Promoting written historical reasoning among undergraduate L2 students

Sendur, K.A.

Publication date
2021

CHAPTER 5

HISTORICAL CONTEXTUALIZATION IN STUDENTS' WRITING

In this study, we investigated the historical reasoning of undergraduate L2 students as measured in their argumentative document-based writing. The study focused specifically on students' performance in written historical contextualization before and after participating in a historical reasoning course. The Content and Language Integrated Learning course was designed using a cognitive apprenticeship model and was based on principles likely to facilitate students' written historical reasoning. Conducted as a quasi-experimental study, students in an experimental condition received explicit instruction in historical contextualization and other features of historical reasoning, while those in the control group participated in a version of the course without a focus on historical contextualization. Students in both the experimental and control groups significantly improved in all of the areas of historical reasoning that we measured. There was not a significant difference between the groups in the area of historical contextualization, but a further qualitative analysis demonstrated traces of the instructional approach in students' writing. Unexpectedly, students in the experimental group were significantly better than the control group in terms of writing claims. Possible explanations for this finding are discussed.

Sendur, K. A., van Drie, J., & van Boxtel, C. (submitted). Historical contextualization in students' writing. Manuscript submitted for publication.

INTRODUCTION

Historical reasoning is an important goal of history education. When reasoning historically, a student "organizes information about the past in order to describe, compare, and/or explain historical phenomena" (Van Drie & Van Boxtel, 2008, p. 89). Because of its importance, instruction in historical reasoning is of great interest to the history education community. Studies have investigated different methods of instruction, such as cognitive apprenticeship and direct instruction (De La Paz et al., 2017; Monte-Sano, 2008), as well as the use of heuristics and different ways of presenting information (Nokes et al., 2007). With instruction, students can learn to demonstrate components of historical reasoning, for example, make written claims and support them with evidence (De La Paz et al., 2017), use and evaluate the reliability of sources (Sendur et al., 2021) (Britt & Aglinskas, 2002; Reisman, 2012a), and corroborate with multiple sources (Nokes et al., 2007). Many of these same studies, however, also demonstrate that learning to reason in history is difficult for students and that not all students perform well. Studies focusing on historical reasoning among L2 learners, the focus of this study, are not common.

Historical contextualization is one particular component of historical reasoning that remains difficult for students in history classes (Monte-Sano, 2010; Reisman, 2012a; Van Drie et al., 2015). Historical contextualization is the reconstruction of the chronology, geography, and social features of the time period in order to situate a source or historical phenomenon (Van Drie & Van Boxtel, 2008; Wineburg, 1991). We also found this difficulty in a previous study (Sendur, et al., 2020). One reason for such difficulty may be that many students view the past through a present lens and therefore may experience difficulty in interpreting history from the context of the past (Grant, 2018). It may also stem from the belief of the past as inherently deficient and change in history as inherently progressive (Lee, 2005). Alternatively, students may lack (well organized) knowledge about historical developments, phenomena, and chronology (Van Boxtel & Van Drie, 2012). Not enough is known, however, about how students contextualize because of the lack of recent research in this area (Reisman & McGrew, 2018). As a key concept in historical reasoning, historical contextualization warrants further research. In the current study, we focus on the content and procedural knowledge needed for historical contextualization when writing a historical argument. This work has the potential to provide insight into how to promote historical contextualization in writing.

THEORETICAL FRAMEWORK

HISTORICAL REASONING

Students' historical reasoning can be studied from the perspective of the framework proposed by van Drie and van Boxtel (2008, 2018). The framework consists of six interrelated aspects that conform to the notions of reasoning in the discipline of history that can be used as a way to study students' oral and written historical reasoning. First, in history students *answer historical questions* about continuity and change, causes and consequence as well as similarities and differences. Students may, for example, analyze the causes of the fall of the Roman Republic. Because questions in history can be considered ill-defined problems, answering such questions often involves *argumentation*, for example, by making claims. Students reason with information from *sources* when answering the historical question by using the sources as evidence in their argument. These arguments also make use of *substantive concepts*, such as patricians and the Roman Empire. When students reason they must also construct the *historical context* by considering the chronology, geography and social characteristics of the time period they are studying. Finally, in the construction of the argument, students also make use of *meta-concepts and related heuristics*, such as understanding what counts as historical evidence, corroboration and considering the usefulness and reliability of the source for the particular question they are answering.

HISTORICAL CONTEXTUALIZATION

Historical contextualization is one aspect of historical reasoning (Van Drie & Van Boxtel, 2008). Historians contextualize by reconstructing the chronological, geographical, and social characteristics of the source, person, event or phenomenon under study (Van Boxtel & Van Drie, 2012; Van Drie & Van Boxtel, 2008; Wineburg, 1991, 1998). Contextualization enables historians to develop an interpretation of a unique event or period that accounts for the general characteristics of the time period being studied (Carr, 1990). By contextualizing, a historian can help make sense of actions by those in the past that may appear counterintuitive to our modern understanding, but were rational to those in the past. For example, when explaining why free Romans volunteered to become gladiators despite the risk of death, it is necessary to understand the implications of living in a militaristic society with limits on those who can participate in the military.

In a study of two historians, Wineburg (1998) identified six aspects that were used to reconstruct the historical context when reading historical documents: 1) spatio-temporal, the chronology and geography of the event, 2) social-rhetorical, comments about the social demands of the event, 3) biographical comments about the life of the person being studied, 4) historiographic comments about other historians' writing, 5) linguistic, comments about the historical meaning of words, and 6) analogical references to different periods of history. One historian was able to use his specialized knowledge of the period to create a context, demonstrating the importance of deep content knowledge in historical interpretation. Even without extensive knowledge of the historical period, however, the other historian was able to use his knowledge of other historical periods and procedural knowledge of contextualization to interrogate the sources and create a context through which to interpret the period. This study demonstrates the importance of both content and procedural knowledge in the act of contextualization.

STUDENTS' PERFORMANCE IN HISTORICAL CONTEXTUALIZATION

Historical contextualization appears to be an aspect of historical reasoning that is challenging for students in the context of history classes. One reason that historical contextualization is difficult is because of a tendency towards presentism. Students commonly believe that people in the past shared the same beliefs and values as those in the present (Lee, 2011; Shemilt, 1984). Similarly, Reisman and Wineburg (2008) note that in many cases, students approach history with a background rooted in popular culture and preconceived notions of the past, leaving them ill-equipped to understand history. This tendency towards presentism is even found in those intending to become history teachers (Wineburg & Fournier, 1994). These beliefs may make it difficult for students to see the importance of historical contextualization.

Another potential reason for students' difficulty is that contextualization requires students to possess and also use a sufficient amount of well-organized historical background knowledge. In a study focusing on students' performance in historical perspective taking, Huijgen, van Boxtel, et al. (2017) used a fictional case study to determine how students decided why an historical actor would behave in a certain manner. They found that students who considered more background knowledge, particularly at least three different types, performed better in a measure of historical perspective taking. Similarly, students who were more successful in dating a primary source used better organized and more extensive historical background knowledge than students who were not as successful (Van Boxtel & Van Drie, 2012). When given training that focused on building students' network of background knowledge and chronological landmarks, a subsequent group of students outperformed those who were not given this type of training. These studies demonstrate that in order to perform contextualization successfully, students need to possess both sufficient and sufficiently well-organized background knowledge.

Written historical contextualization may also be challenging because of the complexity of choosing and effectively integrating relevant background information into a historical essay. When evaluating students' writing for contextualization, studies look for both the inclusion of accurate information about the historical period, such as chronology, location, cultural values, and historical events, as well as its use to situate, strengthen or explain the argument (Monte-Sano, 2010; Monte-Sano & De La Paz, 2012; Nokes et al., 2007; van Boxtel & van Drie, 2013; Van Drie et al., 2015). Such contextualization may take the form of causation by using the background to explain why a certain event or phenomenon may have taken place (Monte-Sano, 2010; Nokes et al., 2007). For example, when writing about Julius Caesar's decision to sponsor extravagant gladiator games as part of his campaign for political office, students should include both information about the role of a candidate's personal fortune in the Roman political system at the time, as well as explain why Julius Caesar's decision was logical. Students who fail to contextualize may either not include background information or their use of the background information may include factual or interpretive errors (Monte-Sano, 2010).

Studies of students' historical writing have shown that there is wide variation in their use of historical contextualization, and that many students do not include contextualization in their writing, possibly because of a lack of awareness of how and why to include it in their writing. One study reported that older and better writers more frequently demonstrated the use of contextualization than younger and poorer writers (De La Paz et al., 2012), whereas another study found that students were able to include elements of contextualization across different writing tasks (Monte-Sano & De La Paz, 2012). In Monte-Sano (2010)'s study, some students incorporated contextualization, but others constructed an inaccurate historical context or one with a flawed interpretation. In the L2 tertiary context, the inclusion of the "circumstances surrounding the historical actor's actions" as evidence led to better essays (Myskow & Ono, 2018, p. 64). In another study of historical reasoning and L2 students' writing, contextualization was seldom used (Sendur et al., 2020). Nokes et al. (2007) found that high school students so rarely used contextualization in their essays that it could not be analyzed as a part of their intervention on the use of heuristics as a learning tool. Other studies (e.g., De La Paz et al., 2014; Van Drie et al., 2015) also mention the use or promotion of contextualization in writing, but do not separately report students' proficiency, making their performance in contextualization unclear. This wide variety in performance, and particularly the lack of contextualization in much of student writing, makes studies that investigate contextualization and approaches towards promoting it in student writing critical.

TEACHING HISTORICAL CONTEXTUALIZATION IN WRITING

In this study we approach teaching written historical reasoning in general and written historical contextualization specifically through a Content and Language Integrated Learning (CLIL) model. CLIL is a pedagogical model that simultaneously teaches language and content (Coyle et al., 2010). CLIL is widely used in the European context, including in history classes (Eurydice, 2006). Extensive research in CLIL has found that it has language-related benefits for students without significantly affecting the acquisition of content knowledge (Pérez-Cañado, 2012). In one case, students in a CLIL history context required additional class time to make content gains similar to their peers (Dallinger et al., 2016), but in another case they outperformed their non-CLIL peers in terms of content knowledge (Oattes et al., 2020).

In both teaching and research contexts, the document-based question (DBQ) is commonly used when assessing students' historical writing (McCarthy Young & Leinhardt, 1998; Monte-Sano, 2010). In this study we focus on students' written historical contextualization in an argumentative DBQ writing task. In a DBQ, a student composes an answer to a historical question through the analysis of multiple sources, often both primary and secondary. DBQ essays, particularly those that are argumentative in nature (Greene, 1994; Monte-Sano & De La Paz, 2012), provide the conditions for students to demonstrate written historical reasoning, including contextualization. Nokes and De La Paz (2018) conclude that such writing may be one of the best ways to assess students' historical reasoning.

Studies have investigated how to best promote students' historical writing using approaches such as text models (Van Drie et al., 2015) and class discussion (van Boxtel & van Drie, 2013). Explicit instruction and cognitive apprenticeship have been shown to be particularly effective in helping students compose essays with features of historical reasoning (De La Paz et al., 2014; De La Paz et al., 2017; Monte-Sano, 2011; Nokes et al., 2007). Explicit instruction is effective in an L2 context as well (Goo et al., 2015; Norris & Ortega, 2000). Studies focusing on contextualization as a part of written historical reasoning, however, are not common. The few studies that include explicit instruction in historical contextualization and writing have had mixed results, with the students in Nokes et al's (2007) study failing to include contextualization after instruction, whereas student in De La Paz et al's (2017) study improved in their written historical reasoning (contextualization was included as a part of the measure). Given the difficulty that students have in incorporating contextualization in their historical writing, studies that integrate both the active use of background knowledge and instruction in procedural aspects of how to include contextualization into students' writing are important.

Huijgen, van de Grift, et al. (2017, p. 163) propose four strategies when teaching historical contextualization: "(1) Reconstructing the historical context, (2) fostering historical empathy, (3) performing historical contextualization to explain the past, and (4) raising awareness of present-oriented perspectives when examining the past." Using these strategies, the authors (2018) designed a historical contextualization intervention study using a combination of a case study, the reconstruction of the historical context, and a historical empathy task. After eight lessons using this approach, students in the intervention group made significant progress in historical contextualization in comparison to a control group. Reisman and Wineburg (2008, p. 203) advocate three strategies to help students learn to contextualize when reading historical sources: "(1) providing background knowledge, (2) asking guiding questions, and (3) explicitly modeling contextualized thinking." In a later study (Reisman, 2012a, 2012b), however, students' contextualization did not improve even with explicit strategy instruction in which the teacher modelled contextualization and students completed graphic organizers that included contextualization questions. This finding led the author to speculate that the complexity of the skill or the lack of practice may have played a role. With students' difficulty in historical contextualization, particularly in writing, and the lack of consistent findings regarding a pedagogical approach, it is important to test the effectiveness of different pedagogical approaches that both utilize and depart from the proposed strategies.

DESIGN APPROACH

In this quasi-experimental intervention study, we adopt aspects of teaching historical contextualization from Reisman and Wineburg (2008) and Huijgen, van de Grift, et al. (2017). We use a cognitive apprenticeship model since it has been shown to be effective in history education (De La Paz et al., 2014; De La Paz et al., 2017). We also integrate writing instruction through the use of text models and focused language practice in order to help students better grasp the genre and language of argumentative writing in history. We investigate the effect of a historical reasoning curriculum based on these principles on students' historical reasoning and writing when taught with an increased emphasis on incorporating historical contextualization into DBQ writing as compared to a control group that does not include a historical contextualization focus. We also conduct a separate qualitative analysis of students' historical contextualization by investigating the different ways in which contextualization is incorporated in their writing.

RESEARCH QUESTIONS

1. How do undergraduate students perform on aspects of historical reasoning (claim, evidence, sourcing, corroboration and historical contextualization) in their document-based writing before and after participating in a course with explicit instruction in historical reasoning?
2. What is the effect of explicit instruction in historical contextualization during a historical reasoning course on undergraduate students' document-based writing?

For our first research question, prior to the course, we would expect that students in this course would perform poorly in historical contextualization as measured in a written source-based argument. We would expect that students in both groups would improve on the different elements of their historical reasoning as measured by their document-based writing. We would expect similar improvements for both groups in the areas of claim, supporting a claim with evidence, sourcing and corroboration. For the second question, we would expect that students in the experimental condition would score higher on historical contextualization than those in the control group.

METHOD

PARTICIPANTS AND CONTEXT

This study was conducted at a small private English-medium university in Istanbul, Turkey during the 2017 fall semester. Undergraduate students typically spend one to two semesters studying English in a pre-university intensive English preparation program before beginning their undergraduate studies. Close to half of these students choose a major in engineering. While all undergraduate students take a series of required history courses, the university does not offer an undergraduate degree in history.

140 students in the intensive English preparation program participated in this study while enrolled in a historical reasoning course taught as a part of the program. See Table 1 for an overview of the students. All students are non-native English speakers who had been placed at the B2 level according to the Common European Framework of Reference for Languages (CEFR), which is the highest level of English taught in the program. English language instructors in the program must meet minimum education and experience requirements, and teach using a highly standardized curriculum.

Table 1

Student Demographics and Intended Area of Study

Experimental	Control
60 students	80 students
34 male	41 male
26 female	39 female
Intended program of study	Intended program of study
41 Engineering and natural sciences	65 Engineering and natural sciences
7 Business	10 Business
12 Arts and social sciences	5 Arts and social sciences

Eleven English language instructors teaching the historical reasoning course in thirteen intact classes took part in this study. Instructors had taught the historical reasoning course between zero and six semesters prior to the study (Experimental instructors had zero to three semesters experience; Control instructors had zero to six semesters of experience). Four instructors with a variety of experience were invited at the recommendation of the course coordinator and accepted to teach in the experimental condition. The remaining instructors were assigned to the control condition. On the basis of their instructor, five of the thirteen classes (n= 60 students) were assigned to the experimental condition and eight classes (n= 80 students) were assigned to the control condition (one instructor in each condition taught two classes each).

All instructors were aware that different conditions existed, but only instructors in the experimental condition were aware of the focus of the intervention. Experimental lesson plans were kept in a location inaccessible to control condition instructors, and experimental condition instructors were specifically told not to discuss their lesson plans with anyone teaching in the control condition.

A historical reasoning course prepared by the first author for the English preparation program was used for the study. Detailed lesson plans and answer keys were prepared for each condition. The lessons specified goals, activities and the pacing of each lesson. In the event that lesson components ran longer than expected, activities central to the study were highlighted as required activities for a given lesson. The course coordinator kept track of the progress of each class and ensured that all required lesson components were completed.

All instructors participated in a training with the first author that outlined the curriculum, focusing on aspects of the curriculum that applied to both conditions. Instructors in the experimental condition participated in a separate training with the first author regarding aspects of the curriculum specific to the experimental condition. All instructors also participated in weekly meetings with the course coordinator to clarify questions regarding the content or lesson plans, as well as to review answer keys and sample essays, as needed.

HISTORICAL REASONING COURSE

Students in both conditions completed similar versions of a 28-hour historical reasoning course that took place for four hours weekly for seven weeks. We first explain the course aspects that were common to both conditions, and then itemize differences between the two conditions. The course used the topic of gladiators in late Republican Rome and the early Empire to introduce students to aspects of historical reasoning and the language of historical argumentation. This topic was chosen partially because it would be studied the following semester in one of the required history courses. Students were made aware of this link between courses so that it would be motivating for students to learn about the topic. Gladiators was also chosen because it was the subject of a recent popular television show.

After an introduction to historical reasoning and Roman gladiators in weeks one and two, respectively, the following four weeks were divided into three units: Roman socioeconomics, Roman politics in the late Republic and early

Empire, and Roman cultural values. The final week consisted of a comprehensive course review.

In terms of historical reasoning, all student received explicit instruction in sourcing and corroboration using a cognitive apprenticeship approach (Collins et al., 1991). Instructors first modelled the concepts. Instruction was then scaffolded so that students were given significant assistance, which was withdrawn as students demonstrated competence. Students were given multiple opportunities to practice each element in the course. For example, when students were first introduced to the concept of sourcing, they listened to a think aloud of the first author analyzing a primary source and took guided notes using a graphic organizer. In subsequent weeks, students worked in small groups and as a class to analyze other primary sources using the same type of graphic organizer. After practicing sourcing, students individually demonstrated their performance in a short assessment based on the Historical Assessment of Thinking (Wineburg et al., 2012).

Instruction in writing a document-based historical argument focused on the use of a claim and evidence to address a historical question (Monte-Sano, 2010). For the purposes of this intervention, student were taught to write an exposition, a one-sided argument (Coffin, 2006). Similar to sourcing and corroboration, students learned through a cognitive apprenticeship model. When students first learned about writing a claim, for example, they used a set of criteria introduced by the instructor to assess the quality of several sample claims, as well as write their own. Later in the course, they worked in small groups to develop a claim based on a small set of primary and secondary sources. Students independently wrote a DBQ essay after the conclusion of the following units using the associated primary and secondary sources: socioeconomics and politics.

Students were provided with sample DBQ answers as models of argumentative writing in history. As a guided practice in historical argumentation, students analyzed these text models. Text models have promise as a pedagogical method of demonstrating the features of a genre in both L1 and L2 student writing (Graham & Perin, 2007; Hillocks Jr, 1986; Hirvela, 2016; Hyland, 2007), including in history (Van Drie et al., 2015). Because students may place too much trust in the text models, Hirvela (2016) recommends examining models of both successful and unsuccessful texts. Therefore, when working with text models, students had two primary tasks: 1) identify features of an argument and historical reasoning, and 2) revise text models with missing or erroneous components. The first task focused primarily on passive recognition of textual features while the second was intended to help students practice text production. For example, in the first text model, students were asked the following: "The claim should account for all of the evidence. What major aspect did this essay forget to account for?"

Students in both conditions used the same set of readings, including primary sources, secondary sources and background information necessary for historical contextualization. Based on the recommendations of Reisman and Wineburg (2008), we included a timeline noting major events and figures studied in the course and sufficient background information about gladiators and Roman history to provide "some familiarity with key developments" (pg. 203). Primary and secondary sources were simplified by course instructors to match students' B2 reading level. Primary sources were excerpted and simplified based on principles by Wineburg and Martin (2009). Secondary sources were chosen or modified to limit the amount of argumentation to encourage students to develop their own arguments rather than using others' arguments (Miller et al., 2016). A one-page summary of highly relevant historical background information was included as the first page of a lesson when a new topic was introduced. The text was written in a neutral textbook style, each subtopic was given a separate heading, and keywords were bolded.

The course was structured as a Content and Language Integrated Learning (CLIL) model integrating both content and language (Coyle et al., 2010).

We integrated language into the course in terms of a focus on both meaning and form, as well as frequent opportunities to engage in written and oral production (de Graaff et al., 2007; Westhoff, 2004). To help students cope with reading comprehension, students read modified sources, annotated their readings, and completed graphic organizers for each primary source. Frequent oral and written output was built into the course in order to build fluency. The course also included an overt focus on language forms useful in historical argumentation, such as modals and citation language (Coffin, 2006; De Oliveira, 2011). Additionally, all students received focused language practice with language models for sourcing, corroboration, and explaining the significance of evidence. For example, when introducing corroboration, students were given the language stem, "[Evidence from Author 1]. A similar point is made by [Author 2], who…" Students used the language stems first to write about points of corroboration they had identified between two texts in that lesson, and later in their DBQ writing.

EXPERIMENTAL CONDITION

Students in the experimental condition received explicit instruction in historical contextualization used in historical reasoning and writing. The instruction focused on two aspects outlined in the literature: 1) engaging students with the background knowledge they could use to contextualize (Huijgen, van de Grift, et al., 2017; Huijgen et al., 2018; Reisman & Wineburg, 2008) and 2) supporting their procedural knowledge of writing so that they could incorporate contextualization into their arguments (Graham & Perin, 2007; Hillocks Jr, 1986; Hirvela, 2016; Hyland, 2007; Van Drie et al., 2015).

Engaging students with background knowledge

Two types of activities were used to help students engage with the background knowledge needed to contextualize: case studies and quote sorting. These activities supported students' contextualization by building the background knowledge students needed in order to contextualize and to helping students see the value of using historical context to ground their claims about the past.

Three discussion-oriented case studies were designed to give students a reason to engage with the historical background information in order to make a well-grounded claim. In a task similar to Huijgen et al. (2018), students completed a series of case studies of fictional historical actors corresponding to each of the three topics: socioeconomics, politics and cultural values. See Figure 1 for a sample case study. The use of both teacher and student-led discussions can help students make gains in reasoning, argumentation, content knowledge and writing (Kuhn et al., 1997; Reznitskaya et al., 2001; Van Drie & Van de Ven, 2017; Wegerif et al., 1999).

The activity took place in two parts. First, students answered a set of guiding questions about the provided historical background. The purpose of these questions was to act as a guided practice in creating a historical context and situating the character from the case study in the past. Second, students attempted to decide how the person would have acted in the situation given the historical context. In two of the three case studies, the historical background information could have supported multiple answers, requiring students to form claims well-supported by the historical context in order to argue their position. This second part was designed as an independent practice with the instructor facilitating (and fact-checking) a student-led discussion.

Figure 1

Case Study for the Unit Roman Politics in the Late Republic and Early Empire

Late Republican Politics Case Study

It is 65 BCE and Julius Caesar is planning to run for praetor in a couple of years. One of your friends is planning to run against him. You heard that Julius Caesar is planning to spend an extravagant amount of money to sponsor gladiator games in honor of his father, and also offer a public feast. Spending that amount of money will make voters forget your friend's name. There is no way he can win. What should he do?

1. Don't worry that Julius Caesar's games are more expensive than your friend's. He has better ideas for policy. Voters will know the better candidate.
2. Sponsor a new law banning so many gladiators in Rome. It's dangerous to have so many gladiators in the city.
3. Borrow money so your friends can offer even better games.

Guiding Questions

1. What would you, as a modern person, want to do?
2. Julius Caesar lived during the late Republic. Briefly describe the state of politics during this time.
3. How does sponsoring gladiator games help a person get elected?
4. Julius Caesar is running for praetor. Who does he need to convince to vote for him? Are his plans likely to be effective?
5. Does Julius Caesar belong to a political party? How important do plans for governing seem to be when deciding whom to vote for?
6. Why might people in Rome consider it dangerous for one person to own many gladiators?
7. What should the person running against Julius Caesar do? Why?

To help students further engage with the historical background knowledge, students in the experimental condition also completed one quote-sorting activity. During this activity, students sorted unattributed quotes about politics from primary sources into either the Roman Republic or Empire based on evidence in the quote. For example, in the following simplified excerpt from Dio Cassius, the student would need to be able to note that politicians were using gladiators as a personal army, and search the historical background to determine that this was indicative of the late Roman Republic.

> Milo (a politician) caused many disturbances, and at last he collected some gladiators and others who agreed with him and fought with Clodius (another politician), so that bloodshed occurred throughout practically the whole city (Loeb edition translation by E. Cary, 39.8).

This activity took place during a lesson on politics, and was specifically designed to help students consider the differences in the political systems during the two time periods because students in a previous version of this course sometimes confused the features of the two systems.

Supporting students' writing

Students engaged in two types of activities that supported integrating historical contextualization into their writing: text models and focused language practice. These activities supported students contextualization by teaching students the procedural knowledge needed in order to integrate contextualization into their writing, and the value of such an endeavor. When engaged in these activities, we emphasized the importance of including relevant background information with the goal of helping the reader understand the argument. We also focused on noting a connection between the background information and evidence as a part of the explanation.

Using a series of guiding questions, all students evaluated text models for argument structure and features of historical reasoning, including sourcing and corroboration, as described above. Students in the experimental condition also evaluated the text models for the use of historical contextualization. For example, in one text model students decided which information to include as historical contextualization and where to place it in the essay. See Figure 2 for an excerpt of another text model and guiding questions students used after writing their first DBQ. The original activity included three sample paragraphs and was designed as a guided practice to help students see the importance of historical contextualization in situating and strengthening the argument by comparing paragraphs with and without effective historical contextualization.

Figure 2

An Excerpt From a Text Model and Guiding Questions Activity Used with Students in the Experimental and Control Conditions

Experimental Condition Questions	Control Condition Questions
1. Compare the paragraphs to the argument structure requirements on page 24 in your course book. a. Find 2 examples of evidence that is relevant, specific and significant. b. Find 1 example that explains the importance of the evidence or links the evidence back to the claim c. Find 1 example of an evaluation of the usefulness of the source. 2. Identify the historical context (if any) in each of the paragraphs. Historical context may include information about: a. Time period b. Location c. Socioeconomic class d. Politics e. Culture and values 3. Evaluate the use of the historical context. Which paragraph has a) historical context that helps explain the argument, b) irrelevant historical context and c) no historical context?	1. Compare the paragraphs to the argument structure requirements on page 24 in your textpack. Evaluate to what extent each paragraph has the following: a. Evidence that is relevant, specific and significant. b. An explanation of the importance of the evidence c. A link between the evidence and the claim d. Corroboration between sources. e. An evaluation of the usefulness of the source
Common Text Model Paragraph 1 The upper classes benefitted from their social class in unimportant ways, such as seating at gladiator games. For example, Livy (34) details that senators were given separate seating for a gladiator show. Suetonius likewise describes Augustus' law that assigned seats based on social class (Aug. 44). In both instances, the seats assigned to the upper classes were better than those given to the lower classes.	

Focused language practice was used for all students to help them with the types of language structures used when writing historical arguments. Students in the experimental condition also practiced the language related to historical contextualization, such as language related to chronology. For example, in one activity students constructed a timeline of major events concerning gladiator games in the Roman Republic and Empire, and then used language models to describe their duration (i.e., for at least 500 years) and sequence (i.e., before the Empire began) in a guided practice.

Students also practiced with language models that demonstrated how to note the importance of or connection between the historical background information and the evidence. For example, as a part of the quote sorting activity described above, students explained their answer using relevant historical background information. They then noted the importance or connection between the quote and related historical background information to Roman politics using language stems such as "this demonstrates…" and "it is reasonable to conclude that…"

CONTROL CONDITION

Students in the control condition did not receive any explicit instruction related to the use of historical contextualization in historical reasoning or writing. Students in both conditions had an identical number of course hours, used the same primary and secondary sources, and completed the same assessments. All students had historical background information and language models for historical contextualization in their coursebook, but students in the control classes did not complete activities that engaged with those aspects.

Three of the seven lessons were identical with no differences between conditions. Four four-hour lessons, corresponding to the weeks on socioeconomics, politics, and cultural values, contained differences during part, but not all of the lesson. See Figure 3 for the timing of a partial sample lesson, which demonstrates how timing was balanced between the two conditions. In some cases, students in each condition used the same (or similar) materials for different purposes. In this case, the time for each activity was the same. This is the case in the text models, such as the one in Figure 2 and described above. In other cases, the experimental condition students had an additional activity, such as the case study shown in Figure 1 and reflected in Figure 3. These activities were also planned for a specific amount of time. To compensate for the amount of time spent in the case study, the control condition completed a more extended version of another task common to both groups.

Figure 3

Lesson Component Timing for a Partial Sample Lesson About Politics in the Late Roman Republic

Experimental Condition	Control Condition
35 minutes: Text model task (experimental condition questions) (See Figure 2) *15 minutes:* Review homework (common task) *15 minutes:* Case Study (experimental condition) (See Figure 1) *20 minutes:* Guided source evaluation *15 minutes:* Writing about source evaluation *5 minutes:* Wrap up discussion (common question)	*35 minutes:* Text model task (control condition questions) (See Figure 2) *15 minutes:* Review homework (common task) *30 minutes:* Guided source evaluation (control condition extended discussion component) *20 minutes:* Writing about source evaluation *5 minutes:* Wrap up discussion (common question)
Total Time: 105 minutes	*Total Time: 105 minutes*

DATA SOURCES AND ANALYSIS

Pre-test and post-test document-based question

In addition to the two DBQs noted above, students completed a pre-test and post-test DBQ. The pre-test DBQ was used to assess their initial historical reasoning and writing. The pre-test was completed during lesson 2 after students had been introduced to the topic of gladiators and the concept of historical reasoning and writing, but before beginning detailed instruction. After completing the course, students completed a post-test DBQ to assess their historical reasoning and writing. See Table 2 for an overview of the pre and post-test questions. Both tasks used a similar question structure that was designed to solicit a claim and an argumentative response. Students answered each question based on course content, which was designed to allow students the possibility of including each aspect of historical reasoning addressed in the course. For example, all primary sources included a head note including background information about the author that could be used in sourcing. Between the two assessments, students participated in the experimental or control versions of the historical reasoning course.

Table 2

Pre-test and Post-test Questions, Sources, and Word Count Requirements

	Pre-test	Post-test
Question	To what extent is "The Gladiator" film clip historically accurate?	It is believed that many gladiators were volunteers. To what extent would it be desirable and/or undesirable for a free man to volunteer to become a gladiator?
Word Count	120-150 words	250-300 words
Sources	1 Primary Source 4 Secondary Sources	5 Primary Sources 2 Secondary Sources
When completed	During Lesson 2	Following Lesson 7

Analysis of historical reasoning in writing

Student responses to the two writing tasks were scored using a five point analytical rubric from a previous study (Sendur et al., 2020). See Appendix F for the rubric. Building on the works of Monte-Sano and De La Paz (2012) and Nokes (2017) the rubric measured the following aspects of historical reasoning and writing taught in the course: claim, evidence, source evaluation, historical contextualization and corroboration. For each category, students were scored based on the highest level they achieved. Students scoring in the 3 to 4 point range for a given feature of historical reasoning demonstrate competence in historical reasoning, while those scoring a 2 show lower levels of performance. Scores in the 0 to 1 point ranges indicate a lack of historical reasoning or significant errors.

The pre-test was scored by a trained research assistant unfamiliar with the study and the first author. The research assistant was trained in three one-hour sessions during which the first author provided sample scored essays with written explanations for the scoring decisions. The research assistant practiced scoring subsequent sample essays under the guidance of the first author and later independently. After resolving scoring discrepancies, the first author and the research assistant scored a set of 20 student essays. Cohen's Kappa ranged from .64 to 1.0 with the claim and source evaluation categories receiving the lowest and highest Kappa, respectively. Cohen's Kappa was not calculated for the category historical contextualization because in the randomly chosen data set both authors had 100% agreement and all essays received the same score. The first author scored the remainder of the essays.

The first and second author coded a round of 21 students responses to the post-test. Cohen's Kappa ranged from .66 to .82 with the claim and source evaluation categories receiving the lowest and highest correlations, respectively. The first author scored the remainder of the essays.

Measure of historical knowledge

Students completed a short closed-book/notes assessment of their historical knowledge immediately after handing in the post-test. The measure assessed students' factual knowledge of the main concepts in Roman history that were available in the coursebook and could have been potentially useful in the post-test. Questions consisted of multiple choice, fill in the blank and ordering concepts. For example, students were asked "Which of these values and jobs describe the ideal Roman man? Circle 3" and "Which social class had the least legal rights?" Students answers' were scored out of 13, with one point for each correct answer.

Individual interest in history survey

Before beginning the historical reasoning course, students completed an 8-question survey to measure their individual interest in history. The survey, used by Stoel et al. (2017), is an adaptation of a survey to measure interest in mathematics (Linnenbrink-Garcia et al., 2010; Pintrich et al., 1993). The survey measures interest with a 6-point Likert scale from strongly disagree to strong agree. Sample items include "History is practical for me to know" and "I like history." Cronbach's alpha for the survey was .92 (n=128). Note that a small number of students had an incomplete dataset for this measure or the measurement of historical knowledge. The number of students is reflected in the results.

Exploring historical contextualization

We also prepared students' post-tests for a further analysis of their written historical contextualization. First, we counted each instance of historical contextualization, as defined by the rubric above. Using each of these incidences of historical contextualization (excluding geography and chronology), we next specified where each was located with respect to the relevant part of the argument. For example, consider a student includes information about the importance of militarism in Roman society to support the argument that people volunteered as gladiators to obtain military-like glory. In this case, we identified where the historical contextualization (the importance of militarism) was located with respect to the argument it supported (people volunteered as gladiators for military-like glory). Chronology and location were excluded from the analysis because students used them as brief notations, such as 'during the Republic' that provided little aid in understanding the argument.

Since students in the experimental group were taught to include historical contextualization in order to help the reader understand the argument, the location can serve as a proxy for the extent to which contextualization may aid in understanding. Contextualization in close proximity to the related argument is more likely to be useful in aiding understanding. Using an iterative process, we identified six different locations that students used to integrate contextualization, as described later in Table 5.

Finally, we determined if a connection between the contextualization and evidence was explicitly noted and/or a conclusion was drawn on the basis of the historical contextualization, and if so, what language was used. We examined this aspect since it was included as an aspect of instruction.

RESULTS

In this section, in response to the first research question, we report on the extent to which students include features of historical reasoning before and after the course. To address the second research question, we present results comparing the performance of students in the experimental and control groups for the feature, historical contextualization. We also compare students' individual interest in history. Finally, we provide a separate analysis of students' claims and historical contextualization to explore the nature of students' performance in the experimental and control conditions.

HISTORICAL REASONING IN STUDENTS' DBQ WRITING

We conducted Kolmogorov-Smirnov tests to see whether the distribution of the pre- and post-test scores deviated from a normal distribution. The pre- and post-test scores for all categories of the document-based writing rubric and the total scores were significantly non-normal (*D*(140) = 0.09 to 0.54, p < .05).

Because the scores were not normally distributed, we conducted a non-parametric Mann-Whitney test to test if the experimental group differed from the control group on the pre-test. The total score on the pre-test in the experimental group (M*dn* = 5.00) did not differ significantly from the total score on the pre-test in the control group (Mdn = 4.00), *U* = 1999.50, *z* = -1.71, *ns*, r = -.14. There were also no significant differences for the five subcategories of the rubric.

We first investigated how students performed before the course. As expected, students performed poorly in written historical contextualization in a DBQ prior to the course. See Table 3 for descriptive statistics for each category. To test our hypothesis that in both conditions students would improve their essay score, we conducted a Wilcoxon test. In the experimental condition, the total score for the post-test (M*dn* = 13.0) was significantly higher than the total score for the pre-test (M*dn* = 5.0), *T* = 1, p < .05, r = -.61 Also in the control condition, the total score for the post-test (M*dn* = 12.0) was significantly higher than the total score for the pre-test (M*dn* = 4.0), T = 1, p < .05, r = - .61. Thus, in both conditions there is a large positive change in the total essay score. Further analysis showed that in both conditions students scored significantly higher on the post-test compared to the pre-test with respect to all five aspects of the rubric (claim, evidence, source evaluation, historical contextualization and corroboration).

Table 3

Medians and Ranges for Pre and Post-test by Condition (Experimental n=60, Control n=80)

	Experimental		Control	
	Pre-test	Post-test	Pre-test	Post-test
	M*dn* (Range)	M*dn* (Range)	M*dn* (Range)	M*dn* (Range)
Claim	3.00 (0-4)	3.00 (0-4)	1.00 (0-4)	2.00 (0-4)
Evidence	2.00 (0-4)	2.00 (1-4)	2.00 (0-4)	2.50 (1-4)
Source Evaluation	0.00 (0-2)	3.00 (0-4)	0.00 (0-1)	3.00 (0-4)
Historical Contextualization	0.00 (0-2)	3.00 (0-4)	0.00 (0-3)	3.00 (0-4)
Corroboration	0.00 (0-4)	3.00 (0-4)	0.00 (0-4)	3.00 (0-4)
Total Score	5.00 (5-20)	13.00 (0-11)	4.00 (0-10)	12.00 (4-19)

Our second hypothesis was that students in the experimental condition with explicit instruction on historical contextualization would outperform students in the control condition on using historical contextualization in their post-test essay. To test this hypothesis, we conducted a Mann-Whitney test. The historical contextualization score in the experimental group (M*dn* = 3.00) did, however, not differ significantly from the score of the control group (M*dn* = 3.00), *U* = 2146.50, *z* = -1.11, *ns*, r = .09. We explored whether the experimental and the control condition differed on other categories of the rubric and on the total score for the essay. We only found a significant difference for the subcategory Claim. Students in the experimental condition (M*dn* = 3.00) scored significantly higher on this aspect than students in the control condition (M*dn* = 2.00), *U* = 1704.50, *z* = -3.08, p < 0.01, r = .26. This is a small to medium effect. Because we conducted tests for several dependent variables, we might increase the chance of making a Type 1 error, and we disregard the possible relations between the dependent variables. Although a MANOVA assumes normally distributed variables, we conducted a MANOVA to check whether the conditions differed along the combination of variables. Pillai's trace showed that the experimental and control group differed significantly with respect to the five subcategories of the rubric, *F*(5,134) -2.53, p < .05. Separate univariate ANOVAs on the five subcategories of the rubric only revealed a significant effect of the intervention on Claim (*F*(1, 138) = 8.88, p < .05).

To summarize, explicit instruction on historical contextualization during the historical reasoning course did not result in better contextualization in the essays, but had a significant effect on the quality of the claims students made. Students who received the explicit instruction on historical contextualization made better claims.

INDIVIDUAL INTEREST IN HISTORY

We used a survey of individual interest in history, as described in the methodology. We conducted Kolmogorov-Smirnov tests to see whether the distribution of the survey of individual interest in history scores deviated from a normal distribution. The scores were significantly non-normal (*D*(128) = 0.10, p < .05) so we conducted a non-parametric Mann-Whitney test to test if the experimental group differed from the control group in terms of individual interest in history. The score in the experimental group (M*dn* =3.81) did not differ significantly from the score in the control group (Mdn = 4.38), *U* = 1716.50, *z* = -1.502, *ns*. We can conclude students did not have different levels of interest in history.

ADDITIONAL QUALITATIVE ANALYSIS OF STUDENTS' CLAIMS

Since the significant difference between the two conditions for claim was unexpected, we explored the nature of the differences between the two conditions. In our rubric, students who wrote a claim that was both arguable and directly answered the question scored in the 3 to 4 range. For example, "Being an gladiator can be desirable or undesirable depending on which social class individual belongs to. For lower class it is highly desirable while for upper class it is not desirable at all (Student 76, post-test)." Those who wrote a statement that was not arguable and without a direct answer scored in the 1 to 2 range. For example, "There are some positive and negative effects of being a gladiator (Student 23, post-test)." In both bands (3/4 and 1/2), students who accounted for much or all of the evidence scored higher than those whose answer only considered a subset of the evidence. Few students omitted a claim (a score of 0).

In exploring students' claims, we noted whether the claim included a specific or vague controlling idea detailing how the student would develop the essay. For example, we distinguished between the specific controlling idea "Volunteering to become a gladiator is very desirable for a free man. Better living conditions, female adoration and fame make the life of a gladiator very desirable (Student 68, post-test)" versus the more vague controlling idea "It was very undesirable for a freeman to volunteer to become a gladiator. Because in the future there are many consequences waiting for them (Student 25, post-test)." Additionally, we noted any conditions placed on the claim, such as a time period or specific segment of society for which the claim held.

Within any given scoring band, students' answers in both conditions were very similar in terms of the presence or lack of a controlling idea or condition. However, approximately 75% of students in the intervention group formulated a claim as opposed to only 44% in the control group. The students in the control group were more likely to write a descriptive statement instead of a claim. This formulation appears to form the basis of the difference between the two conditions.

EXPLORING ALTERNATIVE EXPLANATIONS

In this section, we conduct an analysis that explores potential reasons for these unexpected findings. First, we use the recommendation by O'Neill (2012) to consider whether the design was a factor in these findings. Later, following the model of Barron (2003), we introduce hypotheses that could account for and explain these unexpected findings.

This intervention meets the five criteria O'Neill (2012) proposes to help explain and learn from intervention designs that return unexpected results. Based on this analysis, we conclude that the intervention had a high likelihood of success. O'Neill (2012)'s first criteria is that studies clearly state that they failed to achieve their goals, which we have presented earlier in the results section. We address the second criteria outlining the theoretical basis for why the study should have worked in our description of the course in which we note that we employ a cognitive apprenticeship model, a model which has been shown to be effective in teaching written historical reasoning, including to L2 students (De La Paz et al., 2017). We also use established literature in designing historical contextualization-focused instruction (Huijgen, van de Grift, et al., 2017; Huijgen et al., 2018; Reisman & Wineburg, 2008). Third, our study does not notably depart from the aforementioned literature.

Fourth, the intervention was well implemented, and therefore, had a strong likelihood of success. Our scripted lessons and support for instructors, as described in the methodology, have been effective in this class previously (Sendur et al., 2020), and result in high fidelity to the curriculum. Finally, we have provided significant detail for others to also try to explain why the intervention was unsuccessful. To meet this goal we have included details of our curricular approach in the methodology, including activities in the experimental and control conditions, and example of student work.

Based on the previous section, we conclude that our intervention had a good likelihood of success. Therefore, in this section we explore other reasons that could account for the nonsignificant findings. First, we explore whether there are differences between the two groups that could have accounted for the findings.

We first considered whether students' background knowledge about Roman history was sufficient since a sufficient level of background knowledge is necessary when contextualizing (Reisman & Wineburg, 2008). To test this, we collected a measure of knowledge in history as described in the methodology. We also considered whether the students' background knowledge differed by condition since both background and procedural knowledge are necessary in order to integrate historical contextualization into writing. If students in the control condition have greater background knowledge, it could have compensated for their lack of procedural knowledge. Therefore, our first hypothesis is as follows:

Alternative Hypothesis 1A: Students in both conditions have insufficient background knowledge needed to contextualize.

Alternative Hypothesis 1B: Students in the control group had greater levels of background knowledge needed for contextualization.

We conducted Kolmogorov-Smirnov tests to see whether the distribution of the measure of historical knowledge scores deviated from a normal distribution. The scores for individual questions and the total scores were significantly non-normal (*D*(138) = 0.46 to 0.84, p < .05). Because the scores were not normally distributed, we conducted a non-parametric Mann-Whitney test to test if the experimental group differed from the control group on the measure of historical knowledge. The total score in the experimental group (M*dn* =11.00) did not differ significantly from the total score on the post-test in the control group (Mdn = 11.75), *U* = 2167.50, *z* = -.752, *ns*. There were also no significant differences for the individual questions.

Students in both groups scored relatively highly on the measure and the questions were designed to test knowledge potentially useful in the post-test. Therefore, we conclude that students had sufficient knowledge to contextualize. Furthermore, students in both groups had similar knowledge of Roman history that they could use to contextualize in the post-test. We thus reject Alternative Hypothesis 1A/B.

Since students had sufficient background knowledge, we next considered whether the way we measured historical contextualization was a consideration in our nonsignificant findings or whether the intervention was ineffective in promoting written historical contextualization. The original analytical rubric builds on the works of Monte-Sano and De La Paz (2012) and Nokes (2017). Based on the results of these previous studies, we would have expected the rubric to be sensitive enough to measure written historical contextualization. It is possible, however, that an aspect is missing from the rubric that did not allow us to capture how students integrated historical contextualization into their writing in this study. We also would have expected the study design to be successful given its use of established literature and effective implantation, however, it may have been ineffective for these students.

In this series of hypotheses we explore aspects of our instruction and whether students' in both groups performed similarly. If students performed similarly across all aspects, it is likely that the intervention was ineffective in promoting written historical contextualization. However, if students perform differently, then it is also possible that the analytical rubric is not sensitive enough to capture students contextualization and should be updated based on the findings of this analysis.

First, in our intervention we emphasized the importance of including relevant historical contextualization with the goal of helping the reader understand the argument. Therefore, we would expect different amounts of contextualization in the groups if the rubric is a factor.

Alternative Hypothesis 2: Students in the control and experimental group used different amounts of historical contextualization in their essay.

To test this hypothesis, we counted each instance of historical contextualization, as described in the methodology. See Table 4 for an overview of total instances of historical contextualization. We conducted Kolmogorov-Smirnov tests to see whether the distribution of the instances of historical contextualization deviated from a normal distribution. The distribution was significantly non-normal (*D*(140) = 0.25, p < .05) so we conducted a non-parametric Mann-Whitney test to test if the experimental group differed from the control group. The number of instances in the experimental group (M*dn* =1.00) did not differ significantly from the number in the control group (M*dn* = 1.00), *U* = 2350.00, *z* = -.222, *ns*. We concluded that students in both groups used a similar number of instances of contextualization in their writing, and reject Alternative Hypothesis 2.

Table 4

Total Instances of Historical Contextualization by Condition (Experimental n=60, Control n=80)

Type of historical contextualization	Experimental	Control
Socioeconomic, political, cultural	56	69
Chronological and geographical	23	35
Total contextualization	*79*	*104*

Since students in both groups included similar amounts of contextualization in their writing, we continued to investigate whether other aspects of their contextualization were the same. For our next hypothesis, we examined where students incorporated historical contextualization within the context of their essay. This is important procedural knowledge for students since the placement of the contextualization can be used to help clarify its relevance to the argument, as we illustrate below. For this analysis we identified the location of each instance of historical contextualization, excluding geographical and chronological information, as described in the methodology and below in Table 5.

Table 5

Frequencies and Percentages of the Different Locations of Historical Contextualization (HC) Within Student Post-tests (Experimental n=60, Control n=80)

HC Integration	Experimental	Control
Introduction: HC is included in the beginning of the essay or in the claim, and before beginning the argument.	7 (13%)	6 (9%)
Beginning: HC is integrated at the beginning of the argument, and the related argument follows immediately after the HC.	20 (36%)	26 (38%)
Integrated: HC is integrated into an argument, with elements of the same argument both before and after the HC.	20 (36%)	24 (35%)
End: The HC is the final part of the argument.	8 (14%)	4 (6%)
Conclusion: The HC is at the end of the essay and not used as a part of the preceding argument.	0 (0%)	2 (3%)
Offset: The HC may be relevant to an argument, but is not located adjacent to the pertinent argument.	1 (2%)	7 (10%)

Instruction in the experimental condition focused on including relevant historical contextualization in a manner likely to help the reader understand the argument. When analyzing text models, for example, students discussed possible locations within the model for locating historical contextualization. If the rubric is a factor, then we would expect to see differences in the patterns of contextualization between the groups.

Alternative Hypothesis 3: Students in the control and experimental group used historical contextualization in different places in their essay that was not captured by the original rubric.

In examining the location of students' historical contextualization, we found that when it was placed at the beginning of an argument, it was most likely to receive the highest score. Integrating the contextualization into the argument was also effective, but more evenly spread among score bands. About one third of students in both conditions integrated contextualization in the beginning or middle of the argument. For example, in the following excerpt, Student 6 noted Roman militaristic values (italicized below) and their role in motivating people to become gladiators in the context of an argument for why free men might want to volunteer to become gladiators:

> *Rome had a militaristic society that believed virtus and gloria. Especially man would to show virtus in battle, then they gained gloria by winning victory of battle.* Therefore, free man want to become a gladiator for virtus and gloria. Dunkle states that life of free gladiators took on new meaning. They fought for its courage and achievement. Therefore, they had a honor as well or Roman soldiers (post-test).

While seen less frequently, historical contextualization in the introduction of the essay sometimes received a high score, especially when it was used to limit the claim, or provide background information that could situate a later argument. For example, one student used historical contextualization to explain the significance of the social status of freeman, the subject of his essay.

> Back in the Ancient Roman society it was desirable for freeman to volunteer to become a gladiator. Since social hierarchy is an important part of Roman society, freedman were slaves who had either bought their freedom or been granted it, their right were limited. Freedman may grow his wealth, but most of them were poor (Student 19, post-test).

Contextualization offset from the corresponding argument was less effective in students' writing, because it was more difficult to understand the relationship between the contextualization and the argument. Offset contextualization is counter to the instruction that students in the experimental group received. 10% of the control group contextualization was offset, while only 2% in the experimental group was offset. In the following example, Student 135 commented on the importance of militarism. However, while accurate, the contextualization was not related to the argument about reasons for becoming a gladiator in which it was situated:

> Some free man in Rome volunteered to become a gladiator and treated as a slave, even though they had citizen rights, because it was desirable for them. Rome was a militaristic society. Men were supposed to show his manliness and courage, especially in battle. According to Dunkle (2002) in ancient Rome, number of jobs was limited as a result for some becoming a gladiator seemed positive alternative (post-test).

Based on this analysis, it is likely that the placement of the contextualization matters, and there do appear to be some differences in students' responses on the basis of the condition. This qualitative analysis does not statistically test this hypothesis, and it may be a future area of research.

For our final hypothesis, we examined whether students noted a connection between the historical contextualization and the evidence or drew a conclusion pertinent to their argument. We examine this aspect because, as a part of the experimental condition, students learned to note a connection or conclusion as a part of the explanation and should therefore be present in their writing if the intervention had an effect.

Alternative Hypothesis 4: Students in the control and experimental group noted a connection between the historical contextualization and/or drew a pertinent conclusion to a different extent that was not captured by the original rubric.

For this analysis, we counted whether and how students noted a connection or drew a conclusion, as described in the methodology. See Table 6 for a list of words and phrases students used to signal a connection or conclusion. For example, one student (Student 55, post-test) used the signal phrase 'because of this' to drawn an explicit connection between the social status of gladiators and the implications on their lives: "At these times, different groups had different legal rights. Slaves were accepted as they were under property. Because of this, gladiators had no right to control their bodies even when they are beaten, wounded or killed." Of all instances of historical contextualization, 61% of those in the experimental condition included an explicit connection or conclusion, but only 43% of those in the control condition did so. This also appears to show there may be a difference between conditions.

Table 6

Signal Words and Phrases Used to Indicate Significance of Historical Contextualization

As a result
Because (of)
For this reason
From this perspective
It is reasonable to conclude
Like…also…
Since
So
So from this information
So it can be said that
So…because of
That is why
That leads to the fact
The reason (that is) why
Therefore
This (statement) shows that
Thus
Which shows/demonstrates/means
This is (might be) one of the reasons
Other: significance explained in a sentence and without signal words

PROPOSING A NEW HISTORICAL CONTEXTUALIZATION RUBRIC

The original analytical rubric building on the works of Monte-Sano and De La Paz (2012) and Nokes (2017) could have been expected to be sensitive enough to measure written historical contextualization. This qualitative analysis, however, leads us to conclude that two further components should be included. First, we propose that the location of the contextualization should be proximate to the argument, particularly at the beginning of the argument. Monte-Sano and De La Paz (2012, p. 286) note that one option for proficient contextualization is "integrates context and evidence in an explanation or conclusion." While this can be reasonably understood as requiring proximity, we believe that proximity should be explicitly required since we found students with offset contextualization in our dataset. Second, we propose that an explicit connection between the historical contextualization and the evidence, or a conclusion should be required. This draws on the importance that Monte-Sano (2010) places on causality in contextualization and complements the category used in Monte-Sano and De La Paz (2012, p. 286). We therefore propose the following revisions to the rubric, as outlined in Figure 4.

Figure 4

Proposed Historical Contextualization Rubric

Score	Historical Contextualization (Original)	Historical Contextualization (Proposed)
4	Provides accurate and relevant historical context (temporal, spatial or social features) as support for the claim, evidence or source. The HC is elaborate and used to situate and/or further the claim **or** the HC is less elaborate & explicitly used to situate and/or further the claim.	Historical context (temporal, spatial or social features) is accurate and relevant. The HC is elaborate enough to support and/or situate the argument. HC is proximate to the related argument. A connection/conclusion between the HC and the evidence is explicitly noted.
3	Provides accurate and relevant historical context. It may be used to implicitly situate **&/or** further the argument.	HC is accurate and relevant. The HC is elaborate enough to support and/or situate the argument. HC is proximate to the related argument.
2	Provides historical context that is of limited support for the argument **&/or** has minor inaccuracies. It is not used to situate **&/or** further the argument/argument **&/or** there are errors.	HC may have minor inaccuracies and/or be a limited relevance and/or is not elaborate enough to support and/or situate the argument. HC is proximate to the related argument.
1	Provides historical context that is historically inaccurate **&/or** largely irrelevant.	HC is historically inaccurate and/or largely irrelevant and/or the location of the HC is offset from the related argument
0	Does not note historical context.	Does not note historical context.

DISCUSSION AND CONCLUSIONS

In this study, we investigated the written historical reasoning of students in an historical reasoning course taught using a cognitive apprenticeship model. We compared the performance of students who received explicit instruction in historical contextualization with the performance of students who did not. At the beginning of the course, students in both groups had similarly low levels of performance in written historical reasoning in a DBQ. In line with other studies (Nokes et al., 2007; Reisman, 2012a), historical contextualization was one of the lowest scoring components of historical reasoning. After completing the course, we found that students in both groups made significant progress in their written historical reasoning in all areas that we studied, which confirmed our first hypothesis. This increase in performance is similar to that found in an earlier version of this course (Sendur et al., 2020) and supports studies that have found that cognitive apprenticeship is an effective model for teaching historical argumentation and writing (De La Paz et al., 2014; De La Paz et al., 2017).

Contrary to our second hypothesis, students in the experimental group did not perform significantly better in historical contextualization than those in the control group despite receiving explicit instruction. All students, including those in both the experimental and control groups, significantly improved their historical contextualization. Historical contextualization, however, was among the low scoring aspects of written historical reasoning in both conditions before the intervention. This is in line with the finding by others (Nokes et al., 2007; Reisman, 2012a). This finding highlights the difficulty of historical contextualization for students, and the importance of testing alternatives for teaching it.

We identify three potential explanations for the lack of a significant finding in terms of historical contextualization: 1) the extent of the contextualization instruction, 2) the nature of the case study, and 3) the manner of assessment. First, the duration of the instruction may be one possible explanation for the lack of a significant different between the experimental and control conditions. In our intervention, a part of each of four four-hour lessons was devoted to historical contextualization, split between engaging students with the background information and supporting students' procedural knowledge necessary for writing. Huijgen et al. (2018), on the other hand, were able to devote eight full lessons to instruction. Students may need significantly more practice individual aspects of historical reasoning before they are able to contextualize better. However, shorter interventions focusing on different aspects of historical reasoning using an explicit instruction approach have also shown a positive effect (Britt & Aglinskas, 2002; Van Drie et al., 2015) demonstrating that this intervention was likely of a sufficient length.

Second, the case studies focused on engaging with the historical background information and did not include a writing component. Instruction regarding incorporating historical contextualization into writing was limited to text models and focused language practice, and not explicitly linked to the case studies. It is possible that students were unable to transfer the skills from the case study to their own writing. As writing assignments were focused on a historical question and not a fictional character, it might also have been difficult for students to see the relationship between the case and the writing assignment. Third, historical contextualization was assessed through students' DBQ writing, and centered on the construction of the historical context and the use of historical background information to explain the past. It is possible that other assessments, such as students' case study answers or assessments that do not conflate writing and historical contextualization may show different results. It is also possible that the historical contextualization category of our analytical rubric was not sensitive enough to detect the differences in students' writing. The newly proposed rubric should be evaluated in a further study to determine whether it is more sensitive in measuring written historical contextualization.

Surprisingly, students in the experimental group performed significantly better when writing claims than those in the control group. Specifically, they wrote claims that were more arguable and directly addressed the writing prompt, which conforms to the instruction students received. In contrast, students in the control group tended to write non-arguable restatements of the topic. This difference is particularly surprising since students received identical instruction for writing claims. It is possible that some aspects of the historical contextualization instruction present in the experimental group helped students better make use of the existing claim writing instruction.

Oral dialogue has been shown to help student make gains in reasoning, argumentation, content knowledge and writing (Kuhn et al., 1997; Reznitskaya et al., 2001; Van Drie & Van de Ven, 2017; Wegerif et al., 1999) while explicit instruction is also an effective pedagogy for reasoning and argumentative writing in history (De La Paz et al., 2017; Stoel et al., 2017). We propose that the dialogic nature of the case study combined with an explicit focus on argumentation may have enabled students to better grasp the point of argumentation and develop procedural knowledge needed to make claims more indicative of argumentation, a position supported by Reznitskaya and Gregory (2013).

In the case studies, students engaged in dialogue with their peers and instructor to decide how they believed a historical figure would have acted in a given situation. In a guided practice, students used a series of guiding questions that both grounded them in the historical context and opened up multiple possible interpretations. As a part of an independent practice student-led discussion, the instructor's role was to facilitate the discussion and hold students to high standards of argumentation. These case studies demonstrated that multiple conclusions could be supported by evidence (while other conclusions were unsupportable) and necessitated argumentation to reach a conclusion. The reciprocal roles of proposing and evaluating each others' claims during the independent practice may have enabled students to develop the procedural knowledge, language and content knowledge necessary to formulate a claim indicative of argumentation (Reznitskaya & Gregory, 2013). Further research that compares students' dialogue and writing, as well as investigates the combination of oral dialogue and cognitive apprenticeship is warranted to investigate this possibility.

This study contributes to our understanding of how students incorporate historical contextualization within the context of an argument and the language they use to denote a connection or draw a conclusion, an area of interest to those studying L2 writing in history (Myskow & Ono, 2018). In our qualitative analysis, we identified six different locations within an argument where students included contextualization, and the variety of language that students used to indicate a connection or conclusion. We found that students in the experimental group more frequently explicitly noted a connection or conclusion than those in the control group, and included fewer instances of contextualization offset from the argument. This analysis has practical implications for instruction, as including contextualization at certain points of the argument received higher scores than others. The wide range of language that students used to demonstrate a connection shows that students may be able to use their current knowledge of language to note this in their writing.

This study has several limits. First, while instructors in both conditions had a similarly wide variety of experience, due to the constraints of the course we were not able to randomly assign students and teachers to a condition. Second, due to course constraints, the pretest and posttest used the same style of question, but the number of available sources and required essay length differed. Further studies should consider using a counter balanced essay approach to minimize this effect, such as that used by De La Paz et al. (2014).

In conclusion, this study demonstrates that a cognitive apprenticeship model works well with L2 undergraduate students learning about historical reasoning. Further studies should investigate other ways of assessing students' use of historical contextualization and further explore methods of teaching historical contextualization.