

Reading and writing from multiple source documents in history: Effects of strategy instruction with low to average high school writers

Susan De La Paz^{a,*}, Mark K. Felton^b

^a University of Maryland, USA

^b San Jose State University, USA

ARTICLE INFO

Article history:

Available online 27 March 2010

Keywords:

Historical reasoning
Disciplinary literacy
Argumentation
Pre-writing
Strategy instruction

ABSTRACT

This study examined the effects of historical reasoning strategy instruction on 11th-grade students. Students learned historical inquiry strategies using 20th Century American history topics ranging from the Spanish–American war to the Gulf of Tonkin incident. In addition, students learned a pre-writing strategy for composing argumentative essays related to each historical event. Results indicate that in comparison to a control group ($N = 79$), essays written by students who received instruction ($N = 81$) were longer, were rated as having significantly greater historical accuracy, were significantly more persuasive, and claims and rebuttals within each argument became more elaborated. Importantly, students in the control group read the same primary and secondary source document sets, and received feedback on written essays on the same topics.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Academic literacy is critical to success in American schools and professional life. As young adolescents prepare for the demands of high school and college classrooms, they must learn to read and write increasingly complex and specialized forms of text. They must go beyond telling what they know with text, to engaging in knowledge construction, reasoning, and argument with text (Bereiter & Scardamalia, 1987). And as they make the transition from basic to academic literacy, adolescent writers must adapt to a variety of tasks, rhetorical structures and standards that vary from one discipline to the next (Ackerman, 1991; Beaufort, 2004; Geisler, 1994). Unfortunately, data from the National Assessment of Educational Progress (NAEP) suggest that most adolescent writers are not prepared to make this transition. A recent NAEP report found that fewer than 24% of 12th- graders wrote essays at or above the proficient level—a standard defined as “solid academic performance...demonstrating competency” (Salahu-Din, Persky, & Miller, 2008).

In addition, while the percentage of 12th-graders performing at or above basic levels increased from 74% in 2002 to 82% in 2007, students who write at this level are often unable to provide adequate support for their positions. Moreover, in both years' samples, students' difficulties were most pronounced on writing tasks that

required structured responses to analytical or argumentative prompts, precisely the kinds of disciplinary writing emphasized in secondary, post-secondary and professional settings. In short, the data suggest that there is a large population of students who struggle with the demands of academic literacy in writing. In response, national panels on literacy like Writing Next (Graham & Perin, 2007a) and the National Commission on Writing (Magrath & Ackerman, 2003) have called for increased attention to adolescent writing instruction that is embedded in content area courses.

This need for discipline-based writing instruction is particularly evident in the history classroom. Over the past 15 years, the history curriculum has undergone significant reform, placing greater emphasis on reading and writing from primary source documents. Students must read first- and second-hand accounts of events in history and then write essays that either advance an interpretation of events or advocate a position based on information available to decision-makers at the time. Whether taking the role of novice historian (Wineburg, 2001) or democratic citizen (Barton, 2005), students must use historical evidence drawn selectively and critically from primary source documents to write well structured and well substantiated written arguments. Unfortunately, high school history students' essays tend to list facts rather than argue claims (Rothschild, 2000; Young & Leinhardt, 1998), leave arguments unelaborated (Nystrand & Graff, 2001), and draw on source evidence indiscriminately (Britt & Aglinskis, 2002; Perfetti, Britt, & Georgi, 1995). These findings, echoing those of the NAEP, suggest that a majority of adolescent writers struggle at authoring a simple substantiated argument in the discipline.

* Corresponding author. Address: 1308 Benjamin Bld. College Park, MD 20742, USA. Fax: +1 301 314 9158.

E-mail address: sdelapaz@umd.edu (S. De La Paz).

1.1. Writing intervention research

Writing intervention research has generally followed the cognitive processing model proposed by Hayes and Flower (1980), or as revised by Hayes (1996, 2006). Of the many reasons for this, first, the examination of sub-processes (e.g., goal setting as part of planning) and reciprocal relationships (e.g., revision while planning) has proven ripe for empirical validation. One of the most successful lines of writing intervention research has been a cognitive apprenticeship model of instruction as summarized by Graham and his colleagues in several recent meta-analyses (Graham, 2006; Graham & Perin, 2007a, 2007b). Often times referred to as strategy instruction or as self-regulated strategy instruction (Deshler & Schumaker, 1986; Englert et al., 1991; Harris & Graham, 1996; Wong, 1997; Wong, Butler, Ficzero, & Kuperis, 1997), both models situate writing as a purposeful activity, and apply several heuristics in an expert-novice apprenticeship.

To illustrate, teachers discuss or explain features of writing that are valued, often using models and text structure as tools. Moving beyond direct instruction, teachers model underlying processes such as planning or revising while thinking aloud during the composing process. A key element in instruction is the collaborative, co-constructed nature of student work. Students compose with their teachers, and then compose together in small groups, before attempting to apply the sub-processes alone. Mnemonics often are used to remind students of key steps in composing or of important features of text structure (or both). Teachers also allow students to work towards mastery, i.e., they permit students to attempt to apply the strategy independently on more than one occasion until students are able to do so without scaffolds (such as visible reminders or assistance from teachers). Importantly, large and consistent effect sizes have been reported for strategy instruction research (Graham & Perin, 2007a; weighted $ES = 0.62$) with even greater effect sizes for students in which strategy instruction included self-regulation ($ES = 1.14$).

With few exceptions, the work evaluated by Graham and his colleagues has focused on generic genres such as story, explanation, and persuasion (a particular species of argumentative writing). Of importance to the present study, much of the intervention research on argumentative writing has addressed the need to help students write a structured argument that includes claims, counterarguments and evidence, by teaching students these components as elements of text structure, to be included in an essay. Little attention has been given to the question of how argumentative writing in a specific discipline might place additional demands on the writer (e.g., De La Paz & Graham, 1997; Graham, Harris, & Mason, 2005). For example, document-based writing in history requires a host of strategies for reading texts with rhetorical purposes in mind. To write such an essay, students must be able to represent the arguments that they encounter across documents, compare documents to examine and critique competing claims, and weave together evidence to construct their own line of argument (Britt, Rouet, Georgi, & Perfetti, 1994; Kuhn, Weinstock, & Flaton, 1994). Thus, the writing task for adolescent history students requires not only processes for writing an elaborated argument, but also processes for constructing a single argument from multiple, sometimes conflicting, sources of evidence.

Over the past 15 years, a body of literature that addresses these discipline-specific demands of history writing has developed. These studies document the ways in which students must interpret the writing task, read and evaluate documents, and construct evidenced arguments. Most notably, research has focused on how students build a global argument from the local arguments that they encounter across multiple source documents (Britt et al., 1994), and how they read, understand and cite evidence from multiple source documents in their essays (Britt & Aglinskis, 2002; Young

& Leinhardt, 1998). But much this work has been descriptive and analytical in nature, leaving open questions about the design and efficacy of instructional interventions aimed at improving the quality of students' essays. More research is needed to understand how interventions that address each of the skills of disciplinary writing in isolation can be combined to improve the breadth, depth and quality of students' arguments in document-based essays.

1.2. Integrating disciplinary reading and writing

Our own conceptual framework for supporting disciplinary writing rests on the notion that reading and writing processes must be addressed together and that when they are, they come to reinforce one another. In a landmark study, Young and Leinhardt (1998) outlined the challenges that document-based writing presents for adolescents and investigated the effects of one teacher's curriculum on writing development. They argue that history writing is an embedded activity in which students must learn the purposes, structures and standards of the discourse community (after Gee (1992) and Geisler (1994)). Students must interpret the writing task, read and evaluate documents, and write with a clear and consistent rhetorical plan from start to finish (Young & Leinhardt, 1998). Each of these tasks requires students to fold disciplinary thinking into the writing process, as they represent, select, organize and transform textual information into evidence for a claim. Thus support for writing may help students learn to read multiple source documents with the purpose of identifying and reconciling conflicting points of view. Conversely, support for reading historical documents may help students to develop more sophisticated claims, evidence and counterarguments in their writing, because they have built a sophisticated representation of the arguments found in the texts they have read.

Intervention research in writing, at least in disciplines other than language arts, has largely overlooked this intimate relationship between reading and writing processes. One recent study by De La Paz (2005) attempted to address this oversight. In that investigation, a language arts and social studies teacher coordinated their presentation of distinct self-regulated strategies aimed at pre-writing and historical reasoning in a month-long unit on westward expansion with culturally and academically heterogeneous eighth grade students. Students' responses to document-based questions, when reading multiple source documents, were compared to students in a posttest only control group. Results indicated that students in middle school could accomplish a more sophisticated means for reasoning with documents, and after applying a pre-writing strategy for argumentative essays, produced qualitatively better essays than their peers who did not receive such instruction. Of note, after instruction, students who were in need of special education services wrote essays that were comparable to those written by average and talented writers on most measures.

Despite the improvements evidenced in their writing (demonstrated by gains in length, factual accuracy, and persuasive quality), a close examination of students' written plans and essays produced at posttest revealed that although students learned to use information from the primary and secondary sources in their writing, they did this without citing evidence to support their claims. In other words, students did not learn to situate or explain quotations or other types of evidence from the documents in their papers. As such, students did not develop interpretations that were supported with evidence. Students improved their ability to write persuasive essays, but not their ability to write evidence-based arguments.

1.3. The present study

In this study, the primary purpose was to determine the effectiveness of an integrated reading and writing intervention on

students' abilities to write evidence-based arguments using a cognitive apprenticeship model for instruction (modified SRSD approach). This study draws on and extends previous work by De La Paz (2005) in four important ways. We expected that our instruction would help students learn to develop more sophisticated claims and rebuttals, when they used these elements in their written arguments. We also expected that our intervention would cause students to attend to factual information in documents they were reading, and hoped they would accurately represent this information in their essays. Furthermore, we hoped our intervention would enable students to learn how to use documents to further their arguments.

Second, the prior study (De La Paz, 2005) did not verify outcomes between groups of students who received instruction in historical reasoning and written argumentation with students who had an opportunity to learn from the same materials and write historical arguments without a strategic component to the learning process. We attempted to remedy this by evaluating the efficacy of the combined strategies with a second group of students who receive exposure to the same materials and practice in writing historical essays. In the current study, students in a comparison group received feedback and guided practice in meaningful ways to ensure a more accurate comparison of the experimental instruction than had been accomplished in prior work on this topic. Students in the experimental group learned specific strategies for accomplishing and coordinating the reasoning and writing tasks; however, students in the comparison group benefited from instruction in interpreting documents and writing using methods that were already in place by their teachers.

Third, instruction in the current study was evaluated under more natural teaching conditions as a better test of external validity. Rather than coordinate schedules between a social studies and an English teacher, and provide the lessons to students in a team-taught unit, in the current study social studies teachers provided instruction in both historical reasoning and writing. Social studies teachers are logically positioned as experts when teaching discipline-based writing instruction. Teachers in the intervention condition also decided to distribute the historical reasoning and writing lessons over an entire semester, rather than teach all of the material within a concentrated unit, interspersing other (non-document based) historical content in the intervening weeks.

Finally, students in the current study were older (in the 11th grade) which provided us with several opportunities for changes in our intervention. We upgraded the historical reasoning heuristics to match the needs of more sophisticated learners, and modified specific elements common to self-regulated strategy instruction (e.g., use of self-instructions). Importantly, these changes were done in collaboration with the general education teachers, who based them on the increased levels of independence and cognition seen in their students.

1.4. Hypotheses

1. Students in the experimental group, who receive instruction in analyzing sources and planning argumentative essays, will compose essays with greater use of evidence from documents to further their arguments than students in the comparison group, after adjusting for any pre-existing differences regarding initial writing ability.
2. Students in the experimental group will write argumentative essays with more advanced development of claims and rebuttals, after instruction, after controlling for length of their essays and adjusting for any pre-existing differences regarding initial writing ability.
3. Students in the experimental group will write longer and qualitatively better essays (i.e., greater factual accuracy and overall

persuasiveness), after adjusting for any pre-existing differences regarding initial writing ability.

2. Method

2.1. Participants and setting

The design of the current study was quasi-experimental. A total of 160 11th-grade students (none receiving services for special education or English language development) received instruction from four US history teachers at two schools, in intact classrooms (in a total of 10 different sections). One teacher at each school agreed to have his students serve as the experimental group and another teacher at each school agreed to have his or her students serve as the comparison group. Teachers at the two schools had been assigned different numbers of American history sections; therefore, students at Fulton (a pseudonym) comprised 20% of the experimental condition (one class section) and 32% of the comparison condition (two class sections). Students at San Carlos (also a pseudonym) comprised 80% of the experimental condition (four class sections), and 68% of the comparison condition (three class sections). In the experimental condition, students' ethnicities were reported as: 28.4% Hispanic, 28.4% Asian, 16% Caucasian, 14.8% Filipino, 7.4% African American, and 3.7% Pacific Islander. In the comparison condition, students' ethnicities were reported as: 32.9% Hispanic, 25.3% Asian, 19% Filipino, 11.4% Caucasian, 6.3% African American, and 3.8% Pacific Islander.

Additional information about the schools came from district websites. The enrollment at Fulton was 1880 students, of whom 30.5% met district criteria as socio-economically disadvantaged. Fully 93% of the students graduated during the year in which the study took place, and the school's API score was 708 (a state index measuring the percentage of students who achieve proficiency on state-administered achievement tests in language arts and mathematics; a perfect score on this index = 800). The enrollment at San Carlos was 2380 students, of whom 32% met district criteria as socio-economically disadvantaged. This school reported an 89% graduation rate for the year in which the study took place and an API score of 745. Of the students who graduated, 57% of Fulton's students met requirements for applying to universities in the University of California system, whereas 56% of San Carlos' students met these same requirements. The latter indicator is indicative of students who were capable of entering public 4-year post-secondary schools in California.

Finally, Fulton ran on a block schedule for course delivery, whereas San Carlos utilized a traditional schedule of classes. Students at Fulton completed a yearlong American history course in one semester, attending 90-min class periods each day. They were enrolled in three classes each day, plus a homeroom period. In contrast, students at San Carlos attended seven, 50-min class periods each day, and completed each academic course in one full year.¹

2.2. Demographic comparisons between groups

The students' most recent NCE ranking on Reading and the Language arts portions of the Stanford were available for 113 of the 160 students. There were no significant differences among students in the experimental or comparison group in terms of reading, $F(2, 111) = .180$, $Mse = 129.829$, $p = .673$ (effect size = $-.086$), or language arts $F(2, 111) = .127$, $Mse = 108.656$, $p = .722$ (effect size = $.065$). In addition, grades from the previous semester were available for 123 students in English and 122 in social studies.

¹ While this may appear to be an important distinction, statistical analyses (see Section 3) conducted after the study ended did not indicate different student outcomes that could be attributed to differences in the course structure.

There were no significant differences for grades earned in English $F(2, 121) = 1.372$, $Mse = 2.535$, $p = .244$ (effect size = .21) or grades earned in social studies $F(2, 120) = 1.267$, $Mse = 1.942$, $p = .263$ (effect size = -.19).

Students' scores on the written expression subtest of the Wechsler Individual Achievement Test, administered before the study began, were compared to determine whether the two groups differed significantly in initial writing ability. One one-way analysis of variance test evaluated the relationship between the groups and performance. Students performed at virtually the same levels, $F(1, 158) = .845$, $Mse = 110.594$, $p = .359$ (effect size = .16). See Table 1 for descriptive information regarding students.

2.3. Materials

The composition task chosen for investigation in this study was argumentative essays that involved historical interpretation. This particular genre was selected over expository writing because the participating teachers wanted to prepare students for the SAT, which had just begun to include this type of writing on its exam. As we planned to conduct the study in high school American history classrooms over one semester, the two teachers in the experimental condition selected a pool of six topics based on state content standards for use during instruction and testing. Counterbalancing topics at pretest and posttest was not possible due to the need to teach historical content in chronological order, thus we return to discuss similarities and differences between the first (Spanish-American War) and last topic (Gulf of Tonkin Resolution) at some length at the end of this section. Testing procedures were established to ensure consistent delivery regardless of the condition that students were assigned to or time.

The four topics chosen by the teachers for instructional purposes were as follows: (a) passage of the Eighteenth Amendment to the Constitution and the Volstead Act (i.e., Prohibition), (b) opposition to the first New Deal, (c) the Neutrality Act and America's entry in World War II, and (d) ways to end the Soviet missile crisis in Cuba (see Table 2 for information regarding each document set). We selected several primary source documents including cartoons, speeches, letters, and memorandums, and wrote contextual overviews for each set, using topics that were perceived to have high student interest and to allow teachers to balance coverage of other twentieth century topics without the use of primary sources. Each document set contained a two page historical overview,² a timeline from one of the district's adopted textbooks, situating the event within other American and world events during one or more decades, a writing prompt or historical question, 1–2 cartoons, 1–4 primary textual sources and at most one secondary textual source. Although materials used during instruction were not always the same in terms of type of source, this was not true for materials used for assessment purposes. The pretest and posttest had the same number of cartoons, primary textual sources, and one secondary textual source.

In addition to controlling for type of source, several steps were taken to ensure the pretest and posttest were equivalent in their: (a) general interest level, (b) difficulty, (c) structure, and (d) opportunity for students to respond from either point of view as they covered different historical topics. First, we consulted both district adopted textbooks, a variety of historical databases (including national and university archives), as well as primary sources that the teachers had previously used and searched for new primary sources that could be combined to create a balanced overview of conflicting perspectives or information on each topic, and checked state standards to ensure they were aligned with those content

Table 1
Summary of student characteristics by condition.

Variable	Condition	
	Experimental	Comparison
Gender (N)		
Male	42	44
Female	39	35
Ethnicity (N)		
Hispanic	23	26
Asian	23	20
Caucasian	13	9
Filipino	12	15
African American	6	5
Pacific Islander	3	3
Stanford (M, SD)		
Reading	42.1 (29.1)	44.2 (24.9)
Language arts	50.3 (28.1)	48.3 (30.1)
Recent grades (M, SD)		
Social studies	2.1 (1.1)	2.3 (1.3)
Language arts	3.0 (1.4)	2.7 (1.4)
WIAT (M, SD)	88.8 (12.4)	87.2 (10.4)

Note. Grades converted to 4.0. WIAT = written expression subtest of the Wechsler Individual Achievement Test.

expectations. Then the experimental teachers reviewed the sources and prompts in terms of their appropriateness for students before the study began. The teachers reviewed two or more drafts of each set and made suggestions about additional documents to consider. An American history professor, who was familiar with state standards in social studies at the high school level, then reviewed the teachers' suggestions. Throughout this process, we were especially conscious of the need for students to have accessible materials, and to be able to argue for either position on both sets of controversial events. Revised materials were used for data collection.

Second, after the study ended, a second American history professor and an 11th grade US history teacher who was from one of the participating districts were asked to examine the pretest and posttest materials with the above four criteria in mind. The following is a summary of their comments. In terms of general interest level, the historian thought that the selections for both topics contained material that would be of generally high interest to students. He wrote that, "in the case of the Philippines because, it seems, students usually are fascinated by war and certain parallels might be drawn to the current war in Iraq, and in the case of Vietnam because of its continuing currency. It still engenders emotional reactions from adults and it is very possible that a number of parents could have fought in that war."

In terms of difficulty, the historian noted, "both sets of documents were similarly structured with pro and con arguments distributed proportionately together with supporting documents [cartoons, editorials] generally related to the arguments. The exception to this was the section on yellow journalism which did not seem directly related to the documents [on the Philippines] but more about the origins of the war."

The high school teacher's comments regarding interest and difficulty level were interconnected. She felt that "naturally, interest level is also affected by difficulty" and that,

In several ways, the posttest shows a moderate increase in difficulty over the first. . . The posttest includes slightly more text to read. While each document-based question contains a reference or allusion that students need to understand for full comprehension of the documents, in Doc. 5 of the posttest, 'paper tiger' is much less likely to be familiar to students than William Jennings Bryan's references to US founding ideas in Doc. 3 in the pretest. The posttest also contains more challenging language; although Bryan's speech contains many bracketed synonyms

² Each overview was a 500–1200 word summary from one or more texts and/or textbooks.

Table 2
Summary of document sets (excerpts from all text sources).




Topic/use	Question	Political cartoon	Source attributions
Spanish-American War/pretest	Your task is to take the role of historian and develop a written argument about what happened before the start of the Spanish-American war. If you were living at the time the Spanish-American unfolded, would you have sided with the expansionists or the anti-imperialists?	<p>McKinley serving the Philippine Islands (shown on map, as if a menu) to Uncle Sam</p> 	<p><i>Primary:</i> President McKinley's speech to Methodist church leaders November 1899; Cincinnati Speech by Jennings Bryan January 1899; General (then President) Emilio Aguinaldo Decree for independence from Spain, October 1896 and war message February 1899 against the United States <i>Secondary:</i> Yellow Journalism: the Role of the Press in the United States war against Spain – excerpts from newspaper accounts and periodicals</p>
Prohibition/instruction	The law against alcohol was one of the most controversial issues in the 1920s. If you had been a legislator who was asked to repeal the Eighteenth Amendment, whose side would you have taken?	<p>"The Modern Devil Fish" Showing the liquor traffic trade as an octopus, the knife is the vote for Prohibition</p> 	<p><i>Primary:</i> 1926 Testimony by two officers from the Federal Council of Churches before the Committee on the Judiciary of the United States Senate; 1926 testimony before the Judiciary by Fiorella H. LaGuardia, New York city politician in the House of Representatives; Testimony in 1926 between Senator Reed and Russell Lee Post, a student at Yale University; 1926 testimony before the Committee on the Judiciary by Mrs. Henry W. Peabody, President of the Women's National Committee for Law Enforcement</p>
Opposition to Roosevelt's first new deal/instruction	Many people saw the New Deal as a threat to the relationship between individuals and the government. If you had been living in 1935, would you have supported or opposed Roosevelt's first set of plans to reform the economy and relief for citizens?	<p>On the left is <i>Farm Relief Bill</i>, a farmer says, "Let 'er go, Mr. President" while it crushes taxpayers, businessmen, and consumers. On the right is <i>Café Roosevelt</i>, where the menu says, ready soon: National economic readjustments, increased prices for farm products, etc.</p> 	<p><i>Primary:</i> Senator Huey Long's "Share Our Wealth" Speech, 1935</p>

Table 2 (continued)

Topic/use	Question	Political cartoon	Source attributions
Neutrality and entry to WWII/ instruction	This question is about the question of whether or not the US should have entered WWII. If you had been living before the start of the Second World War, would you have supported the war or have supported the isolationists?	A Neutrality Act that will keep us out of any war. . .and will scrub floors and do the dishes in its spare time	Primary: Franklin D. Roosevelt's "I hate war," address at Chautauqua, New York, August 1936; Winston Churchill's "Letter to President Roosevelt, May 1940;" George A. Dondero's speech: "Are We Being Led Into War?" from the Congressional Record, 1941; Signed Memorandum to the President of the United States, December 1942 by a Delegation of Representatives of Jewish Organizations
Cuban missile crisis/ instruction	If you had been an advisor to President John F. Kennedy October, 1962, would you have supported the naval blockade or an airstrike as a means to end the conflict over Soviet missiles in Cuba?	Kennedy with sign, "For Russia if we are attacked by Cuba"	Primary: Dobrynin's Cable to the Soviet Foreign Ministry, 27 October 1962; Memo of Meeting, Wednesday, October 17th, at 8:30 a.m., and at 4:00 p.m., attended by Rusk, Ball (each part of the time) Martin, Johnson, McNamara, Gilpatric, Taylor, McCone, Bohlen, Thompson, Bundy, Sorensen, Dean Acheson (for a short time); Memo from Acting Secretary of State Ball to Kennedy October 2, 1962; Declassified military draft on advantages and disadvantages for air strike against offensive missile bases and bombers in Cuba
Gulf of Tonkin resolution/ posttest	Your task is to take the role of historian and develop a written argument about the Gulf of Tonkin Incident in 1964. If you were a member of Congress at the time this event unfolded would you have voted for or against the Gulf of Tonkin Resolution? Please choose and defend one point of view in a well-developed opinion essay	Johnson and Ho Chi Minh on Vietnam Escalator, "Our position has not changed at all"	Primary: President Johnson's message to Congress, August 5, 1964; Senator Wayne Morse's speech on Senate floor, August 5, 1964; Two interviews: one with Admiral Stockdale, 5 June 1996, the other with Robert McNamara, June 1996 Secondary: Opinion piece from Washington, D.C. <i>Evening Star</i> , 5 August 1964

to help the student with vocabulary, Johnson's contains a number of rather difficult words that go unexplained, and Stockdale's interview is full of rambling, jargon-laden language that is quite difficult to decode.

Both the historian and teacher noted that the cartoon in the posttest was not dated, which was difficult for students. In addition, because the "Vietnam escalation" cartoon did not clearly identify its characters, students may or may not catch the double meaning of "position."

The historian then considered the issue of whether students would be able to respond either "pro" or "con" with each document set. His comments regarding the pretest materials were that,

"The documents give the students a very clear opportunity to formulate their arguments and opinions on either side of this issue." The teacher agreed, stating, "Each assignment does present a nicely balanced selection of pro, con, and mixed points of view. Both present a mixture of opposing motivations."

The historian summarized his review with the statement, "Taken at face value, both sets of materials are evenly balanced both in tone and type of source. The only apparent imbalance is the nearness of Vietnam to the public consciousness and the remoteness of the Spanish-American War, which could make the Vietnam selection have a higher interest level." The high school teacher summarized her comments by noting that both sets of materials had "a good balance of documents of varying

interest levels; both contain a mixture of dry and emotional rhetoric, and both contain implications of intriguing manipulations behind the scenes.”

Assessment procedures were as follows. Teachers followed identical testing procedures in the experimental and control groups. Teachers reviewed contextual information for one full class period, to introduce each topic. Students then had one full class period to read the document sets and a second full class period to write their response to the historical question/writing prompt. To help students with vocabulary, difficult words were italicized, and synonyms were presented in square brackets (e.g., “*orator* [speaker]”). Students were also told during the assessment that they could ask for definitions of words they did not know.

2.4. General instructional procedures

To be included in the final participant pool, students in both the experimental and comparison conditions: (a) provided parental consent to access academic records, (b) completed the written expression portion of the Wechsler Individual Achievement Test (WIAT; Psychological Corporation, 1992) before instruction began, (c) were present for both the pretest and posttest, and (d) wrote two essays during instruction. No attempt was made to ensure mastery of learning, or to eliminate students from the final participant pool on the basis of limited writing output (e.g., unfinished work) during instruction.

Teachers in both conditions administered the pre- and posttests under the same conditions, which took 3 days (or 150 min, in the case of Fulton, where the school ran on a block schedule) to complete. On the first day of testing, teachers used the contextual materials as the basis for establishing basic concepts for each topic. They supported the materials with supplementary materials that were available in their classrooms (such as maps) and explained key vocabulary in depth. The second and third days were reserved for students to read the sources,

make notes, and write their persuasive essays in response to the prompt. The documents and students' notes were available throughout testing.

Following the pretest, teachers in the experimental condition and teachers in the comparison condition each used two of the instructional topics for students to read and write independent responses. Although students in the experimental condition had reasoning and writing strategies modeled for them on two additional topics, across the two conditions students wrote the same number of essays, and received written feedback (using rubrics specific to each instructional condition, as described later) regarding their performance. Thus, we were able to control for time spent writing and practice effects associated with that experience. Teachers in the comparison condition selected which of the four topics to use for this purpose, and they chose different topics, for various pedagogical reasons. Table 3 provides an overview of the two experimental conditions.

2.5. Experimental condition

The two social studies teachers provided instruction in both the historical reasoning and argumentative writing strategies without collaboratively teaming with an English teacher. As such, instruction was spread over several weeks, allowing us to intersperse different topics for each stage of instruction and provide the teachers time to cover other topics that were aligned with the state standards for their grade level. We modified De La Paz's (2005) historical reasoning strategy for use with older students (see Fig. 1) based on teacher input and a variety of public domain resources on teaching students to use historical documents (e.g. from the National Archives' document analysis worksheets).

2.5.1. Experimental group

Students in the experimental condition learned specific strategies for reconciling primary and secondary accounts that

Table 3
Summary of procedures for experimental and comparison groups.

Stage	Experimental group	Comparison group
<i>Pre-testing:</i> Spanish-American War	<ul style="list-style-type: none"> 150 min divided into 3 days or segments: (a) teacher describes background, and class reviews contextual sources, (b) students read sources independently, and (c) students plan and write independently 	<ul style="list-style-type: none"> 150 min divided into 3 days or segments: (a) teacher describes background, and class reviews contextual sources, (b) students read sources independently, and (c) students plan and write independently
<i>Instruction</i> Similarities	<ul style="list-style-type: none"> Instruction with primary sources was used intermittently. The existing textbook was used for other topics during instruction Teachers provide contextual information about each topic before students read primary and secondary sources Students compose two essays (and 2 days (100 min) were allotted for reading sources and composing each essay) An expectation for production of content was a multi-paragraph theme Students received grades from the regular teacher for work in the study 	<ul style="list-style-type: none"> Instruction with primary sources was used intermittently. The existing textbook was used for other topics during instruction Teachers provide contextual information about each topic before students read primary and secondary sources Students compose two essays (and 2 days (100 min) were allotted for reading sources and composing each essay) An expectation for production of content was a multi-paragraph theme Students received grades from the regular teacher for work in the study
Differences	<ul style="list-style-type: none"> Teachers used the topic of Prohibition to model the planning strategy and discuss a sample essay Teachers used the topic of opposition to the first New Deal to model the historical reasoning heuristics Teachers used the topic of the Neutrality Act and entry into WW2 for an oral debate. Students independently planned and wrote an essay on this topic Teachers used the topic of the Cuban Missile Crisis for students to have independent practice in reading and writing 	<ul style="list-style-type: none"> The San Carlos teacher used the Prohibition topic for students' first writing opportunity. The Fulton teacher used her textbook to cover this topic The San Carlos teacher used his textbook to cover the New Deal. The Fulton teacher used the opposition to the New Deal as her students' first writing opportunity The San Carlos teacher used his textbook to cover this topic. The Fulton teacher used the topic of the Neutrality Act and entry into WW2 as her students' second writing opportunity The San Carlos teacher used the Cuban Missile Crisis for students' second writing opportunity. The Fulton teacher used TCI materials to cover this topic
<i>Post-testing:</i> The Gulf of Tonkin incident	<ul style="list-style-type: none"> 150 min divided into 3 days or segments: (a) teacher describes background, and class reviews contextual sources, (b) students read sources independently, and (c) students plan and write independently 	<ul style="list-style-type: none"> 150 min divided into 3 days or segments: (a) teacher describes background, and class reviews contextual sources, (b) students read sources independently, and (c) students plan and write independently

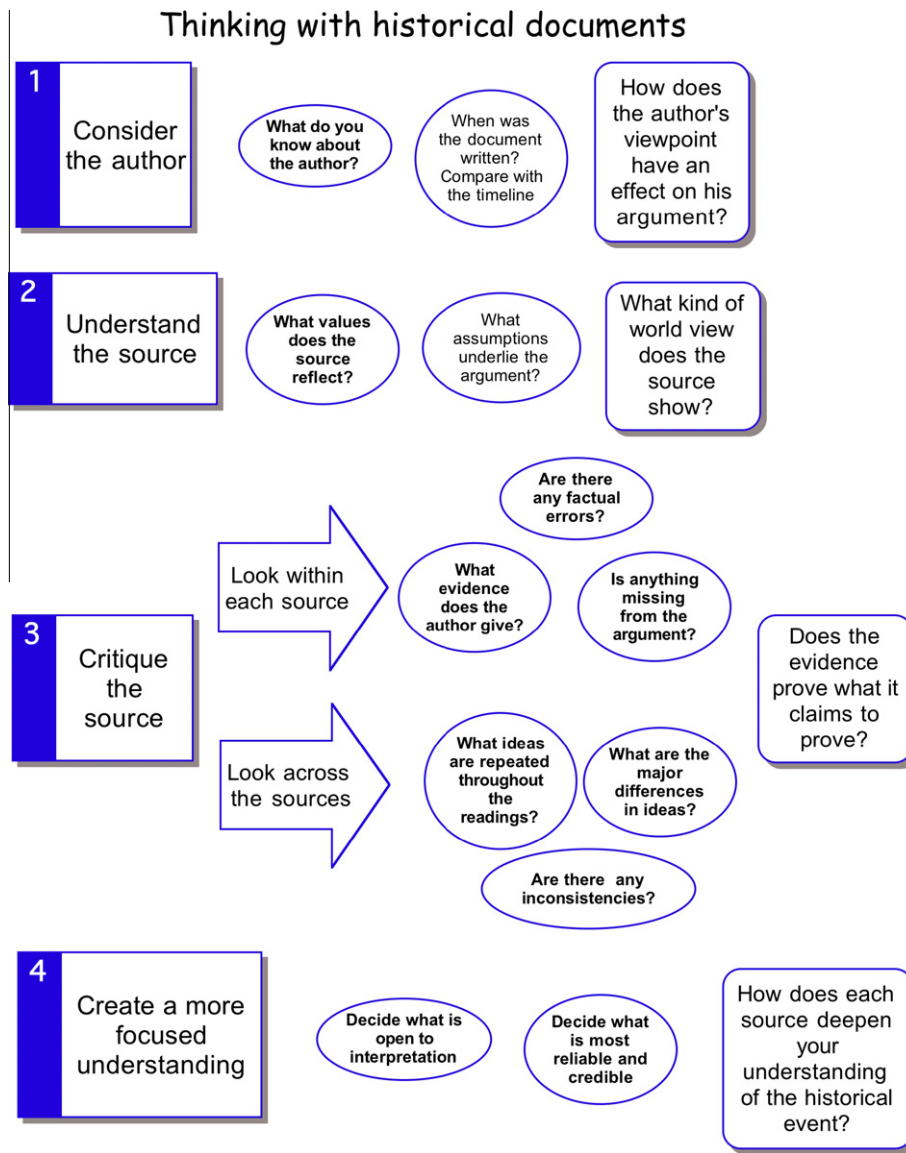


Fig. 1. Historical reasoning strategy.

contained conflicting information or conflicting points of view to build an understanding of complex historical events. They also learned to plan and compose argumentative essays. The strategies, skills, and knowledge were taught using a generic form of strategy instruction, in which teachers provide think-aloud demonstrations followed by verbal scaffolding as a vehicle for helping students gain independence in using strategies and supporting skills (Deshler & Schumaker, 1986; Englert et al., 1991; Schumaker & Deshler, 1992; Wong, 1997; Wong et al., 1997). Strategy instruction has been shown to benefit adolescent students in mainstream classrooms (e.g., Yeh, 1998) and a recent meta-analysis of writing instruction for adolescent students reported that this form of instruction has been especially helpful for struggling writers, with an average weighted mean effect size of .62 (Graham & Perin, 2007b). Students did not use self-regulatory statements in the current study for two reasons. First, the high school teachers felt it was not age-appropriate as students in the 11th grade were thought to be self-directed learners. In fact, once instruction began, teachers referred students to materials they had been exposed to from their English classes that would support writing in the social studies. They did not review

this material, however. Second, there were no students with special needs in the study (although close inspection of the WIAT written expression test results indicated that nearly half of the students in each condition scored between one and two standard deviations below the mean). It should be noted that teachers modeled self-regulatory statements involving problem definition (e.g., "Since I decided to put my thesis statement first, I will write it as the beginning of my introductory paragraph") and planning (e.g., "OK, my next step is to...") throughout modeling and demonstrations.

Five³ stages provided the framework for instruction: *develop background knowledge*, *describe it*, *model it*, *support it*, and *independent performance* (Harris & Graham, 1996). Each of these stages is italicized and placed in parentheses when it occurs in the following discussion of the reasoning and writing program. Each stage of instruction involved one or more class sessions. Because the teachers were teaching two strategies sequentially, we designed an iterative

³ A sixth stage, *memorize it* was not employed.

sequence involving the first three stages twice, once for each strategy, using the topic of Prohibition.

Students involved in this study had each independently read the materials on the Spanish-American War for their pretest; therefore, after the instructional phase of the study had begun, the social studies teacher presented students with two sample argumentative essays (in support of the expansionists and in support of the anti-imperialists) to help students see how the historical content could be developed and supported for either point of view (*develop background knowledge*). He also reviewed the documents with the students and explained how they would learn to read and analyze, then write about similar document sets over the remaining weeks of the semester.

About 2 weeks later, students were given materials on the 18th Amendment, which banned the manufacture, sale, and transport of alcoholic drinks in 1919 (Prohibition). Contextual information included origins of the Prohibition movement as well as its opposi-

tion. The social studies teachers began with a five-minute writing prompt on the effects of alcohol in modern society, and followed it with a discussion of the question, “should alcohol be made legal or illegal?” This was followed by the historical context on Prohibition, including the timeline, and whole class reading and discussion of key terms in the material. In reading the primary sources, teachers expanded on why each side believed what they believed, and students determined whether each source was “in favor” or “against” Prohibition, labeling each document for future reference. Finally, students were shown how to highlight, or “Mark up in some way” evidence that backed up the argument.

Teachers introduced the historical reasoning strategy (see Fig. 1) and an expanded sourcing handout (see Fig. 2; c.f. Britt & Aglinskis, 2002) on successive days. The first component of the historical reasoning strategy, *Consider the Author*, provided a way for students to engage in sourcing (Wineburg, 1991), which in this study was conceptualized as determining not only what was

Consider the author

What do you know about the author?

The **occupation** or **credentials** (President, Senator, Railroad man, etc.) shows the author’s **position** and may give hints about trustworthiness.

Consider how the author came to know about the events. Was the author an eyewitness or have firsthand information? **Or**, was the author relying on hearsay?

When was the document written?

The **date** lets you know how much information the writer had.

If the author wrote the document **after the event occurred**, s/he has the chance to pick and choose what to include in the account.

If the document was written **much later**, does s/he cite official records or put the event into a broader perspective?

How does the author’s viewpoint have an effect on his argument?

The author’s **motivation** in writing the document influences its content.

You should **evaluate the author’s opinion** to see the extent that it seems biased or provides a full and complete account of the events.

Understand the source

What values does the source reflect?

You can guess the values by first identifying the **type or form** of document. Check if the document is a personal letter, an official record, autobiography, scholarly book, sworn deposition, an actual treaty, etc.

The type of document tells whether it is a record without interpretation, interpretation evaluated by informed peers, a concise overview or a summary of agreed-upon information, or mere entertainment.

What assumptions underlie the argument?

Try to **find assumptions**. If you find assumptions use them to critique the source.

What kind of world view does the source show?

World view means the overall opinion about the topic. To figure this out, ask questions about the facts, and come up with your own conclusions.

Fig. 2. Expanded sourcing handout.

known about the author (occupation, position, and how he came to know about the events) but also when the document was written (immediately or some time after the event occurred), which might give an understanding of the author's credibility, and whether there was an obvious motivation underlying the document.

The second component, *Understand the Source*, helped students look more deeply at the values and assumptions in the source. We wanted students to consider the kind of world view the source showed, by identifying the form of document (e.g., personal letter vs. official record) and clues associated with its form (e.g., a record with interpretation or a summary of agreed-upon information). Knowing that high school students might struggle with understanding what "assumptions" and what one's "world view" meant, we defined it simply as the overall opinion of the topic, and attempted to prompt students who were capable of thinking at this level to determine if they could determine one from the source. The teachers found that the analysis of the cartoons could readily be used for this component.

The third component, *Critique the Source*, was designed to teach students the process of corroboration, which involves comparing the details of one source against those of another before accepting its trustworthiness (Wineburg, 1991). Students were taught to look *within* each source and *across* the sources, responding to three questions for each element: (a) What evidence does the author give? (b) Are there any factual errors? (c) Is anything missing from the argument? (d) What ideas are repeated throughout the readings, (e) What are the major differences in ideas? and (f) Are there any inconsistencies? This component ended with a more general question: Does the evidence prove what it claims to prove?

The fourth component, *Create a More Focused Understanding*, prompted students to consider what was open to interpretation, what was most reliable and credible, and how each source deepened their understanding of the historical event. While these components have been presented sequentially, it is important to note that the teachers presented them in context (*describe and model it*) by referring to different documents to highlight the relevance of each component.

After describing and modeling the historical reasoning strategy, the social studies teachers provided an overview about the purpose as well as a description of the writing strategy (*describe it*). Instruction was similar to a previous strategy (De La Paz, 2005; De La Paz & Graham, 1997); the mnemonic STOP reminded students to consider and generate ideas on both sides of an argument before deciding which side to support in their essay. The steps of the mnemonic prompted them to *Suspend* judgment, *Take* a side, *Organize* (select and number) ideas, and *Plan* more as you write.

The teacher showed students a sample structure for writing multi-paragraph essays based on Karras' (1994) suggestions (described in De La Paz, 2005), gave them a copy of transition words from their high school writer's resource handbook, and a sample essay, "Let the Noble Experiment Continue!" We used this structure as teachers requested it, and believed it to be relevant because studies of history classrooms reveal that writing instruction of any kind is uncommon, even among exemplary teachers (Applebee & Langer, 2006; Grant, 2003; Nystrand, Gamoran, & Carbonaro, 1998; Young & Leinhardt, 1998). Finally, the social studies teachers "worked backwards" from the sample essay, showing students how he had located "evidence" in the documents and used this material to create a plan for writing his essay, and how it exemplified elements of text structure. On a subsequent day, the teacher used additional essays, previously written by students, to point out what they were missing or what worked well. Students reviewed the planning process and asked questions.

After describing and modeling both strategies, teachers returned to their general history curriculum for about 2 weeks. They introduced content related to President Roosevelt's New Deal, and

the country's debate over its proposed benefits and drawbacks using the historical context and primary sources that had been prepared for their use. Students used this set of materials to work in small groups (*support it*) applying the historical reasoning strategy components for its analysis. While teachers prompted students to complete each step of the historical reasoning strategy, they did not have students plan or write a corresponding essay. Teachers did however provide students with a rubric that they would be using for subsequent grading of future essays. This rubric integrated text structure with use of evidence, in a modified multi-paragraph theme. Students were to be graded on their ability to present a topic sentence, reason, and use of evidence, as well as present (with evidence) but refute an opposing point of view (with new evidence).

During the final stage of instruction (*independent performance*), students used both strategies to read historical documents and write essays on two topics (Neutrality and entry to World War II and the Cuban Missile Crisis) but received needed assistance from the instructor in applying it. The two teachers used slightly different procedures for introducing and engaging students' interest in the final topics. To illustrate, regarding the topic of the United States' entry into World War Two, one teacher used a formal debate (pro-isolation and pro-involvement) to engage his students, whereas the other teacher created a powerful PowerPoint presentation, integrating movie montage and a song by Woody Guthrie to pique his students' curiosity about the event. In both cases, however, students worked through the cartoon and primary sources, and then planned and wrote an argumentative essay. Students' essays were graded and those grades were discussed when the essays were subsequently returned.

The final document set used during instruction provided the second opportunity for students in the experimental condition to apply both strategies independently. As before, the teachers reviewed contextual information first, explaining who individuals such as Fidel Castro, Nikita Krushchev, Robert and John F. Kennedy, Anatoly Dobrynin were. The teachers reviewed the need to cite evidence from the documents (and one teacher, in particular, told students to refer to the documents by number as a citation) as that had been a general weakness from their first attempt, a few weeks earlier. Students were encouraged to write "For Blockade" or "For Airstrike" on each document, as they read the sources, and to learn about both points of view. Students discussed the sources the next day, and the teachers emphasized that they would need to know both positions so that they could find an opposing view and argue for it, but then oppose it. To reinforce this point, they were also asked to create two plans for composing (one from each point of view), before settling on the one they would use for their final essay. As they had done during the first independent practice attempt, one full class period was allocated for writing the second practice essay.

2.6. Comparison condition

Students in the comparison group did not receive instruction in either the historical reasoning or the argumentative writing strategy. They did however read and write on two topics, between the initial and final testing sessions, receiving instruction from their teachers that we summarize on the basis of classroom observations. Students in the comparison group followed the same content standards, used the same textbooks, and completed the pretest and posttest at the same time as students in the experimental group. Whereas students in the experimental group used two additional sets of primary and secondary sources for modeling and demonstration purposes, students in the comparison group studied the same topics using other materials that their teachers preferred (e.g., introducing the Cuban Missile Crisis with a video segment

from the docudrama, “The Missiles of October,” see Table 3 for other differences).

We made a series of decisions to ensure that students in the comparison group received instruction that could be compared to the experimental group. First, teachers in the comparison condition taught students to reason with the same primary and secondary sources as students in the experimental group. They accomplished this through large group discussions, in which students and the teacher analyzed the documents in turn, and made annotations as insights were made. With respect to writing instruction, teachers were allowed to instruct students in ways that reflected their preferences. In these schools, teachers reviewed a fairly traditional multi-paragraph theme (thesis statement, followed by three arguments, and a conclusion). Students were to choose one side of the issue in question and back it up with three reasons (one teacher told his students that this was not the only way to structure an essay but that it was the most effective for the current purpose due to its solid structure⁴). They were to use material from the documents as well as information about the topic that was not in the documents (e.g., from their textbooks). Historical content from the topics was again integrated into the discussion about writing, and students were encouraged to weigh the importance and validity of the evidence from the documents as well as to recognize that all evidence is not equally valid.

Teachers in the comparison classrooms emphasized other connections between writing in social studies and writing in English and referred students to school handbooks on the writing process. Several pages were devoted to ideas such as “adding support” to a topic sentence, “developing paragraphs with evidence” and using “evidence” to clarify points made in a thesis or topic sentence. The handbooks also provided rules for using quotes, such as “always identify the source of the quote” and “quotations are support for a topic and should be sandwiched between the quote introduction and its explanation.” One handbook provided a particularly useful page on transitions that we distributed to all students in the comparison condition.

Although the two control classrooms were similar in many respects, they were not identical. For example, comparison students did not always use the same materials to study each topic. For example, at San Carlos the teacher used the document set that we designed for the Cuban Missile Crisis topic, whereas at Fulton, the teacher used materials published by Teachers Curriculum Institute (Bower, Lobdell, & Owens, 2004) to cover the same topic. The teachers had agreed to participate in the study and provide opportunities for document use, practice in writing, and to give students feedback on their writing, however, they had differing preferences for instructional materials some of the time. In essence, this description serves to point out that despite differences in how teachers covered the six topics, students in this condition functioned as a comparison group because they read two sets of primary and secondary source materials at different time points, without strategic instruction in making sense of the documents or in writing argumentative essays from them. Although it might have been preferable to have comparison classrooms that were identical, this is extremely difficult to achieve in a school setting, especially when a study involves more than one school district.

Therefore, when introducing the historical reading and writing instruction in the comparison condition, teachers engaged in whole class, week-long explorations of each topic, reviewing the historical context in 1 day, then taking two subsequent days to read and explain ideas, vocabulary, and images in each primary

source, as students took turns reading paragraphs and discussing the concepts and events. They distributed copies of the a rubric for analyzing document-based essays, explained key concepts, and led discussions on its use with essays that students were to write after reading new document sets. Teachers reiterated the importance of responding to the prompt repeatedly; noting as they reviewed documents how key ideas could be used for writing a response. Students were instructed to note important ideas on their documents.

In summary, the comparison condition primarily involved guided, group practice in interpreting primary and secondary source documents, instruction in a basic format for writing multi-paragraph essays, independent practice in writing argumentative essays, and feedback (using a rubric for analyzing document-based questions) designed to support the learning of these skills. Although students in the experimental condition also received instruction in analyzing primary and secondary source documents, planning and writing a multi-paragraph-essay and practiced writing such essays, they also learned to independently use an integrated set of strategic processes, skills, and knowledge for developing historical essays.

2.7. Treatment validity

To ensure that the assessment and instruction was implemented according to plan, we instituted the following procedures. First, the first author observed, or sent a research assistant to observe teachers on days they provided lectures, demonstrations, or modeling, for at least one class period. This included observations of both experimental teachers and both comparison teachers. She also met with the teachers who implemented the experimental intervention at least once weekly, and exchanged written emails 2–3 times during the week. Written field notes were recorded for each observation, detailing the interaction between teacher and students. In rare cases when this was not possible, teachers audio recorded class sessions in which they described or modeled experimental or comparison procedures. After the study ended, an undergraduate student typed summaries of the lessons, which were then compared with written lesson plans (in the case of the experimental condition) or used to verify what the comparison teachers did when they assigned the two topics for practice in reading and writing responses from historical documents. Observations confirmed that assessment procedures were followed carefully by all teachers, and that three of the four teachers (two experimental and one comparison) executed all steps of instruction as planned. The only exception to this finding was a single instance in which we lost written records of how the fourth teacher introduced one session (the second writing session of the second comparison teacher). It should be noted that in all other observations, she followed assessment or instructional procedures as planned.

2.8. Measures

2.8.1. Essay length

All essays were scored in terms of number of words written. This number included all words that represented a spoken word regardless of spelling. Two undergraduate students independently counted the number of words from both sets of essays independently (i.e., pretest and posttest; $r = .99$).

2.8.2. Quality

Two male graduate students who were completing their single subject credentials in social science were recruited to score essays in terms of their overall persuasiveness and historical accuracy. Both were unfamiliar with the purpose, students, and conditions

⁴ To be fair, we did not recognize a limitation in the instructional procedures for students in the comparison group until after the study ended. Teachers did not set a goal for students to generate rebuttals to potential counterarguments in this condition.

in the study independently scored all written papers, after identifying information had been removed, using a holistic rating scale. The students were told simply that 11th-grade students wrote the essays, and that all papers came from two schools.

Each student read the primary source documents and essay questions before scoring the papers. They were also told to read each paper once to get a general impression of the essay to judge and whether it was persuasive, and to read it a second time to identify specific elements: the writer's ability to: (a) interpret the documents and incorporate outside information related to the documents, (b) provide a clear opinion on the topic, (c) support a position with accurate facts, examples and details, (d) weigh the importance, reliability and validity of the evidence, (e) analyze conflicting perspectives presented in the documents, (f) weave documents into the body of the essay, and (g) include a strong introduction and conclusion. The raters' final qualitative judgment considered specific criteria for each of these dimensions (see Appendix A for the rubric).

The score (ranging from a low of 0 to a high of 6) reflected the overall quality of the essay. The raters were also given a representative sample of a low, average, and above-average scoring essay as guides or anchor points for scoring. These essays were obtained from 11th grade students at the same school who were not in the final pool of participants. Interrater reliability (Pearson product-moment correlation) for this measure was .90. Differences between raters were resolved through discussion; the resulting score was based on a mutually agreed upon rating. Final scores were used for all data analyses.

2.8.3. Argument analysis

All essays were scored in a multi-stage process for the development of arguments. Following Toulmin (1958) we define an argument as discourse that is used to support a conclusion on an issue. The central element in an argument is the "claim," a statement that is advanced to support the conclusion, which in turn can be elaborated through the use of argument elements, like data, warrants, backing and rebuttals. In Toulmin's model, these additional elements, are essential to making an argument acceptable by explaining how the author of an argument marshals evidence to support the claim, and details the conditions under which he or she holds the claim to be true. The advantages of using this model to analyze arguments are twofold: (1) it offers a means of identifying the number of claims in support of a conclusion and (2) it provides a means of judging the quality of an argument, by examining the degree to which the grounds for a claim have been elaborated with argument elements. However, one important limitation to using Toulmin's model for our present purposes is that it is not developmental. Toulmin's model can be used to test whether an argument element is effective in providing grounds for a claim, but it cannot be used to distinguish between levels of quality when an element is not effective. For our present purposes we have adopted and adapted Toulmin's framework to allow for a more fine-grained analysis of argument development.

In the first stage of data analysis, essays were coded to identify all claims in favor of or against the position. For the pretest topic, there were 11 possible pro-side and 11 possible con-side claims; and for

the posttest topic, there were 10 possible pro-side and 10 possible con-side claims. In cases where writers misunderstood the prompt and did not take a position on the topic, no codes were assigned. In the second-stage each claim was then coded for its level of development. Claims were assigned one of four levels (described below), independent of essay topic or position taken, based on the degree to which they elaborated grounds for accepting the claim. In the third stage, claims that opposed the position taken by the writer were then identified and coded for the degree to which they were responded to in a rebuttal. Opposing-side claims were assigned one of four levels for rebuttal (described below) independent of the topic, position taken or level of the claim. Student work samples have been included in Appendix B to illustrate how codes were assigned to typical pretest and posttest essays.

The second author, who was unfamiliar with identifying codes labeling students' papers, developed the coding schemes for the argument analysis based on 15% of the data and coded the entire data set. He developed a scoring manual for the pretest and posttest and content elements and for describing the levels of development among claims and rebuttals (see Table 4). A student who was unfamiliar with the purpose, students, and conditions in the study read the scoring manual and practiced scoring essays, using 25% of the papers that had been scored for her as a teaching tool. Training required 5 h, three hours for the pretest papers and two for the posttest papers. She then independently scored a randomly selected set of 25% of the remaining papers. Interrater reliability was .86 overall (.90 at pretest and .81 at posttest) and was calculated as a percentage of exact agreements.

2.8.4. Claims

We first counted the total number of claims, but controlled for length of essay when doing so, as longer essays were likely to permit more claims to emerge. This was done by dividing the number of distinct claims made on either side of an issue in each essay by the number of words in that essay and multiplying the result by 100, to equal the number of arguments per 100 words. Thus, the term "number of claims" refers hereafter to the number of claims—regardless of the degree of their elaboration—controlled for length.

More importantly, we designed a coding scheme for this study to examine the development of students' claims that was intended to be sensitive to advances in novice argumentation. Much of the extant work on argumentation is based on Toulmin's (1958) *The Uses of Argument*, and focuses on the use of evidentiary elements like data, warrants and backing to justify a central claim. However, novice arguers often fail to use these evidentiary elements, and instead use explanations to elaborate claims (Kuhn, 1991). Although explanations do not substantiate an argument, they do represent an appreciation of the importance of plausible and coherent claims in developing a position (Kuhn, Shaw, & Felton, 1997). Thus, the rationale for the levels of quality in our coding scheme is based on evidence that elaborating claims with explanations is an intermediary stage towards developing the ability to substantiate claims with true argument elements.

There are four levels in our coding scheme for the development of claims. At level 1, no claims are used in support of a position.

Table 4
Levels of argument quality.

Level 1	Level 2	Level 3	Level 4
<i>Claims</i>			
No claim is advanced	Claim appears in a list or quote without explanation	Claim is paraphrased, explained or grounded in a historical quote	Claim builds on or substantiates another claim
<i>Rebuttals</i>			
No opposing claims are presented	Opposing claims are presented but not addressed	Opposing claims are addressed with simple counter-claims	Opposing claims are addressed with elaborated counter-claims or critique

Most often, this level was assigned when a student has misunderstood the prompt and failed to take a position on the topic. At level 2, the student lists or quotes claims found in the textual materials without explaining their meaning or relevance. In other words, at level 2 the student simply provides unelaborated claims in support of a conclusion. For example, in the sample pretest (in Appendix B), student A supports US expansion in the Pacific at the turn of the 20th century by claiming that it would allow the country to control trade routes (P2.1) and gain refueling stations (P2.3). These two claims were each assigned a level 2, because they appear without any explanation or support.

At level 3, the student elaborates a claim by explaining its meaning or relevance to the argument. This may be accomplished by clarifying a claim, expanding on it, discussing it, illustrating it with examples (either personal or academic), or using it to counter an opposing-side claim. In other words, at level 3, the student shows movement towards an appreciation for argument elaboration by trying to establish the credibility, if not the grounds, of the claim that they present. For example, in the same essay cited above, student A justifies expansionism by arguing that the country could also gain raw materials from subjugated nations (P2.2). Here the claim is assigned a level 3 because student A goes on to provide a related historical example (expansion into Cuba) to illustrate the point.

At level 4, a student goes beyond explaining a claim to supporting it with evidence (what Toulmin calls “data”) or subordinate claims (“warrants”). At this level, argument elements, which are identifiable in Toulmin’s model, are provided to support a claim. For example, at the posttest (also in Appendix B), student A takes a position in favor of passing the Gulf of Tonkin resolution and supports it with the claim that the US had to step in to contain communism. The claim is first explained (“*Minh and his Viet Minh were clearly looking to take over S. Vietnam and Laos*”) and then supported with both evidence (“*We have found N. Vietnamese shell fragments on decks of our destroyers*”) and a warrant that ties the evidence back to the claim (“*This is a clear act of communist threat to S. Vietnam.*”) level 4 claims are more sophisticated because they establish both the explanatory and the evidentiary strength of a claim and thereby add an element to the structure of the student’s argument. Interrater reliability for assigning the level of claims was .81 (exact percentage) at pretest and .86 at posttest.

2.8.5. Rebuttals

We first counted and analyzed the number of rebuttals that students wrote as we were interested in knowing the degree to which this was a common occurrence. However, we then explored their level of development by ranking rebuttals according to degree of sophistication. Like claims, rebuttals were divided into four levels, representing the degree to which an opposing-side claim was addressed. In Toulmin’s model, a “rebuttal” is a response that neutralizes an opposing-side claim that runs counter to the writer’s own conclusion. To produce an effective rebuttal, an individual must first acknowledge an opposing-side claim, and then either refute it or dismiss it as not relevant to the present case. Again, intermediary approximations of a successful rebuttal are not defined in Toulmin’s model. However, Felton and Kuhn (2002) have found a common approximation of true rebuttal, in which individuals respond to an opposing-side claim with a claim for their own side. In such cases, the additional own-side claim does not directly neutralize the opposing-side claim. However, it suggests that the individual is aware of the need to respond to opposing-side, even if they fail to neutralize it effectively.

Thus, at the first two levels of our coding scheme were assigned to text in which a response to an opposing-side is absent. Level 1 was assigned to essays in which opposing-side claims did not appear at all and level 2 was assigned to an opposing-side claim that

was cited, but not responded to in subsequent text. Level 2 is more sophisticated than level 1 in that the student at least acknowledges alternative perspectives in the text, even if he or she has not provided a response. Level 3 was assigned to opposing-side claims that are addressed with a counter-claim. In these instances, the opposing-side claim is not disproven so much as disregarded in the face of another claim in support of the author’s own side. The shortcoming of this response is that the initial opposing-side claim is never addressed directly with a refutation or dismissal based on relevance. While the strength of the opposing-side claim is not diminished at level 3, at least it has been challenged by what the student sees as a stronger or more compelling claim in support of his or her own position.

At level 4, the opposing-side claim is weakened with evidence or claims that challenge its truth, validity or relevance. As with the coding scheme used for claims, level 4 rebuttals represent the addition of argumentative element that adds to the structure of the student’s argument. For example, at the posttest student A rebuts the opposing-side claim that “*the United States is too much involved in foreign policing*” but asserting that “*if we are not [involved in foreign policing], then we become Isolationist and other countries suffer like in WWII. We left Europe alone and what happened? Hitler came to power.*” While there are certainly flaws to this line of reasoning, it is assigned a level 4 because the student has used historical evidence and counterfactual reasoning to undermine the opposing-side claim. Interrater reliability for coding the level of rebuttals was .84 (exact percentage) at pretest and 1.00 at posttest.

2.8.6. Document use

Two variables were developed to capture the extent to which students used documents in the development of their essays. First, we counted the total number of documents students cited. Second we ranked students’ use of documents, in their essays. In some essays, students made no reference to documents at all (coded as 0). In other essays, students referred to documents by mentioning it or referring to the author (1). Increasing levels of sophistication included use of documents and direct quotes from the documents. In these latter cases students made a claim or gave a point of view and followed it with a quote to substantiate a claim, or used a quote and then analyzed the quote (both valued at 2). Finally in fewer instances, students wrote a quote and used it to further an argument (3).

3. Results

The unit of analysis used across variables in this study was each student’s individual score. We did not model instructional groups within treatment conditions because of sample limitations.

3.1. Pretreatment comparability

At pretest, students’ essays were compared to determine whether the two groups differed significantly. Eight separate one-way analysis of variance tests were conducted to evaluate the relationship between the two instructional conditions and essay length (i.e., number of words written), claims, rebuttals, overall quality (see Table 5 for means and standard deviations for these measures), and document use (total number and highest level). We also examined the percentage of well-developed claims and well developed rebuttals (receiving scores of 3 or higher).

At pretest, students assigned to the comparison condition wrote papers that were longer, $F(1, 158) = 19.15$, $MSe = 206282.22$, $p = .000$ (effect size = $-.57$), with more claims per 100 words, $F(1, 158) = 8.93$, $MSe = 1.10$, $p = .003$ (effect size = $-.49$), and higher

Table 5
Means and standard deviations for length, claims, rebuttals, and document use.

Condition		Experimental		Comparison	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Length	Pretest	195.32	77.92	267.14	124.86
	Posttest	327.86	101.38	281.59	123.17
Claims ^a	Pretest	1.47	1.09	1.96	1.00
	Posttest	1.58	0.74	1.68	.99
Rebuttals ^a	Pretest	.67	1.20	.76	1.02
	Posttest	.82	.94	.31	.63
Document use ^a	Pretest	.53	.93	.84	1.19
	Posttest	1.97	1.2	1.38	1.36

^a Number of elements per essay.

in overall quality, $F(1, 158) = 13.34$, $MSe = 16.07$, $p = .000$ (effect size = $-.55$). Students in the comparison group also wrote papers with better claim development $F(1, 143) = 18.46$, $MSe = .55$, $p = .000$ (effect size = $-.83$).

In contrast, students in the experimental condition wrote essentially the same number of rebuttals as students in the comparison condition $F(1, 142) = 0.23$, $MSe = .126$, $p = .629$ (effect size = $-.09$), with comparable development of rebuttals $F(1, 158) = .34$, $MSe = .235$, $p = .56$ (effect size = $-.09$). The total number of documents cited by students at pretest was comparable in both groups $F(1, 142) = 0.36$, $MSe = 3.10$, $p = .55$, (effect size = $-.11$) as was the highest level of document use $F(1, 142) = 2.96$, $MSe = 1.12$, $p = .09$, (effect size = $-.26$).

Two groups did not differ on the WIAT or other demographic indicators before instruction began, however these results favor students in the comparison group for half of the dependent measures. Thus, we used students' pretest scores as a covariate for determining the relative effectiveness of the intervention for relevant posttest measures, and when ANCOVA was an appropriate procedure (i.e., essay length and the number of claims per 100 words). In each of these cases, the ANCOVA assumption that the regression slopes were homogeneous was met. To estimate the practical significance of effects, we computed effect sizes by dividing the difference between the means (or means that were regressed and adjusted for covariates) by the standard deviation of the comparison group.

We used ordinal regression to model results for all ranked measures, regardless of whether groups may have differed before instruction began. This procedure was applied for the following measures: quality (however scores were first recalculated from a low of 1 to a high of 4, due to minimal occurrences at the extreme values), degree of claim, degree of rebuttal, and document use. In completing the ordinal regression procedure, one first checks for outliers by examining descriptive results for the frequency distribution of each variable (this also helped us decide when to collapse one or more ordinal values). Next, we examined the test of parallel lines, which is a check of an assumption that slope coefficients are the same across all response categories. This assumption was met for the first three variables. We collapsed our fourth variable, document use, into two values, "no reference to a document" and "citation of a document and/or use of a quote," in order for the assumption for parallel lines to be met, thus limiting the generalizations we are able to infer from the results. Finally, in all cases the overall model was significant, meaning that we were able to reject the null hypothesis that a model without predictors was good as a model with predictors.

Finally, we used repeated-measures ANOVA to explore results for two measures at posttest (the number of rebuttals in students' papers and the number of documents students cited). This was because both variables are scale variables, and there was no pre-instruction finding favoring the comparison condition. This

procedure is more stringent than a posttest ANOVA and less likely to produce spurious results.

3.2. Essay length

Results of the ANCOVA applied to the posttest scores demonstrated a statistically significant main effect for group, $F(1, 157) = 23.52$, $MSe = 10132.10$, $p = .000$ (effect size = $.66$). The adjusted mean scores indicate that students in the experimental group wrote longer posttest essays than students in the comparison group.

3.3. Quality

Results of the ordinal regression for quality, with scores re-ordered into a low ranking of 1 (instead of 0) and a high ranking of 4 (instead of 5, no student received a higher score) are presented in terms of the probability of events occurring; in this case, estimating the likelihood that higher scores are associated with a student's assignment in the experimental group where instruction was delivered. Table 6 displays probabilities associated with each rating and group membership. The results show that students are twice as likely to earn the highest rated score for quality (4) in the experimental group. In contrast, students are twice as likely to earn the lowest score (1) in the comparison group. Students are also more likely to earn better scores (3) in the experimental group and less proficient scores (2) in the comparison group. When the top two scores for each group are combined, there is a 57% chance of earning a 3 or 4 for students in the experimental group as opposed to a 39% chance of earning the same score for students in the comparison group.

3.4. Claims

Results of the ANCOVA applied to the posttest scores demonstrated that groups were not significantly different at posttest, when the number of claims produced per essay were controlled for the length of each essay, $F(1, 158) = .48$, $MSe = .75$, $p = .488$ (effect size = $-.10$). The adjusted mean scores indicate that students in both the experimental and comparison group wrote essentially the same number of claims per 100 words in their posttest papers. We view this finding positively as students in the comparison group wrote essays with more claims at pretest.

Results of the ordinal regression for the development of students' claims, with values ranging from 1 to 4 (see Table 4 for an explanation) are also presented in terms of the probability of events occurring. Table 7 displays probabilities associated with each rating and group membership. These results show that it is relatively rare to receive the highest rating for development of one's claim, but that students who were instructed in the experimental group were three times more likely to do so. They were also more likely to show no claim development when in the comparison group (recall that a level 1 in this case represents no claims in the students' paper). Students were nearly equally likely to show a moderate development of claims, with a 77% probability of earning a 3 for students in the experimental group and a 62% probability of earning the same score for students in the comparison group.

Table 6
Estimated probabilities for quality.

Group		Estimated probability			
		1.00	2.00	3.00	4.00
Experimental	Mean	.10	.33	.36	.21
	Std. deviation	.07	.11	.07	.13
Comparison	Mean	.19	.42	.28	.11
	Std. deviation	.13	.07	.10	.07

Table 7
Estimated probabilities for degree of Claim.

Group		Estimated probability			
		1.00	2.00	3.00	4.00
Experimental	Mean	.02	.08	.77	.12
	Std. deviation	.02	.06	.01	.08
Comparison	Mean	.11	.23	.62	.05
	Std. deviation	.09	.09	.14	.09

Taking the two highest levels together, students in the experimental group had an 89% chance of demonstrating elaborated claims as opposed to a 67% chance in the comparison group.

3.5. Rebuttals

A 2 (instructional group) \times 2 (trials) repeated-measures ANOVA design was used to evaluate the relationship between the instructional conditions (experimental and comparison) and the number of rebuttals in students' essays to determine whether scores differed significantly at posttest. There was a time by group interaction $F(1, 138) = 6.94$, $MSe = .94$, $p = .009$, (effect size = .79). Table 5 presents descriptive information. Examination of this data reveals that students in the comparison group wrote more rebuttals on pretest essays. On the posttest measure, however, the majority of students in the experimental group wrote essays with more rebuttals.

Results of the ordinal regression for the development of students' rebuttals, with values ranging from 1 to 4 (see Table 4 for an explanation) are also presented in terms of the probability of events occurring. Table 8 displays probabilities associated with each rating and group membership. These results show that the probability of writing a paper without a developed rebuttal was at least 44%. However, this was probability increased to 78% for students in the comparison condition. The chance of writing papers with developed rebuttals was higher for students in the experimental condition, with students being three times more likely to write the most developed rebuttals in the experimental condition. In all, only 22% of the students in the comparison group wrote essays that included opposing claims, with or without simple counter claims or elaboration.

3.6. Document use

A 2 (instructional group) \times 2 (trials) repeated-measures ANOVA design was used to evaluate the relationship between the instructional conditions (experimental treatment and comparison) and the number of documents students cited in their essays to determine whether scores differed significantly at posttest. Table 5 presents descriptive information. There was a main effect for trials $F(1, 135) = 71.14$, $MSe = 9.39$, $p = .000$, and a main effect for group $F(1, 135) = 17.35$, $Mse = 11.53$, $p = .000$. More importantly, there was a time by group interaction $F(1, 135) = 26.60$, $Mse = 9.39$, $p = .000$ (effect size = 1.42). While groups were not significantly

Table 8
Estimated probabilities for degree of rebuttal.

Group		Estimated probability			
		1.00	2.00	3.00	4.00
Experimental	Mean	.44	.12	.26	.19
	Std. deviation	.06	.01	.01	.05
Comparison	Mean	.78	.07	.10	.05
	Std. deviation	.06	.01	.03	.02

different at pretest, on the posttest measure, the majority of students in the experimental group wrote essays with more document citations and use of quotations to further their arguments.

Results of the ordinal regression for students' use of documents, two values (1 and 2) are analyzed in terms of the probability of events occurring. Table 9 displays probabilities associated with each rating and group membership. This measure refers to "no reference to a document" and "citation of a document and/or use of a quote." We reduced the original four-level ordinal variable to a simplified view of document use in order for the assumption for parallel lines in the analysis to be met. Thus, while losing some level of sophistication in knowing how students' use documents, we can assess the probability of whether students use them at all (or a high versus low use) in their essays. These results show that the probability of referring to a document and/or using a quote in his or her essay was 83% for students in the experimental condition, whereas the chance of this happening was just over the level of chance, for students in the comparison condition. Thus, we attribute students' use of documents to the instruction they received that is under study in the present investigation.

4. Discussion

This study examined the effects of integrating disciplinary reading and writing strategies on poor and average high school writers' argumentative essays. The results demonstrated here are consistent with those reported by Yeh (1998) and De La Paz and Graham (2002), who developed interventions that taught students to write argumentative essays in English classes in which instruction focused on text structure, use of reasons and claims, and presenting counterarguments. We contribute to this body of research first by verifying a student's advantage of having participated in our intervention with respect to writing elaborated claims and rebuttals (the probability was about 1/3 higher for writing elaborated claims and three times higher for writing elaborated rebuttals for students in the experimental group), despite an initial advantage in development of claims and overall writing quality for students in the comparison group at pretest. Moreover, our analysis of students' writing specifies a more nuanced distinction of these elements within Toulmin's model, which is warranted for developing writers who are beginning to substantiate claims (for example, by trying to establish the credibility, if not the grounds, of the claim that they present) or disprove opposing-side claims in rebuttals (e.g., by providing a simple counter claim in the absence of direct refutation).

Thus, after instruction, early writing researchers might have inferred that students' improvements were merely due to their increased knowledge of text structure and the writing task, which Hayes (1996) viewed as influences from the task environment. However, in our study, students' writing has disciplinary meanings – their argumentative essays revealed the kind of thinking (albeit on a simpler level) that one would expect of historians. To illustrate, students' writing demonstrated that they understood relationships between series of events that they had read about in the primary and secondary sources (c.f., Shanahan & Shanahan,

Table 9
Estimated probabilities for degree of document use.

Group		Estimated probability	
		1.00	2.00
Experimental	Mean	.17	.83
	Std. deviation	.01	.01
Comparison	Mean	.46	.54
	Std. deviation	.01	.01

2008). As such, our results extend De La Paz's (2005) study in which two general education teachers taught middle school students to read and write argumentative essays on controversial historical topics. In the current study, students in the experimental group wrote essays with more document citations and used quotations to further their arguments at posttest, despite students in the comparison group having a relative advantage in this regard at pretest. Unfortunately, a limitation in the stability of our data precludes further analysis of our data regarding students' use of documents.

4.1. Writing behavior

As expected, the instruction had a positive effect on the writing performance of the participating high school students. After instruction, students in the experimental group were much more likely to be able to use cite documents or quotes, or use quotes to further an argument after instruction. This was not a subtle change in growth among students in the experimental group, as students in the comparison group had outperformed them at pretest on the number of documents cited and the level of document use. To illustrate, consider this example, *"That if one raises any questions or expresses any criticism of the policies of our country in the field of foreign policy, one's very patriotism is subject to question." What Senator Wayne Morse is trying to tell us is that we let our pride, ideals, and dreams cloud our thinking. That we cannot accept the fact that another country has a different point of view toward their lives. We let our ego get the best of us and that is how the great War started.*" Comments like these, found in greater numbers and levels of sophistication in the posttest essays of the experimental, illustrate the ways in which writing can be enriched by the meaningful integration of documentary evidence. Effect sizes for writing measures that were evaluated on a scale (length and number of documents used) were moderately large (.38 and .59, respectively) on the final writing probe.

Students in the experimental group also wrote argumentative essays with more advanced development of claims and rebuttals, after instruction, after controlling for the length of their essays. We had not predicted how the intervention would affect the number of claims and rebuttals students generated, and found that these results were less clear. Whereas students in the comparison group wrote essays with more claims at pretest, students in the experimental group wrote essays with essentially the same number of claims at posttest. In contrast, whereas students in both conditions wrote essays with essentially the same number of rebuttals at pretest, students in the experimental group wrote essays with more rebuttals at posttest. It may be argued that students in the experimental group were encouraged to include rebuttals in their essays while students in the comparison group were not. However, we note the presence of rebuttals in essays written by both groups of students at the posttest. What is more interesting to note is the level of claim and rebuttal development among students in the experimental group.

The posttest materials in Appendix B show one of the better examples of an essay written by a student in the experimental group (41% of the students wrote essays where more than half of the claims in their essays were well developed, this was nearly twice the 22% found in the comparison group). The example highlighted here shows a passage in which a student not only explains, but also substantiates a claim with other claims, evidence and text. This excerpt contains not only a well-elaborated argument (*The Truman Doctrine is to contain the threat of communism; Minh and his Viet Minh were clearly looking to take over S. Vietnam and Laos. That is 2 more communist countries we would have to deal with. Not to mention if we were to go to war with the Soviet Union*) but also one that is hierarchically organized with multiple sub-

claims being used to support a broader claim (*The Resolution would let the president control the spread of communism without starting a war. We have found N. Vietnamese shell fragments on decks of our destroyers. This is a clear act of communist threat to S. Vietnam. This happened on Aug. 2, and on Aug. 4 it was reported another attack occurred. We need to contain them or we will add communism to the growing party*). Finally, the essay contains a rebuttal in which the student advances an opposing-side claim (*Maybe the United States is too much involved in foreign policing*) and a rebuttal (*If we are not, then we become Isolationist and other countries suffer like in WWII*). This posttest argument represents more elaborated argumentation than seen at the pretest (Appendix B), and demonstrates how the student has moved from a one-sided argument that simply states claims, to a two-sided argument that organizes historical evidence to elaborate claims. The addition of a level 4 rebuttal at the posttest also suggests that the student has developed a global representation of the arguments advanced in the documents and sees how contradictory claims must be addressed. To accomplish this level of argumentation, the student must not only possess an argument structure for writing the essay, but also a means of sorting through historical claims and evidence to represent divergent historical perspectives. While level 4 rebuttals occurred in about 20% of the experimental students papers, only 5% of the students in the comparison group wrote rebuttals at this level of quality. The intervention did more than merely prompt students to provide rebuttals, a simple prompt to include a response to a counter argument would not prepare students to produce rebuttals of this sophistication. This finding gives credence to the belief that the intervention was responsible for improvement in use of rebuttals in the experimental group; however the possibility that the intervention prompted some students in the experimental group to include rebuttals at posttest cannot be ruled out entirely. Future research is recommended to confirm our results.

4.1.1. Limitations

An unexpected finding was that students in the comparison condition received lower scores on their posttest measures of quality and on the development of their claims than they did at pretest. The most plausible reason for this finding appears to be based on the limitation in the study, noted by the American history professor and high school history teacher that the posttest materials were slightly harder than those presented at pretest. We believe that students in the experimental group were better equipped to deal with these more difficult materials, after learning the historical reasoning and writing strategies, and their performance was not negatively impacted. In contrast, students in the comparison group had engaged in group discussions that emphasized understanding of specific historical content rather than strategic processes that could be transferred to new learning situations (i.e., different source materials). Hence, their performance suffered when asked to read more difficult materials, and to respond in writing to an historical essay prompt at posttest.

Moreover, in any quasi-experimental study it remains possible that uncontrolled factors contributed to the results. To illustrate, although we have no evidence to suggest learning experiences were significantly different between conditions, we did not formally assess students' knowledge about the posttest topic before administering the final probe. Thus, a competing hypothesis that cannot be completely ruled out is that students in the experimental group were more knowledgeable about the posttest topic. However, we do not think this is likely, for two reasons. First, our fidelity data provides some assurance that teachers in both conditions used parallel materials for topics that required the use of multiple perspectives, and that students in both conditions read similar multiple sources and wrote arguments on a series of controversial topics over time. In addition, while teachers used the

sources differently, when reading and writing with them in the experimental and control classrooms, students in each condition had access to essentially the same content, for approximately the same amount of time, across one academic semester. Moreover, when they were not using the materials in the study for instruction, teachers in both conditions carefully followed state-mandated content standards.

Our second reason for believing the students in the experimental group were not more knowledgeable than students in the comparison group stems from the fact that there was one comparison and one experimental teacher at each school, and the schools were in different districts. It would have been more likely we would have found knowledge effects that were attributable by school rather than by condition given our design.

There are inherent limitations to the strength of conclusions drawn from research conducted with non-random samples. However, conducting intervention research in public schools is expensive, and under increasing scrutiny from a variety of stakeholders. Another feature of this methodology is that it masks within group differences, in both the experimental and comparison conditions. Issues of non-responders, aptitude-treatment-interactions, and so forth were not explored in this study, nor did we attempt to explore how students who were learning English or students who were in need of special education services responded to the intervention.

4.2. Application of cognitive strategy instruction in the regular classroom

The current study extends the work of De La Paz (2005) and provides additional verification that strategy instruction can be applied successfully with typically developing adolescents in general education social studies classrooms to teach students to use evidence from what they read and transform it to build global, evidence-based arguments. Students in the experimental condition learned two strategies, reasoning with primary and secondary source documents, as well as planning an argumentative essay involving historical facts and claims, from one social studies teacher over the course of an academic semester. Lessons were integrated with other content over the course of the semester, at a rate of about one lesson per 2 weeks of “regular” instruction in 20th century historical content. Students learned how to consider several aspects of the sources they were given to read, and to corroborate and contextualize aspects across sources with events of the time period in which they were situated. They also learned how to use evidence from these sources as a means for substantiating their claims in their written arguments. The current design did not allow us to parse out whether the success of the intervention was due to improvements in planning, use of evidence, or in the development of students’ claims and rebuttals, but from our theoretical viewpoint, we believe a combination of these factors was most likely.

4.3. Educational implications

The results of this study show that with explicit instruction, teachers can shape new understandings for what students expect to write and how they perform in history classrooms. When given clear expectations regarding what it means to engage in disciplinary literacy activities, repeated exposure to document-based questions coupled with direct instruction in historical thinking processes, text structure for writing historical essays, as well as a systematic teaching process that transfers responsibility for learning from teachers to students, then low to average high school writers can achieve demonstrably high levels of writing proficiency, as compared to peers who do not receive this form

of instruction. Throughout the unit, students shared knowledge, read texts and outlined arguments, and in so doing, they used historical reasoning to accomplish authentic, purposeful and integrated tasks. Our results suggest that students developed sophisticated task representations for writing because they experienced firsthand how reading and writing strategies converge to accomplish clearly defined goals in historical writing. In this way, the inquiry process provided focus and made the purpose of reading, pre-writing and writing strategies transparent to students (Kress, 1993; Roth, 1998). We believe that scaffolding historical reasoning enhances writing because students read documents with the purpose of identifying and contrasting conflicting viewpoints. The work of disciplinary thinking about the documents allowed them to develop more advanced and integrated claims and rebuttals, and it lead them to cite sources more readily and more appropriately.

Certainly, if given additional time beyond the intervention shown here, one can expect even greater evidence of historical thinking in students’ writing. The federal government’s Teaching American History grants have enhanced many teachers’ content knowledge and resulted in the development of pedagogical frameworks that are intended to be integrated into a year-long curriculum (Mandell, 2008); certainly, one continuing challenge is to validate additional ways that students can demonstrate deep disciplinary understandings in other types of writing assignments and for other pedagogical purposes.

Acknowledgement

The authors wish to thank AERA for funding that permitted this study as well as Robert Senkewicz, historian, and the teachers and students from San Jose and Cupertino Union School Districts for their help on the project.

Appendix A

A.1. Scoring rubric for historical opinion essays

To receive full credit, the essay will need to:

- Accurately interpret the documents and incorporates outside information related to the documents.
- Provide a clear opinion on the topic.
- Support a position with accurate facts, examples and details.
- Weigh the importance, reliability and validity of the evidence.
- Analyze conflicting perspectives presented in the documents.
- Weave the documents into the body of the essay.
- Include a strong introduction and conclusion.
- The essay is persuasive.

Reduce credit if the response:

- Does not recognize the reliability, validity, or perspectives of the documents.
- Reiterates the content of the documents with little or no use of outside information.
- Discusses the documents in a descriptive rather than analytic manner.
- Lacks an introduction or conclusion.
- Includes inconsistencies in claims or reasons, and irrelevant information.

A.1.1. Scoring rubric

6 Exceeds expectations.

- The paper displays a thorough understanding of the topic and related issues.

- Reasons indicate connection of facts that goes beyond what is presented in the readings.
- The writer deals with an opposing opinion with refutation or alternate solutions. Refutation explicitly recognizes opposing views and provides one or two reasons against those arguments. An alternate solution proposes a compromise position or alternative way of addressing the arguments of the opposition.
- Presents a strong introduction and conclusion.
- Well structured, well written; proper spelling, grammar and mechanics.

5 Strong.

- Paper states a clear opinion and gives a context for that opinion.
- Shows an ability to analyze, compare, and contrast issues or events.
- All historical information is accurate, and clearly relates to the question.
- Paper is free from inconsistencies and irrelevancies that would weaken the argument.
- Shows an adequate interpretation of content across sources.
- The essay is generally well organized and gives an introduction and conclusion.
- Clearly written and coherent; some minor errors in writing.

4 Competent, developed.

- Paper states an opinion and gives reason(s) to support the opinion, plus some elaboration of at least one reason OR
- May include one reason that is well developed using information that could be convincing.
- Uses most documents correctly; recognizes that all evidence is not equally valid OR
- Attempts to analyze issues and events.
- Most historical information is accurate, and generally relates to the question.
- At most 1–2 errors in writing standard English detract from the essay's meaning.

3 Emerging.

- Paper states an opinion and gives reason(s) to support the opinion, plus some elaboration of at least one reason OR
- May give three or more reasons with no elaboration.
- Shows basic, though simplistic, understanding of the topic and related issues.
- Weaker organization; some errors in writing detract from essay's meaning.
- Has a vague or missing introduction and/or conclusion.

2 Low, minimally developed.

- Paper states an opinion, and gives some support, but the reasons are not explained.
- The reasons may be of limited plausibility.
- Shows little understanding of the topic and related issues.
- Poorly organized; many errors in standard English.

0 Undeveloped.

- Paper states an opinion, but no reasons are given as support OR
- Reasons are unrelated to or inconsistent with the opinion or they may be incoherent OR
- Facts do not relate to the documents (i.e., relates to personal knowledge on the topic) OR
- Strings random facts together in a weak narrative that lacks focus.

- Details are weak or nonexistent.
- Disorganized; littered with errors in standard English.

0 Not rated.

- Completely ignores the question OR
- Paper responds to the topic in some way, but does not provide an opinion on the issue.
- Includes so many indecipherable words that no sense can be made of the response.
- Ignores or misuses the documents.
- Lacks any organization; little attempt made; blank paper.

Appendix B

B.1. Participant A: pretest essay

To grow as in expand or to make bigger is what you'd would want if you had a company, correct? Your company would have a better business, more money. Things and times would be good. Who wouldn't want that? I would bet that we all want that even if it meant breaking some laws. That wouldn't be right, but our own self-greed for money would drive use to do so. That is exactly what the United States did. I for one am for that expansion of the United States. I am all for it because trade, naval purposes and Manifest Destiny.

If the United States could expand into island nations *we would control some trade routes* [P2.1 (level 2)]. In order for us to be able to trade easier was to have *docking or refueling stations* [P2.3 (level 2)] at certain islands. We also thought that we can *get raw materials in order for us to trade. For example, we got involved with Cuban and Spanish War because we had an idea that if we help the Cubans defeat the Spanish we could get raw materials from them* [P2.2 (level 3)].

Meanwhile, our naval fleet was only 12 in the world. We had to make them stronger to take these island nations. When we take them *our Pacific fleet would be tremendously stronger. We could have posts in Hawaii, Cuba and the Philippines. Not to mention our navy would grow and we would become a world power* [P3.1 (level 3)].

Finally, the idea of *Manifest Destiny was still in our hearts* [P1.2 (level 2)]. We could not stop expanding. It was something that we didn't want to do. You can't just stop expanding, after you have done it for the past 80+ years.

In conclusion I am for expansion. I think it would be the best thing for the United States to do if they want to become a world power.

B.2. Participant A: posttest essay

In 1954 the Geneva Accords divided Vietnam on the 17th Parallel. The North was controlled by Ho Chi Minh and the South was controlled by Dien Diem (an anti-communist). In 1963 Diem was assassinated and fear of communist takeover in the South was rising. American naval forces patrolled the Gulf of Tonkin. There were "deliberate attacks against US naval vessels." This led to the Gulf of Tonkin Resolution. I am extremely in favor of the resolution for two reasons, the Truman Doctrine and we are sending aid to the South to prevent attack.

The Truman Doctrine is to contain the threat of communism. Minh and his Viet Minh were clearly looking to take over S. Vietnam and Laos. That is 2 more communist countries we would have to deal with. Not to mention if we were to go to war with the Soviet Union. The Resolution would let the president control the spread of communism without starting a war. We have found N. Vietnamese shell fragments on decks of our destroyers. This is a clear act of communist threat to S. Vietnam. This happened on Aug. 2, and on Aug. 4 it was reported

another attack occurred [P1.1 (level 4)]. We need to contain them or we will add communism to the growing party.

The Geneva Accord was to divide the North and South along the 17th parallel and to establish separate countries. Since the assassination of Diem it has been nothing but chaos. The United States is already sending supplies to the South so why not give the president [the power] to send an attack if he had to? We have sent lots of money and our naval forces are being shot at as if an act of war. We need to give the President power if we wish to stop the gun fire at our ships and planes [P2.1 (level 3)].

Maybe the resolution isn't such a great idea. Look, if the president sends troops to Vietnam, then it could cause a communist/anti-communist war. We can sometimes "force every issue into the context of freedom and communism." Maybe the United States is too much involved in foreign policing [C3.3 (level 3)]. But if we are not, then we become Isolationist and other countries suffer like in WWII [P1.3 (level 3); (Rebuttal level 4)]. We left Europe alone and what happened? Hitler came to power. The spread of communism is growing rapidly. We must contain it.

In conclusion, in order for us not to have another Hitler, we must stick to the Truman Doctrine and contain communism. We must not allow Minh to get out of control. The power must go to the President because if it does not then our government, the congress, will take too long to take action. Also, they all might not agree. Give the power to the President. We need the Gulf of Tonkin Resolution.

References

- Ackerman, J. M. (1991). Reading, writing, and knowing: The role of disciplinary knowledge in comprehension and composing. *Research in the Teaching of English*, 25, 133–178.
- Applebee, A., & Langer, J. (2006). *The state of writing instruction in America's schools: What existing data tell us*. Albany, NY: Center on English Learning and Achievement.
- Barton, K. C. (2005). Primary sources in history: Breaking through the myths. *Phi Delta Kappan*, 86, 745–753.
- Beaufort, A. (2004). Developmental gains of a history major: A case for building a theory of disciplinary writing expertise. *Research in the Teaching of English*, 39, 136–185.
- Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Bower, B., Lobdell, J., & Owens, S. (2004). *Bring learning alive! The TCI approach for middle and high school social studies*. Palo Alto, CA: Teachers' Curriculum Institute.
- Britt, M. A., & Aglinskis, C. (2002). Improving students ability to identify and use source information. *Cognition and Instruction*, 20, 485–522.
- Britt, M. A., Rouet, J.-F., Georgi, M. C., & Perfetti, C. A. (1994). Learning from history texts: From causal analysis to argument models. In G. Leinhardt, I. L. Beck, & C. Stainton (Eds.), *Teaching and learning in history* (pp. 47–84). Hillsdale, NJ: Lawrence Erlbaum Associates.
- De La Paz, S. (2005). Reasoning instruction and writing strategy mastery in culturally and academically diverse middle school classrooms. *Journal of Experimental Psychology*, 91, 310–311.
- De La Paz, S., & Graham, S. (1997). Effects of dictation and advanced planning instruction on the composing of students with writing and learning problems. *Journal of Educational Psychology*, 89, 203–222.
- De La Paz, S., & Graham, S. (2002). Explicitly teaching strategies, skills, and knowledge: Writing instruction in middle school classrooms. *Journal of Educational Psychology*, 94, 687–698.
- Deshler, D. D., & Schumaker, J. B. (1986). Learning Strategies: An instructional alternative for low-achieving adolescents. *Exceptional Children*, 52, 583–590.
- Englert, C., Raphael, T., Anderson, L., Anthony, H., Stevens, D., & Fear, K. (1991). Making writing strategies and self-talk visible: Cognitive strategy instruction in writing in regular and special education classrooms. *American Educational Research Journal*, 28, 337–373.
- Felton, M., & Kuhn, D. (2002). The development of argumentative discourse skills. *Discourse Processes*, 29(2&3), 135–153.
- Gee, J. P. (1992). *The social mind: Language, ideology and social practice*. New York: Bergin and Garvey.
- Geisler, C. (1994). *Academic literacy and the nature of expertise: Reading, writing, and knowing in academic philosophy*. Hillsdale, NJ: Lawrence Erlbaum.
- Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 187–207). New York: Guilford.
- Graham, S., Harris, K. R., & Mason, L. (2005). Improving the writing performance, knowledge, and self-efficacy of struggling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology*, 30, 207–241.
- Graham, S., & Perin, D. (2007a). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools – A report to Carnegie Corporation of New York*. Washington, DC: Alliance for Excellent Education.
- Graham, S., & Perin, D. (2007b). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99, 445–476.
- Grant, S. G. (2003). *History lessons: Teaching, learning and testing US high school classrooms*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Harris, K. R., & Graham, S. (1996). *Making the writing process work: Strategies for composition and self-regulation*. Cambridge, MA: Brookline.
- Hayes, John R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 1–27). Mahwah, NJ: Erlbaum.
- Hayes, John R. (2006). New directions in writing theory. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 28–40). New York: Guilford Press.
- Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Hillsdale, NJ: Erlbaum.
- Karras, R. W. (1994). Writing essays that make historical arguments. *OAH Magazine of History*, 8, 54–57.
- Kress, G. (1993). Genre as social process. In B. Cope & M. Kalantzis (Eds.), *The powers of literacy: A genre approach to teaching writing* (pp. 22–37). Pittsburgh, PA: University of Pittsburgh Press.
- Kuhn, D. (1991). *The skills of argument*. Cambridge, MA: Cambridge University Press.
- Kuhn, D., Shaw, V., & Felton, M. (1997). Effects of dyadic interaction on argumentative reasoning. *Cognition and Instruction*, 15(3), 287–315.
- Kuhn, D., Weinstock, M., & Flaton, R. (1994). Historical thinking as theory-evidence coordination. In M. Carretero & J. F. Voss (Eds.), *Cognitive and instructional processes in history and the social sciences* (pp. 377–402). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Magrath, C. P., & Ackerman, A. (2003). *The neglected "R": The need for a writing revolution. The National Commission on Writing*. New York, NY: College Entrance Examination Board.
- Mandell, N. (2008). Thinking like a historian: A framework for teaching and learning. *OAH Magazine of History*, 22(2), 55–63.
- Nystrand, M., Gamoran, A., & Carbonaro, W. (1998). *Towards an ecology of learning: The case of classroom discourse and its effect on writing in high school English and social studies*. Albany, NY: National Research Center on English Learning and Achievement (ERIC Document ED 415525).
- Nystrand, M., & Graff, N. (2001). Report in argument's clothing: An ecological perspective on writing instruction in a seventh-grade classroom. *The Elementary School Journal*, 101(4), 479.
- Perfetti, C. A., Britt, M. A., & Georgi, M. C. (1995). *Text-base learning and reasoning: Studies in history*. Hillsdale, NJ: Lawrence Erlbaum.
- Roth, W. M. (1998). *Designing communities*. Boston: Kluwer.
- Rothschild, E. (2000). The impact of document-based questions on the teaching of United States history. *The History Teacher*, 33, 495–500.
- Salahu-Din, D., Persky, H., & Miller, J. (2008). The Nation's Report Card: Writing 2007 (NCES 2008-468). National Center for Education Statistics, Institute of Education Sciences, US Department of Education, Washington, DC.
- Schumaker, J. B., & Deshler, D. D. (1992). Validation of learning strategy interventions for students with LD: Results of a programmatic research effort. In B. Y. L. Wong (Ed.), *Contemporary intervention research in learning disabilities: An international perspective* (pp. 22–46). New York: Springer-Verlag.
- Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review*, 78(1), 40–59.
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge, England: The University Press.
- Wineburg, S. (1991). On the reading of historical texts: Notes on the breach between school and the academy. *American Educational Research Journal*, 28, 495–519.
- Wineburg, S. (2001). *Historical thinking and Other unnatural acts: Charting the future of teaching the past*. Philadelphia, PA: Temple University Press.
- Wong, B. Y. L. (1997). Research on genre-specific strategies in enhancing writing in adolescents with learning disabilities. *Learning Disability Quarterly*, 20, 140–159.
- Wong, B. Y. L., Butler, D. L., Ficzer, S. A., & Kuperis, S. (1997). Teaching adolescents with learning disabilities and low achievers to plan, write, and revise compare-and-contrast essays. *Learning Disabilities Research and Practice*, 12, 2–15.
- Yeh, S. (1998). Empowering education: Teaching argumentative writing to cultural minority middle-school students. *Research in the Teaching of English*, 33, 49–84.
- Young, K. M., & Leinhardt, G. (1998). Writing from primary documents: A way of knowing in history. *Written Communication*, 15, 25–68.