Investigation

<table>
<thead>
<tr>
<th>Question</th>
<th>Source documents</th>
<th>RL</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was the characterization of American life depicted in Happy Days a myth or reality for “All” Americans?</td>
<td>1. America in the 1950s: Background</td>
<td>6.8</td>
<td>610</td>
</tr>
<tr>
<td></td>
<td>2. President Eisenhower’s State of the Union Address</td>
<td>6.9</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td>3. Anna Blanck: VP of Gimbel Brothers Dept. Store</td>
<td>7.0</td>
<td>177</td>
</tr>
<tr>
<td>Pretest/Posttest–Version B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was the characterization of American life depicted in Happy Days a myth or reality for “All” Americans?</td>
<td>1. America in the 195’s: Background</td>
<td>6.8</td>
<td>610</td>
</tr>
<tr>
<td></td>
<td>2. President Truman’s Farewell Address</td>
<td>6.5</td>
<td>182</td>
</tr>
<tr>
<td></td>
<td>3. Article by Journalist Ches Washington</td>
<td>6.3</td>
<td>195</td>
</tr>
<tr>
<td>Investigation 1: Indian Removal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did the United States government have a right to remove the Cherokee Indians from their land?</td>
<td>1. Indian Removal: Background</td>
<td>6.7</td>
<td>436</td>
</tr>
<tr>
<td></td>
<td>2. President Jackson’s Address to Congress</td>
<td>6.5</td>
<td>144</td>
</tr>
<tr>
<td></td>
<td>3. Statement from the Missionaries</td>
<td>7.2</td>
<td>216</td>
</tr>
<tr>
<td>Investigation 2: Mexican American War</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did the United States government have a reasonable (or unreasonable) argument for going to war with Mexico?</td>
<td>1. Mexican American War: Background</td>
<td>7.6</td>
<td>136</td>
</tr>
<tr>
<td></td>
<td>2. James K. Polk: Message on War with Mexico</td>
<td>7.7</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td>3. Statement by Representative Joshua Giddings</td>
<td>7.4</td>
<td>157</td>
</tr>
<tr>
<td>Investigation 3: Gulf of Tonkin Incident</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>If you were a member of Congress at the time this event unfolded would you have voted for or against using force in response to the Gulf of Tonkin Incident?</td>
<td>1. The Gulf of Tonkin Incident: Background</td>
<td>7.2</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>2. President Johnson’s Address to Congress</td>
<td>7.5</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td>3. Statement by Senator Wayne Morse</td>
<td>7.3</td>
<td>179</td>
</tr>
</tbody>
</table>

Note.  RL = Reading levels according to the Flesch-Kincaid Grade Level Readability test.

On the last day of the week, teachers directed students to compose an historical argument, using their document sets and notes in a group setting. Students in both conditions were provided a modified version of De La Paz and Graham’s (1997) DARE (Develop a stance about the controversy, Add evidence from the documents to support your stance, Rebut arguments from the other-side, and End by restating your stance) as a heuristic for including important elements in historical arguments, based on literature that showed students in elementary (Reznitskaya, Anderson, & Kuo, 2007) and middle school (Nussbaum & Edwards, 2007) and middle school.

Table 3
Summary of Procedures for Experimental and Comparison Conditions

<table>
<thead>
<tr>
<th>Stage and topic</th>
<th>Experimental group</th>
<th>Comparison group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretesting:</td>
<td>*12-item multiple choice test on historical information on the four topics, and</td>
<td>*12-item multiple choice test on historical information on the four topics, and</td>
</tr>
<tr>
<td>1. Historical content test</td>
<td>*3-day administration: (a) teacher provides background and reviews sources, (b) students read sources, and (c) students write independently</td>
<td>*3-day administration: (a) teacher provides background and reviews sources, (b) students read sources, and (c) students write independently</td>
</tr>
<tr>
<td>2. Compose argument on America in 1950s</td>
<td>*Instruction with primary and secondary sources along with background information from the textbook.</td>
<td>*Students participated in 2 days of discussion in small groups of 6 to 8 students for each topic</td>
</tr>
<tr>
<td>Similarities in instruction:</td>
<td>*Students participated in 2 days of discussion in small groups of 6 to 8 students for each topic</td>
<td>*Students composed a historical argument after each investigation</td>
</tr>
<tr>
<td>1. Indian Removal</td>
<td>*Students participated in 2 days of discussion in small groups of 6 to 8 students for each topic</td>
<td>*Students composed a historical argument after each investigation</td>
</tr>
<tr>
<td>2. Mexican American War</td>
<td>*Students composed a historical argument after each investigation</td>
<td>*Students used the text structure heuristic to facilitate the production of historical essay elements for Investigations 1 and 2</td>
</tr>
<tr>
<td>3. Gulf of Tonkin Incident topics</td>
<td>*Students used the text structure heuristic to facilitate the production of historical essay elements for Investigations 1 and 2</td>
<td>*Teachers use generic questions to facilitate group discussions and enhance student understanding of historical documents and overall content learning.</td>
</tr>
<tr>
<td>Differences in instruction:</td>
<td>*Teachers use Walton et al.’s (2008) argumentative schemes and critical questions to facilitate group discussions and enhance student understanding of historical documents and overall content learning.</td>
<td></td>
</tr>
<tr>
<td>Posttesting:</td>
<td>*12-item multiple choice test on historical information on the four topics, and</td>
<td>*12-item multiple choice test on historical information on the four topics, and</td>
</tr>
<tr>
<td>1. Historical content test</td>
<td>*3-day administration: (a) teacher reviews background, (b) students read sources, and (c) students write independently</td>
<td>*3-day administration: (a) teacher reviews background, (b) students read sources, and (c) students write independently</td>
</tr>
<tr>
<td>2. Compose argument on America in 1950s</td>
<td>*10-item open-ended test on the sources used in the America in the 1950s documents</td>
<td>*10-item open-ended test on the sources used in the America in the 1950s documents</td>
</tr>
<tr>
<td>3. Reading comprehension test</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
had difficulties transferring new ideas from discussions to their writing and because this scaffold prompted students to develop essays with essential argumentative elements in prior work (De La Paz, 2005; De La Paz & Felton, 2010). Importantly, this tool was faded from instruction after students completed the second historical investigation in both groups, to minimize confounding due to the use of multiple interventions (Shadish et al., 2002) and to promote internalization of the argumentative text structure. There was some evidence of this occurring as teachers reported about one third of all students wrote DARE on their paper and used it as a guide for transforming evidence from the last set of investigation materials to their resulting historical arguments.

**Experimental and comparison group procedures.** Teachers led discussions about the historical controversies with one group of students in each condition. Three volunteer teachers also led discussions with one small group in each classroom. To control for teacher effects, classroom and volunteer teachers were randomly assigned to groups before the start of instruction, then rotated conditions and groups at the end of each investigation cycle.

**Experimental condition.** Students in the experimental condition learned the argument from expert opinion and argument from consequences schemes and five critical questions (Walton et al., 2008). During the first investigation, one class session was used to introduce terminology and establish a rationale for learning argumentation schemes. Teachers shared how these tools could help students make more informed decisions about the historical controversies and gain a better understanding of each document. Furthermore, in addition to developing a standpoint and adding evidence in support of a position—practices taught in the district’s existing curriculum—teachers emphasized that students would now learn to acknowledge, then identify why another perspective was less convincing when writing arguments. The language in the five critical questions was modified to align with authors and language in the documents for each investigation. For example, the term authors was replaced with the actual authors in the primary source documents (i.e., President Jackson, the Missionaries, President Polk, etc.). Similarly, we replaced the actual outcome in question in each historical investigation (i.e., removing the Cherokee Indians, entering war with Mexico, and sending troops to Vietnam) with the word decision for critical questions about argument from consequences.

Next, teachers reintroduced the historical question, “Did the United States government have a right to remove the Cherokee Indians from their land?” and modeled how to use schemes and critical questions to examine each source. Students began to identify and discuss evidence regarding the president’s argument that the government had the right to remove the Cherokee Indians. Teachers introduced the argument from expert opinion and three accompanying critical questions: (a) “Is the author an expert on the historical topic?; (b) Is he a reliable source?”; And, (c) Is what the author is stating based on sound evidence (are their statements based on first-hand or second-hand accounts?) to prompt students to evaluate whether Jackson was accurate in describing the Cherokee in his message to Congress.

Similarly, teachers explained that the argument from consequences scheme required students to consider both positive and negative consequences of possible future actions. Returning to Jackson’s message, students were asked two critical questions for the argument from consequences,1 (a) “What are the good/positive consequences in following through with this decision?” and (b) “What are the bad/negative consequences . . .”). As students shared answers to the questions, teachers provided more general prompting (i.e., tell me a little more about that, does anyone agree or disagree with these comments), recorded notable points on the board, and directed students to make notes.

Teachers used the same procedures the following day (i.e., Day 4 of the lesson sequence) to help students analyze the second primary source document. The first part of instruction centered on identifying the author’s position in relation to the historical question and locating evidence in the document that supported their beliefs. Teachers then reviewed both argumentative schemes and led discussions about the document using the same five critical questions outlined above. The final 10 min of each discussion session was used to answer students’ questions, reexamine the meaning of the argumentative schemes and critical questions, and to review notable ideas from discussions. Procedures used to facilitate discussions on Days 3 and 4 of the instructional sequences were consistent across all three historical investigations.

However, teachers reported needing less time to review the schemes and critical questions in subsequent discussion sessions, leaving more time for conversation.

**Comparison condition.** Students in this condition also discussed the historical controversies on the third and fourth day of each week in small groups. These students also engaged in six discussions, or about 240 min across topics. However, students in this condition discussed responses to a traditional set of questions to promote understanding of text. The questions prompted students to identify the historical actors, determine the author’s purpose and position, identify main or big ideas, and record details that supported main or big ideas: (a) what happened, (b) when did it happen, (c) who were the major persons involved, (d) why did the author present his or her message, and (e) where did the author stand on the historical controversy? These questions gave a clear contrast to those used in the experimental condition with an additional advantage that students were familiar with them. As in the experimental condition, students discussed a document per day for 2 days per week.

**Fidelity of Treatment**

Participating teachers engaged in three, 1-hr training workshops before the study began, with a 1-hr review before each historical investigation. Training focused on use of instructional materials (e.g., scripted lesson plans and PowerPoint presentations) and procedures for delivering instruction were explained and modeled. The process for administering and supporting students as they constructed argumentative essays was also modeled and reviewed. After the participating teachers asked questions, they practiced how to facilitate whole class instruction and small group discussions from scripted lesson plans. Teachers were observed and monitored by the first author in both whole class instruction, and small group discussion. The participating teachers continued these

---

1 Teachers used two critical questions for the argument from consequence scheme focusing only on positive and negative consequences of specific historical decisions to keep the number of questions comparable across conditions.
approximations until they felt comfortable and achieved 95% accuracy or greater on the instructional checklists.

To ensure that instruction was implemented with integrity, we developed a checklist measure of fidelity of implementation based on De La Paz et al. (2014), and aligned with scripted lessons to be delivered on Days 1 through 5 for each historical investigation. As with Reznitskaya et al. (2007), we encouraged teachers to use the scripts as much as possible, and to limit follow-up questions to general prompts such as, “talk a little more about that [student’s name],” or “does anyone agree, or disagree with [student’s name]?” Although lesson plans and procedures differed each day, quality indicators focused on: (a) implements instructional procedures, (b) demonstrates an understanding of content, (c) addresses questions and monitors student understanding throughout the lesson, and (d) interactions between teacher and students illustrate mutual respect. Each indicator was rated on a 0–100-point scale where 0 = Disagree, 25 = Disagree, 50 = Mutual, 75 = Agree, 100 = Strongly Agree.

For whole class instruction that occurred on Days 1, 2, and 5, the first author and one of the retired teachers observed the six classroom teachers three times over the 3-week study, for a total of 18 observations to assess fidelity of implementation of specific performance indicators. Scores across each lesson were averaged to estimate a fidelity score for each teacher. On average, the six classroom teachers implemented 95% of the lesson protocols during whole class discussions with 92% effectiveness. Cronbach’s alpha was used to measure interrater reliability on the four quality indicators between raters. Results indicated a Cronbach’s alpha of .92 for implementing instructional procedures; .96 for demonstrating an understanding of content; .95 for addressing questions and monitoring student understanding throughout the lesson; and .98 for interactions between teacher and students that illustrate mutual respect.

Similarly, group discussions on Days 3 and 4 were audio recorded so they could be compared with scripted lessons for both conditions. A reading specialist who was unfamiliar with the design and purpose of the study and the first author randomly selected one third of the tapes and evaluated fidelity of implementation. Findings on the fidelity of implementation indicators showed, on average, that teachers implemented instructional procedures with 100% accuracy. Interrater reliability on the four quality indicators between raters was .91 for implementing instructional procedures using Cronbach’s alpha; .94 for demonstrating an understanding of content; .87 for addressing questions and monitoring student understanding throughout the lesson; and .97 for interactions between teacher and students that illustrate mutual respect. The findings on the five quality indicators show that, across teachers, there was a high degree of fidelity in the implementation of whole class instruction and group discussion lessons.

**Dependent Measures**

We used three standardized measures to assess preinstructional reading and writing ability: the GMAT-4 and the Essay Composition subtest of the WIAT-III, and students’ PSSA English Language Arts scores. These standardized measures were later used as covariates. We developed an historical knowledge test, and a reading comprehension measure to test students’ general learning. We also developed three writing measures were used to evaluate students’ (a) historical thinking, (b) overall quality, and (c) essay length. Finally, we administered a delayed posttest 8 weeks after the study ended, and scored it for evidence of students’ historical thinking.

**GMAT-4.** This is a norm-referenced, standardized test of reading achievement (Wayman, Wallace, Wiley, Tichá, & Espin, 2007). Level 3 of the GMAT-4 includes two specific subtests: a vocabulary subtest, with 45 items, and a comprehension subtest, with 48 items (MacGinitie et al., 2002). Vocabulary is assessed through questions that refer to brief sentences with embedded target vocabulary items. The comprehension subtest consists of 11 passages; five are fiction, three are related to social science content, and three have natural science content. The GMAT-4 was group administered in each social studies classroom. Students were given 35 minutes to complete the comprehension portion and an additional 20 minutes to complete the vocabulary portion. Classroom teachers scored the GMAT-4 using the published scoring guide. Test–retest reliability coefficients on the vocabulary and comprehension subtests were .95 and .96, respectively (MacGinitie et al., 2002).

**WIAT-III—essay composition subtest.** The WIAT-III is a diagnostic, standardized achievement test designed for students in grades prekindergarten (PK) through 12 and includes 16 subtests designed to measure listening, speaking, reading, mathematics, and writing skills (Psychological Corporation, 2009). Students were provided a prompt and given 15 minutes to compose an essay. Two experienced school psychologists scored essays using published rubrics for determining word count, theme development and text organization, and grammar and mechanics. Total test–retest reliability coefficients for the essay composition subtest are .87 and .88 for sixth and seventh grade, respectively. Age-based reliability coefficients for students ages 12–14 were .78 for theme development and text organization, .84 for grammar and mechanics .84 for word count, respectively (Psychological Corporation, 2009).

**PSSA reading.** The PSSA is a standards-based, criterion-referenced assessment used to measure attainment of academic standards (PSSA, 2010). Every Pennsylvania public school student in Grades 3 through 8 and Grade 11 is assessed in reading and math on this test. Students’ scores from the prior school year were therefore used as an indicator of reading ability. Determinations whether students have achieved grade-level appropriateness in the area of reading are based on five assessment anchors: (a) comprehension and reading skills, (b) interpretation and analysis of fiction and nonfiction text, (c) components of text, (d) literary devices and concepts, and (e) organization of nonfiction text. Scaled scores are reported as advanced, proficient, basic, or below basic. On the most recent sixth grade reading assessment, interrater agreement on the four open-ended questions reportedly ranged from .70 to .79, and interrater agreement on the seventh grade assessment ranged from .71 to .77. The overall test score reliability values for the PSSA have been historically consistent, with scores in the low 0.90s for reading (PSSA, 2010).

**General learning.** We developed two measures to test students’ general learning. These included: (a) historical knowledge test, and (b) reading comprehension test.

**Historical knowledge test.** We created a 12-item multiple-choice test that focused on historical content that students learned during the intervention, and delivered it before and after the
instructional phase of the study. This measure assessed students’ knowledge of content from all four sets of primary documents, America in the 1950s, Indian Removal, the Mexican American War, and the Gulf of Tonkin Incident. Questions were based on the primary sources and varied in difficulty, which aligned with procedures reported by Reynolds and Perin (2009) and Nuthall and Alton-Lee (1995). The first author and one lead social studies teacher at each grade level first generated 10 questions for each historical topic (40 total questions). Through discussion, the team revised and eliminated duplicate questions to ensure that all content was consistent with main ideas and purpose in each source. The final 12 questions were then formatted to align with the type of questions from the teachers’ manual of the school district’s social studies textbook (Bass, 1995).

A former middle school social studies teacher who retired from the school district the previous year reviewed items. The teacher was asked to read the primary source documents and examine the questions generated for content and construct validity. Minor adaptations were made to several of the questions and three questions were removed, resulting in three questions on each historical topic. Reliability was calculated using Cronbach’s alpha (=.87 at pretest, .93 at posttest). Moreover, a bivariate correlation with social studies grades = .524 (p = < .001), suggesting adequate criterion validity and representative of students’ knowledge about U.S. History.

Reading comprehension. It was important to determine the extent to which students understood the set of primary and secondary source materials on life in America during the 1950s, which they were asked to read independently at pretest and posttest, as these documents were to be the basis for composing historical arguments. Thus, a short test was developed to assess students’ factual comprehension of two primary and one secondary source, and was administered once, at posttest. When developing this measure, the first author and lead social studies teachers initially wrote 15 open-ended questions that aligned with key points, concepts, and vocabulary. A middle school reading specialist who was unfamiliar with the study reviewed each question and suggested revisions. In follow-up discussions, the team reduced the comprehension test to 10 questions, and the reading specialist simplified language in a final revision. Cronbach’s alpha = .78, indicating acceptable levels of reliability among the 10 items on the measure. Moreover, criterion validity was established through bivariate correlations with GMAT-4 = .672 (p = < .001), and PSSA Reading scores = .633, (p = < .001), suggesting the test was representative of students’ reading comprehension.

The first author and six teachers of record scored the open-ended responses for the reading comprehension test. Responses were scored using the general scoring rubrics created for evaluating open-ended questions on the reading portion of the PSSA (cf. PSSA, 2010). Each question was scored on a 0–3 scale, where 0 = a response that provides insufficient, or inaccurate material in all aspects, 1 = an incomplete answer (e.g., indicating either a misunderstanding, or no text-based details); 2 = a partial answer to the task (e.g., indicates some awareness, and at least one text-based detail, or 3 = a complete answer (e.g., a correct answer with text-based support). Because of their familiarity with the general scoring rubric and extensive experience scoring open-ended responses, little training was needed to obtain a high level of agreement between the first author and teachers. As suggested in earlier work (Rushton, Brainerd, & Pressley, 1983), the scores by both raters were averaged. Agreement between raters was tested with Cronbach’s alpha. In total, interrater agreement within 1 point was found to be .91.

Writing. Three measures were used to evaluate students’ historical writing. These included (a) historical thinking, (b) writing quality (using PSSA persuasive writing assessment criteria), and (c) essay length. The first two measures were scored using an analytic scoring method that involved rating students’ historical writing on several traits. Specific scoring criteria were established to increase the reliability of scoring and provide a criterion-based assessment (Diederich, 1966). Research suggests that analytic rubric scores can be scored reliably, are sensitive to instruction, and share a significant relationship with performance on standardized assessments (Hammill & Larsen, 1996).

Historical thinking. We used an analytic rubric to judge the extent to which students’ essays exhibited historical thinking, basing judgments on four elements: substantiation, perspective recognition, contextualization, and rebuttal. The analytic quality of each of the four components was based on a 3-point scale using a previously developed rubric (Monte-Sano & De La Paz, 2012; De La Paz et al., 2014; see Table 4). Briefly, substantiation refers to the evidence-based nature of those arguments. Perspective highlights attention to people’s views in the past, including authors of texts and their credibility. Contextualization requires students to ground their arguments in the time and place under consideration, avoid anachronism, and identify relationships between events. Rebuttal refers to opposing side claims, and can presented but not addressed, or be addressed with simple to elaborated counter claims, or critique. Criterion validity was established through bivariate correlations with the WIAT-III r(151) = .480, p = < .001, and students’ social studies grades r(151) = .327, p = < .001. This suggests a significant relationship between students’ historical writing, their general writing ability, and their performance in social studies.

The first author taught two independent raters to score students’ essays in two, 2-hr sessions, by first reviewing the historical background and primary source documents and explaining distinctions between each level on the rubric (see online supplemental Appendix A for an illustration of how students’ papers were scored). Both raters were former social studies teachers who were familiar with the historical topics. The first author and raters independently scored 20 papers, which was followed up with discussions and additional scoring practice. After training, the two raters scored all pretest and posttest essays. The average of their scores was used in analyses of each element. Interrater agreement, calculated by Cronbach’s alpha was .86 (range = .81–.89) across all elements.

Writing quality. This outcome was scored using the PSSA persuasive writing rubric (PDE, 2010). Each essay was separately scored using the PSSA quality index and considered a generic quality measure. The PSSA quality index captures five dimensions of effective writing: focus, content, organization, style, and conventions. The quality of the five dimensions is based on a 4-point scale (PSSA, 2010). We use bivariate correlations with WIAT-III to establish criterion validity r(151) = .548, p = < .001, and to provide evidence that the test outcomes are representative of students’ writing ability.
Table 4

Analytic Rubric of Historical Thinking

<table>
<thead>
<tr>
<th>Substantiation</th>
<th>Perspective recognition</th>
<th>Contextualization</th>
<th>Rebuttal</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. (a) Facts/quotes are explained AND linked explicitly to a conclusion. AND Explanations are consistently accurate.</td>
<td>(a) Evaluates the authors’ perspectives (e.g., discusses reliability/trustworthiness) OR (b) Reconciles multiple authors’ perspectives (e.g., compares the ideas in the documents)</td>
<td>(a) Integrates background information and evidence from the documents in an explanation or conclusion. OR (b) Uses background information and evidence from the documents together to draw a conclusion or make an inference.</td>
<td>Opposing side claims are clearly presented and drawn from the documents. Writer does not simply consider an opposing side but offers an explicit rebuttal, evaluation of evidence, or reconciliation of opposing views.</td>
</tr>
<tr>
<td>2. (a) Facts/quotes are presented without explanation, but are explicitly linked to a conclusion. OR (b) Facts/quotes are explained but not explicitly linked to a conclusion. AND At least one explanation is accurate.</td>
<td>(a) Describes author’s perspective in a way that recognizes text as the author’s point of view Note: The author’s name does not have to be mentioned</td>
<td>(a) Describes background information alongside the evidence from the documents without explicitly connecting them.</td>
<td>Opposing side claims are presented and drawn from the documents, but there is no explicit justification for choosing one side over the other. The author clearly chooses a position, but stops short of explaining why his/her position follows from what is presented (i.e., opposing sides elaborated but not explicitly rebutted or reconciled).</td>
</tr>
<tr>
<td>1. (a) Facts/quotes presented without explanation (may implicitly support a conclusion that is stated). OR (b) Facts/quotes are presented with inaccurate explanations throughout.</td>
<td>(a) Mentions the author(s) (e.g., “According to Johnson...” “The author says...”)</td>
<td>(a) Mentions background information in the documents (example)</td>
<td>(a) Opposing side claims are not drawn from the documents. OR (b) An opposing side claims are distinguished or acknowledged, but not elaborated on.</td>
</tr>
<tr>
<td>0. (a) Minimal evidence OR (b) Irrelevant evidence OR (c) Transcription of document (and nothing else) Note: There may be a claim</td>
<td>(a) Presents evidence from documents as student’s own perspective OR (b) Treats documents as authoritative (e.g., “Document 1 says...” “It says in the document...”)</td>
<td>(a) Minimal background information mentioned OR (b) Student uses anachronisms (e.g., makes a chronological mistake or uses information from another time period w/o noting the different era)</td>
<td>(a) No mention of opposing side claims. OR (b) No clear overarching position, so no clear treatment of opposing sides on the issue of the prompt.</td>
</tr>
</tbody>
</table>

The first author trained two new raters to use the PSSA scoring index to avoid reader fatigue and bias that might have come from having the same individuals read the same essays with more than one coding scheme. The two raters were English Language Arts teachers in the district who had extensive experience scoring sixth and seventh grade students’ writing using the PSSA index. Therefore, training time was used to identify common areas of strength and weakness in students’ writing. After training, the raters scored all of the essays in the data set independently. Cronbach’s alpha was .93 (range = .89–.96 across the four dimensions).

**Essay length.** All essays were also scored on the total number of words written. Length of an essay may be considered a measure of automaticity in writing (Quinlan, 2004; Kobrin, Deng, & Shaw, 2007) and is positively correlated with overall writing ability (Gregg, Coleman, Davis, & Chalk, 2007). For the purposes of this study, the essay length included all words that represented a spoken word regardless of spelling. Length was determined by counting the total number of words in students’ essays. The first author counted the total number of words in students’ essays at pre- and posttest. Reliability checks were completed by one of the retired elementary school teachers. Cronbach’s alpha was computed on a randomly selected pool (15% of the papers) for total number of words written. Reliability for this measure was .98.

**Results**

Table 5 shows the means scores and standard deviations for historical knowledge, reading comprehension and all writing measures at pretest and posttest. Because participants were randomly assigned both to conditions and then randomly assigned again to groups of six to eight students with one instructor for each small group, mean scores for each of the 12 experimental small groups and the 12 comparison small groups at pretest and posttest were used for subsequent analyses. As a further check on the appropriateness of this strategy, we determined intraclass correlations for each writing outcome using individuals as the unit of analysis at pretest. The results indicated that between 19 and 29% of the variance in the outcomes occurred between teachers, thus confirming the need for comparing group outcomes.

To examine whether historical learning varied as a function of the types of discussions students engaged in, we first conducted a series of one-way multivariate analysis of covariance (MANCOVA) on the two general learning measures, historical knowledge, and reading comprehension; and the three writing measures, writing quality, historical thinking, and essay length. The independent variable was treatment condition. If the main effects for treatment condition were significant, follow-up tests (ANCOVAs) were used to identify which knowledge measures...
### Table 5

Means, Standard Deviations, and *p*-Values of Dependent Variables by Condition at Pretest, Posttest, and Delayed Posttest

<table>
<thead>
<tr>
<th>Variable</th>
<th>Experimental (N = 12)</th>
<th>Comparison (N = 12)</th>
<th>p</th>
<th>Experimental (N = 12)</th>
<th>Comparison (N = 12)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>HK</td>
<td>6.25 (0.55)</td>
<td>6.03 (0.64)</td>
<td>.369</td>
<td>9.85 (0.92)</td>
<td>9.20 (1.09)</td>
<td>.008**</td>
</tr>
<tr>
<td>RC</td>
<td>***</td>
<td>***</td>
<td></td>
<td>8.68 (0.47)</td>
<td>8.57 (0.42)</td>
<td>.104</td>
</tr>
<tr>
<td>WQ</td>
<td>13.12 (1.96)</td>
<td>13.53 (2.67)</td>
<td>.671</td>
<td>16.56 (1.19)</td>
<td>16.20 (1.49)</td>
<td>.235</td>
</tr>
<tr>
<td>Focus</td>
<td>3.18 (0.49)</td>
<td>3.22 (0.48)</td>
<td>.825</td>
<td>3.74 (0.47)</td>
<td>3.65 (0.63)</td>
<td>.321</td>
</tr>
<tr>
<td>Organ</td>
<td>2.23 (0.51)</td>
<td>2.46 (0.66)</td>
<td>.356</td>
<td>3.12 (0.89)</td>
<td>3.20 (0.93)</td>
<td>.378</td>
</tr>
<tr>
<td>Content</td>
<td>2.62 (0.51)</td>
<td>2.70 (0.71)</td>
<td>.756</td>
<td>3.53 (0.66)</td>
<td>3.44 (0.78)</td>
<td>.655</td>
</tr>
<tr>
<td>Style</td>
<td>2.17 (0.50)</td>
<td>2.26 (0.55)</td>
<td>.627</td>
<td>2.95 (0.97)</td>
<td>2.81 (0.88)</td>
<td>.543</td>
</tr>
<tr>
<td>Convent</td>
<td>2.86 (0.45)</td>
<td>2.90 (0.59)</td>
<td>.864</td>
<td>3.13 (0.81)</td>
<td>3.20 (0.81)</td>
<td>.755</td>
</tr>
<tr>
<td>HT</td>
<td>4.40 (2.22)</td>
<td>4.26 (2.47)</td>
<td>.890</td>
<td>8.61 (1.11)</td>
<td>7.41 (1.27)</td>
<td>.019*</td>
</tr>
<tr>
<td>Subs</td>
<td>1.55 (0.48)</td>
<td>1.53 (0.38)</td>
<td>.922</td>
<td>2.39 (0.89)</td>
<td>1.93 (0.89)</td>
<td>.029*</td>
</tr>
<tr>
<td>Perspect</td>
<td>0.90 (0.87)</td>
<td>0.84 (0.98)</td>
<td>.872</td>
<td>2.38 (0.16)</td>
<td>1.92 (0.16)</td>
<td>.065</td>
</tr>
<tr>
<td>Context</td>
<td>0.89 (0.31)</td>
<td>0.84 (0.47)</td>
<td>.730</td>
<td>1.48 (0.11)</td>
<td>1.21 (0.11)</td>
<td>.069</td>
</tr>
<tr>
<td>Rebuttal</td>
<td>1.05 (0.94)</td>
<td>1.10 (0.99)</td>
<td>.988</td>
<td>2.58 (0.11)</td>
<td>2.06 (0.11)</td>
<td>.002*</td>
</tr>
<tr>
<td># of W</td>
<td>106.04 (33.89)</td>
<td>108.51 (37.98)</td>
<td>.868</td>
<td>183.82 (30.24)</td>
<td>176.36 (26.01)</td>
<td>.696</td>
</tr>
<tr>
<td>Delayed</td>
<td></td>
<td></td>
<td></td>
<td>7.94 (1.89)</td>
<td>6.80 (2.02)</td>
<td>.000</td>
</tr>
</tbody>
</table>

Note. HK = Historical knowledge; RC = Reading comprehension; WQ = Writing quality; HT = Historical thinking; and # of W = Number of words; Delayed HT = Historical thinking scores on a delayed posttest essay administered 8 weeks after instruction ended. *Treatment-control difference is statistically significant (p < .05). ***indicates that no pretest was administered. The p values are based on follow-up series of ANCOVAs where pretest scores served as the covariate.

---

were statistically different by condition (Huck & McLean, 1975). Covariates met assumptions of randomization, linearity, and homogeneity of regression slopes for each outcome measure as described below.

### General Learning Outcomes

We ran a one-way MANCOVA on historical knowledge and reading comprehension to determine whether gains in students’ learning were contingent on their discussion condition. The covariate on the historical knowledge measure was students’ pretest historical knowledge scores. Because the reading comprehension measure was not administered at pretest, we used students’ scores from the reading comprehension section of the GMAT-4 as covariates. Levene’s test showed that the assumption of homogeneity of variances was not violated for either measure, with *p* = .651 for historical knowledge, and *p* = .740 for reading comprehension.

In this analysis, the combined writing measures were significantly affected by discussion condition, after controlling for differences in students’ initial writing ability Wilks’ *λ* = .60, *F*(1, 22) = 3.86, *p* = .028, partial η² = .412. Follow-up ANCOVAs revealed a main effect for condition on historical thinking *F*(4, 19) = 3.83, *p* = .019, partial η² = .446, but not for overall writing quality or length of essay. These findings suggest that students’ posttest essays were similar in terms of generic writing ability, but that students who were guided by a disciplinary form of discussion more consistently included elements of historical thinking in their writing than students in the comparison condition.

### Writing Outcomes

Next, we performed a one-way MANCOVA on historical thinking, overall writing quality, and length of essay. The covariates were relevant pretest scores. Levene’s test indicated that the assumption of homogeneity of variances was not violated for the dependent measures (*ps* > .05) and Box’s M test suggested that variance-covariance matrices were similar across cells, *F*(6, 3507) = .169, *p* = .985. The Shapiro-Wilk’s test indicated that measures were normally distributed for each condition (*p* > .05).

In this analysis, the combined writing measures were significantly affected by discussion condition, after controlling for differences in students’ initial writing ability Wilks’ *λ* = .60, *F*(1, 22) = 3.86, *p* = .028, partial η² = .412. Follow-up ANCOVAs revealed a main effect for condition on historical thinking *F*(4, 19) = 3.83, *p* = .019, partial η² = .446, but not for overall writing quality or length of essay. These findings suggest that students’ posttest essays were similar in terms of generic writing ability, but that students who were guided by a disciplinary form of discussion more consistently included elements of historical thinking in their writing than students in the comparison condition.

### Evidence of Historical Thinking in Students’ Writing

We then attempted to identify the factors responsible for the differences in students’ historical thinking scores. To do so, we ran a new set of ANCOVAs to examine differences between discussion conditions on each of the four elements of historical thinking. Students’ pretest scores on each element served as the covariates. These tests showed that compared with students in the comparison condition, students in the experimental group included significantly more so-
Evidence of Historical Thinking in Discussions

Two audiotaped discussions on the Gulf of Tonkin Incident were transcribed and analyzed for evidence of students’ historical thinking. Audio files for each condition were randomly selected from a pool of 10 discussion sessions that occurred within the 4 sixth grade classrooms. This topic was chosen because students had developed greater independence in applying what they had learned the previous two weeks and teachers had grown more effective as facilitators.

We analyzed an excerpt of each form of discussion to highlight differences between students in the two conditions in terms of their historical thinking. In particular, we examined transcripts for instances of substantiation, rebuttal within conversations. Using language in the analytic rubric (Monte-Sano & De La Paz, 2012), when students made a statement, or expanded upon an existing statement and provided facts and/or quotes from the documents to support, or further the statement in dialogue we counted it as an instance of substantiation. A statement was counted as a rebuttal when students cited evidence in the documents as a means to counter or question claims being made by another student or the teacher.

Similar to our findings with historical writing, students in the experimental group provided more instances of substantiation than students in the comparison condition (n = 15, and n = 8, respectively) as well as more rebuttals (n = 10, and n = 2) in the 30-min discussion sessions. Although the majority of students’ rebuttals in the experimental condition were presented in the form of questions, we believed the act of probing further into statements or ideas posed by other participants provides a clear illustration that students not only learned the critical questions, but were actively using them to engage in authentic historical inquiry.

Two excerpts from the transcribed discussions are provided to further illustrate differences in students’ dialogue. The portion of the selected conversations occurred at approximately the same time during the lesson (i.e., right after the historical question was reintroduced, and the primary source document was read aloud by the teacher), and focused on the same discussion point—the author of the document. Although conversations centered on President Johnson, they illustrate how schemes and critical questions served as tools for students to develop an understanding about the Tonkin Incident.

Excerpt 1

T–Let’s think here guys, what is the author’s purpose for presenting this message to Congress, what are the President’s goals here? S1–He wants to attack the Vietnamese because they attacked us. T–So he’s appealing to Congress to get permission to attack? S3–Basically, he wants to go to war because we were attacked a couple of times when we were over there and [unintelligible] even though we shouldn’t have been, the President wants to defend us. T–Does anybody think he has any other motive here? Remember we talked about stopping Communism at this time in history. Does Johnson want to stop Communism as well? S2–He probably does, but what does it matter, either way we are trying to stop Communism or we are attacked, the President wants to stop them over there.

This selection provides a representative example of the dialogue from students in the comparison condition. Asking students generic comprehension questions facilitated their understanding of the source documents, but did not promote a deep level of historical thinking. For example, asking students to think about the author’s purpose demonstrated that S1 and S2 inferred that the President’s purpose in presenting his message to Congress was to “go to war because we were attacked a couple of times when we were over there ...” Though factually accurate, responses from S1, S2, and S3 did not go beyond comprehending the document literally, and contained little to no elements of debate (i.e., opposing-side views, rebuttals, or counterarguments).

Excerpt 2

T–So let’s talk about Johnson then. Did he have the best interests of the US in mind, or is he . . . kind of jumping to conclusions and sending people to war? Should we believe what he’s saying? S1–It’s very hard to say. Yeah, he is blowing it out of proportion. Was it really just one shot, or was it a ton of people shooting, and did they shoot on purpose? Were they shooting at one person or were they just randomly
shooting? S2—But they said, he said, he was on top [flying over], looking down on there and he said that he saw it go the other way [that there was not any shooting or attacks]. I think he said it was either nothing or we shot them. T—Yeah that was Captain James Stockdale... S2—He actually said he didn’t see anything. T—Yeah, Stockdale said he didn’t see anything. So then is Johnson a reliable source? Can we count on what he’s saying? S1—Possibly. T—Possibly? Why? S1—This time the president actually could be, because [...] he was trying to defend the country. S5—... Lyndon Johnson was just trying to get what this report says, he was just trying to defend this country, which actually makes him pretty reliable. He was trying to defend his own country.

In this excerpt, S1 responds to the teacher’s question about President Johnson’s motivation for presenting his message to Congress with a series of rhetorical questions that communicates his belief that Johnson was indeed blowing the incident out of proportion, and S2 highlights the inconsistencies between statements made by Johnson and Captain James Stockdale. We believed that students did so because the argument from expert opinion scheme prompted them to reason about the credibility of each source. In other words, students were likely drawing on the critical question: Is the information President Johnson is stating about the attacks in the Gulf of Tonkin similar or different with what other sources have stated? The rebuttal provided by S2, which was followed by the teacher prompt for students to reconsider the reliability of Johnson, forced S1 and S5—who supported the actions of Johnson—to substantiate their position with additional evidence.

This led us to conclude that although students in both conditions were exposed to the meaning of rebuttal through the text structure heuristic, only by explicitly teaching students to examine and record inconsistencies between sources and/or authors (critical question #2), and investigating whether statements made by the historical authors were based on sound evidence (critical question #3) were students able to understand the value of challenging and/or highlighting limitations in the other-side’s argument. By extension, we believed that the consistent insertion of rebuttals through critical questions, and through examining both the good and poor consequences of certain decisions allowed students in the experimental groups to search more actively for information that substantiated their thinking on the historical question.

**Summary**

These excerpts show tangible differences in how students in each condition responded to their teacher’s questions and the depth to which they considered President Johnson’s potential motives. Students also differed by the extent to which they engaged in rebutting evidence from other-side claims during discussions. Whereas students in the comparison group seemed to accept what the documents stated with minimal interrogation, students in the experimental group grappled more extensively with evidence in the documents to determine a response to the historical question.

The tendency for adolescents and young adults to accept information that aligns with existing moral or social beliefs or to reject information supporting another side in an argument has consistently been observed in the discussions and written essays of middle school, high school, and college-level students (Perkins, 1985, 1989; Perkins, Bushey, & Farady, 1986; Perkins, Farady, & Bushey, 1991). Often referred to in the psychological literature as a confirmation bias (Nickerson, 1998), Perkins and colleagues believed students’ biases toward my-side was one of the most problematic aspects of students’ spontaneous thinking and, as a result, their ability to form objective, logical conclusions in the case-building process. Although it would be difficult to know whether teaching schemes and critical questions altered students’ initial beliefs toward certain controversial historical events, it was clear that doing so taught them to consciously seek evidence on both sides of the question, evaluate it objectively, and draw conclusions that the evidence, in the aggregate, seemed to dictate. More importantly, the data indicated that teaching students to ask critical questions taught them to give primary attention to other-side arguments for the purpose of pointing out their weaknesses.

**Delayed Posttest**

Eight weeks after the study ended, all Grade 6 and 7 students in the district wrote an end-of-the-year historical argument. Two teachers at each grade level created a 2-week historical investigation that aligned with the middle school curriculum on the Battle at Gettysburg, focusing on who had the advantage in fighting the battle at Gettysburg in July of 1863. Students read document sets, which included the historical question, background information, and two primary source documents. Classroom teachers reviewed documents and background information in one 40-min class period, but students wrote arguments without discussing the documents or working in small groups. Grade-level teams including reading specialists and regular and special education teachers scored each essay using the analytic rubric of historical thinking.

Of the original sample of 151 students, scores on the end of the year essay were available for 100 students (mean scores were available for 8 of the 12 experimental small groups and 8 of the 12 comparison small groups). An ANOVA indicated these groups included relatively similar numbers of students from the original conditions $F(1, 99) = 0.96, p = .920$. Furthermore, no significant differences were found between the two conditions at the time of the delayed posttest on gender, $F(1, 99) = 0.04, p = .843$; social studies grades, $F(1, 99) = 0.01, p = .920$; PSSA Reading, $F(1, 99) = 0.23, p = .634$, GMAT-4, $F(1, 99) = 0.15, p = .768$, or WIAT-III-Essay Composition, $F(1, 99) = 0.76, p = .387$.

Table 5 presents the means and standard deviations in historical thinking on students’ delayed posttest measure. We ran an ANOVA to examine the effects of our intervention on the delayed posttest. The results revealed a significant effect for condition $F(1, 14) = 7.27, p = .017$, partial $\eta^2 = .342$. Students who learned argument schemes and to ask critical questions wrote essays that were judged to contain more sophisticated historical elements than students in the comparison condition 2 months after instruction ended. Finally, to test the stability of the historical thinking measure over time, bivariate correlations were run to examine the relationship between posttest and delayed posttest scores. The findings indicated a strong, positive relationship between the two measures ($r = .915, p < .01$) suggesting the historical thinking measure has adequate test–retest reliability.

**Discussion**

The purpose of this study was to compare the effectiveness of facilitating discussions using specific heuristics proposed by Wal-
ton and his colleagues and a traditional form of questioning on students’ abilities to acquire historical knowledge and write historical arguments. In these interventions, we manipulated whether students participated in disciplinary or generic discussions. When tested on knowledge of historical information, students who participated in disciplinary discussions learned more content than students whose discussion emphasized general understanding of sources. Moreover, students who used argumentation schemes and critical questions in their discussions wrote arguments that contained greater evidence of historical thinking than students who participated in traditional discussions. These results are noteworthy as students in both conditions received the same background knowledge and materials, and learned to include important argumentation elements. Finally, our results show that both types of discussion helped students with basic literacy outcomes, as they reached comparable levels of reading comprehension and wrote essays that were similar in length and overall writing quality.

Our results are mirrored in studies by Nussbaum and Edwards (2011) and Song and Ferretti (2013), who showed that learning argumentation schemes and critical questions led to gains in students’ written arguments. We extend prior work on Walton and colleagues’ framework by demonstrating positive effects for its use with a disciplinary literacy task in which adolescents wrote arguments about historical controversies. In the present study, two argumentation schemes (expert opinion and consequences) helped students to learn more historical content and to demonstrate important aspects of historical thinking in their essays, by showing greater substantiation of evidence and more sophisticated rebuttals. Finally, our results extend findings on the benefits of disciplinary discussions on reading (Reisman, 2012) by demonstrating its benefits for disciplinary writing.

We concluded that one of the primary reasons students in the experimental condition may have learned more historical content was because they learned to suspend initial judgments and gained a more comprehensive understanding of the historical controversies. That is, teaching students to think critically about the author taught them not to look for evidence in support of prior beliefs, resulting in a more purposeful evaluation that neither privileged nor excluded information on the basis of existing moral interpretations. This skill is counter to most students’ existing mental representations (Vosniadou, 1994), which often act as a barrier to learning (Chi, 2005; Lombardi, Sinatra, & Nussbaum, 2013). If true, this may explain why students in the comparison condition were often willing to accept authors’ positions without question.

The data also provided an indication that explicitly teaching students to consider and discuss responses to the critical questions that accompany the argument from expert opinion was important in facilitating students’ growth in historical reasoning. In particular, students who engaged in disciplinary discussions learned to actively search for and identify strengths and limitations in each historical perspective, and that historical interpretations must account for all available evidence (Hexter, 1971). This was apparent as early as the first investigation. For example, when discussing Indian Removal, teachers reported that students who participated in disciplinary discussions observed that Jackson was in Washington when he wrote to Congress, and thus, may not have been as familiar with the progress the Indians were making as the missionaries who lived among the Cherokee people. This type of reasoning prompted students to look at other parts of the text (i.e., source information) and to question the beliefs that most of them held at the beginning of the discussion. Students seemed to study each document more closely as a result, and considered motives and facts that went unnoticed by peers in the comparison condition.

By extension, although identifying inconsistencies in the author’s statements during discussions (i.e., Jackson’s inability to make first-hand observations) taught students to think more critically about whether they could trust what was being communicated in the documents, it perhaps more importantly taught them how to effectively articulate limitations in the “other-sides” argument in their writing. Referred to in Walton’s (1996) work as a claim that shifts the burden of proof, the rebuttal is an integral part of historical writing (Coffin, 2006; Gaddis, 2002; Mink, 1987) and an important goal in historical argumentation (DiCamillo & Gradwell, 2012). Thus, we believe that disciplinary discussions not only encouraged students to think about the other-side, but also served as a scaffold for rebutting other-side arguments when they wrote about the historical controversies. This finding seems especially robust as students in the comparison condition were prompted to include and develop rebuttals in their papers through using a text-structure heuristic that included rebuttal and other “same side” historical essay elements.

In addition, and as it relates to our findings for substantiation, while supporting claims with evidence from documents is only one element of historical writing (Reisman & Wineburg, 2008), we believe that greater levels of substantiation in students’ essays was directly connected to the overall development of their historical thinking. For example, we found that students who included rebuttals in their writing also earned higher scores for substantiation. Our interpretation of this relationship was that once students became aware of how to rebut other-side arguments—and how others might challenge their ideas in the same manner—they seemed to work more diligently to systematically collect information that substantiated their position and prepare for discussions where their ideas or positions were often questioned, or even challenged. It seemed clear in our analysis that these practices were transferred to and displayed in students’ argumentative writing. This may be especially useful in future instruction, if students use critical questions to deliberate more fully on evidence about particular historical situations.

Finally, asking students to consider the argument from consequences scheme and critical questions during discussion also seemed to facilitate historical thinking. Students used this scaffold to interpret and integrate information from different sources, speculate about alternate actions, and reformulate information in support of a point of view. For example, when discussing the Gulf of Tonkin Incident, one audiotape excerpt revealed that a teacher asked her students about the positive and negative consequences about using force in response to the reported sinking of an American ship. Students began by stating that there were “always good and bad consequences in every war” and that “we could lose a lot of good people being beaten over a war.” When asked to think further about the negative consequences of using force in response to the attacks, other students suggested, “this goes with every single war there’s ever been, the relationship [deteriorates] between the countries” and that “it takes a lot of supplies, a lot of people,” and “...time, money, and medication...” The teacher pursued the question further by asking, “does the positive consequences of using force over there outweigh the negative consequences, yes or no?” Students’ final comments were nuanced and
reflected an understanding not only of the particular incident but of the events that led up to the Gulf of Tonkin Incident, stating that, "I think they kinda do" and "We can always rebuild, we can always get those resources, and literally the France thing [formerly French Indonesia] and communism, we’re taking more of a bigger picture."

Limitations

The findings reported in this investigation are not without limitations. We first acknowledge that while students in the experimental condition learned more historical content and wrote arguments with stronger evidence and more sophisticated rebuttals, they did not show improved perspective recognition relative to students who participated in generic discussions. This finding may have been due to similarities between the experimental and comparison condition as students in both groups answered questions about authors’ perspectives. Students in both conditions reached comparably high scores for perspective recognition.

Moreover, students in the current study did not differ statistically in terms of their ability to contextualize their arguments. This may have been due to the fact that students in both conditions learned about background information on the first day of each investigation and that none of the critical questions in the experimental condition prompted students to think about past events in accordance with events or perspectives from the specific time periods that they were learning about. Contextualization, or the ability to recognize that beliefs, institutions, and overall values belong to a specific sociohistorical moment (Carr, 1964) is typically difficult for students to grasp (Husbands, 1996; van Drie & van Boxtel, 2008; Reisman & Wineburg, 2008; VanSledright, 2002) and has been difficult to improve in other interventions with adolescents (De La Paz, 2005; Nokes, Dole, & Hacker, 2007). Further research is needed to explore how Walton’s critical questions can be further elaborated, or even new critical questions added, to help students to link related events and situate primary sources in appropriate time, place, and setting.

We also acknowledge limitations in our decision not to type students’ papers and to correct errors in spelling before scoring writing-dependent measures, and in the absence of a generalized learning-dependent measure. Critics contend that scoring essays without such corrections does not allow us to rule out a potential presentation bias from raters who were charged with scoring overall writing quality and historical thinking (Graham, Harris, & Hebert, 2011). However, we believe this limitation does not undermine our results because students generally master handwriting by fourth grade, we included benchmark papers during training from students with a range of handwriting and spelling proficiency, and because the available research also suggests that presentation effects are less problematic with older students (De La Paz et al., 2013). Finally, although we were unable to obtain a generalized learning measure, students did complete an end-of-the-year writing sample after the study ended. Results from analysis of their performance indicated that students who participated in disciplinary discussions retained their advantage in historical thinking 2 months after instruction ended; a finding that mitigates the impact of this limitation.

Conclusion

At the most basic level, the goal of argumentative dialogue is for one or both participants to prove that their beliefs about a certain issue are true. At the same time, while one participant works to prove his or her thesis is valid, the goal of the other participant is to raise doubts, or to shift the burden of proof (van Eemeren & Grootendorst, 1992; Walton, 1996). The outcomes reported here support teaching students to use argumentation schemes and critical questions as a means to facilitate the latter objective (Andriessen, 2006; Nussbaum, 2008), while also extending prior findings by demonstrating benefits in using this approach to learning historical content and promoting more advanced historical reasoning. Although formulating counterarguments and rebuttals to other-side arguments requires substantial epistemological sophistication (Nussbaum & Kardash, 2005), Walton et al.’s (2008) framework provides a promising tool for adolescents to use during discussion, and as a cognitive scaffold for writing historical arguments.

Our findings have implications for practitioners as 43 states have now adopted the CCSS. Beginning in sixth grade, students are expected to write arguments that advance a claim that is best supported by evidence from multiple primary and secondary sources (CCSS, 2010), integrating aspects of a text that reveal an author’s point of view or purpose (CCSS.ELA-Literacy.RH.6–8.6). Unfortunately, many teachers feel unprepared for the specialized demands called for by the CCSS (Murphy & Regenstein, 2012), especially in regards to disciplinary thinking. We believe our work, and that of others in history education (Barton, 2008; DiCamillo & Gladwell, 2012; Monte-Sano, 2010; Reisman, 2012; VanSledright, 2004) provide viable ways to explore, resolve, and inform teachers about the questions and challenges for teaching content instruction with a disciplinary emphasis.

Finally, to our knowledge this is the first study to apply Walton et al.’s (2008) framework as a means to enhance historical writing. Given its short duration, it was encouraging to observe meaningful effects in improving students’ historical writing. More extensive investigation is needed to explore the application of other argumentation schemes and critical questions in history classrooms by examining the degree to which argumentation schemes may be used to help students to analyze important causative factors in history (i.e., economic, political, religious), arguments about historical actors’ motivations, goals, and tactics, and so on. Additional research is also needed to explore the relationship among historical topics, primary source materials, and the utility of specific argumentation schemes and critical questions. Our prior work (De La Paz et al., 2012) revealed the importance of topic (i.e., central historical question) and materials (e.g., documents that include contrasting perspectives and information) in determining which argumentation schemes were relevant. This is likely true because interlocutors use specific argumentation strategies to accomplish particular rhetorical purposes (van Eemeren & Grootendorst, 1992; Walton, 1996; Walton et al., 2008). Finally, future research should explore new sets of critical questions to help students to read evidence from the perspective of those who created it, to place that evidence into context when constructing interpretations of historical events, and to explore benefits in using this approach with other aspects of historical reasoning.
References


Members of Underrepresented Groups: Reviewers for Journal Manuscripts Wanted

If you are interested in reviewing manuscripts for APA journals, the APA Publications and Communications Board would like to invite your participation. Manuscript reviewers are vital to the publications process. As a reviewer, you would gain valuable experience in publishing. The P&C Board is particularly interested in encouraging members of underrepresented groups to participate more in this process.

If you are interested in reviewing manuscripts, please write APA Journals at Reviewers@apa.org. Please note the following important points:

- To be selected as a reviewer, you must have published articles in peer-reviewed journals. The experience of publishing provides a reviewer with the basis for preparing a thorough, objective review.

- To be selected, it is critical to be a regular reader of the five to six empirical journals that are most central to the area or journal for which you would like to review. Current knowledge of recently published research provides a reviewer with the knowledge base to evaluate a new submission within the context of existing research.

- To select the appropriate reviewers for each manuscript, the editor needs detailed information. Please include with your letter your vita. In the letter, please identify which APA journal(s) you are interested in, and describe your area of expertise. Be as specific as possible. For example, “social psychology” is not sufficient—you would need to specify “social cognition” or “attitude change” as well.

- Reviewing a manuscript takes time (1–4 hours per manuscript reviewed). If you are selected to review a manuscript, be prepared to invest the necessary time to evaluate the manuscript thoroughly.

APA now has an online video course that provides guidance in reviewing manuscripts. To learn more about the course and to access the video, visit [http://www.apa.org/pubs/authors/review-manuscript-ce-video.aspx](http://www.apa.org/pubs/authors/review-manuscript-ce-video.aspx).