Historical contextualization in students' writing

Sendur, K.A.; van Drie, J.; van Boxtel, C.

DOI
10.1080/10508406.2021.1939029
Publication date
2021
Document Version
Final published version
Published in
Journal of the Learning Sciences

JOURNAL OF THE LEARNING SCIENCES
2021, VOL. 30, NOS. 4–5, 797–836
https://doi.org/10.1080/10508406.2021.1939029

Routledge
Taylor & Francis Group

Historical contextualization in students' writing

Kristin A. Sendur[1,2], Jannet van Drie[1], and Carla van Boxtel[1]

[1]Research Institute of Child Development and Education, University of Amsterdam; [2]The Center for Individual and Academic Development, Sabanci University

ABSTRACT

Background: This study focused on undergraduate L2 students' performance in written historical reasoning, particularly 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.

Methods: 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' historical reasoning was measured based on their argumentative document-based writing.

Findings: 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.

Contributions: This study makes contributions in terms of operationalizing and measuring written historical contextualization, particularly among L2 undergraduate students.

ARTICLE HISTORY
Received 23 March 2020
Revised 10 May 2021
Accepted 13 May 2021

Historical reasoning is an important goal of history education. When reasoning historically, a student "organizes information about the past in order to

CONTACT Kristin A. Sendur kristin.sendur@gmail.com Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, Netherlands.

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 (Britt & Aglinskas, 2002; Reisman, 2012b; Sendur et al., 2021), 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, 2012b; Van Drie et al., 2015). Historical contextualization has been defined as 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., 2021). 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 Boxtel and Van Drie (2018) and Van Drie and Van Boxtel (2008). 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 Boxtel & Van Drie, 2018; 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 counter-intuitive 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.

Written historical contextualization

Studies primarily define historical contextualization in terms of reading historical sources. Studies involving written historical contextualization define the term similarly in terms of the use of chronological, geographical, and social characteristics in order to create a historical context (De la Paz et al., 2017; Monte-Sano, 2010; Monte-Sano & De la Paz, 2012; Nokes et al., 2007; Van Drie et al., 2015). One challenge in the field of history education research is defining and operationalizing written historical contextualization. When it is operationalized to score written work, researchers similarly look for accurate information about the historical period (Monte-Sano, 2010; Monte-Sano & De la Paz, 2012; Nokes et al., 2007; Van Boxtel & Van Drie, 2013; Van Drie et al., 2015).

Some studies look beyond the inclusion of background information. In these studies, there is an implicit expectation of a causal link or a conclusion drawn based on the background information. Nokes et al. (2007) uses Wineburg's (1998) six aspects of contextualization in order to "understand how or why the event took place" (coding document). The coding examples predominantly include causal language (so, since) explicitly linking the background information to a conclusion. Monte-Sano (2010) focuses on evidence in context, with background information used to "situate and evaluate the evidence" (p. 548). De la Paz et al. (2017) similarly notes a causal element in their operationalization of proficient historical contextualization. With the exception of Monte-Sano (2010), however, these studies do not make explicit the expectation of a causal link or conclusion when defining written historical 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 toward 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 toward 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 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 a historical actor would behave in a certain manner. They found that students who considered more background knowledge 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 less successful students (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. In addition to including background information, 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 omit background information or their reasoning about 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 historical contextualization, and that many students do not contextualize when 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 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 constructed a historical context, 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 evident (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) promote 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 toward promoting it in student writing critical.

Teaching written historical contextualization in L1/L2 settings

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).

Teaching written historical reasoning in the CLIL context may differ from an L1 context since L2 proficiency can play a role when students read and write about history. The vocabulary of history, while not overly technical (Martin, 1991), is still challenging for some L2 students (Bernier, 1994). The demands of source-based writing, both in terms of reading sources and writing with them, is also an important consideration. When reading, the structure of textbooks and primary sources can both present a challenge (Schleppegrell & De Oliveira, 2006; Wineburg & Martin, 2009). As a result, students may need additional support when reading sources in history.

The impact of L2 proficiency when writing in an L2 is not entirely clear. Similar to L1 students, experience in writing may play a role in writing proficiency (Cumming et al., 2016). However, there do appear to be some differences between L1 and L2 writers, such as how students paraphrase sources (Van Weijen et al., 2019). Interpretation is an important aspect of writing in history. When students stay close to the original text when paraphrasing, the effectiveness of their interpretation may be compromised.

Additional support, such as form-focused instruction, may be important in helping students cope with the demands of source-based writing in an L2.

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, 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 students 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, Huijgen et al. (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 modeled 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, 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. A separate qualitative analysis of students' historical contextualization investigates 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 expected that students in this course would perform poorly in historical contextualization as measured in a written source-based argument, but that students in both groups would improve on the different elements of their historical reasoning

as measured by their document-based writing. In the areas of claim, supporting a claim with evidence, sourcing and corroboration, we anticipated similar improvements for both groups. For the second question, we expected 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.

One hundred forty 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.

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

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

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. 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

In this section we first explain the course aspects common to both conditions, and then itemize differences between the two conditions. Students in both conditions completed similar versions of a 28-hour historical reasoning course that took place for four hours weekly for seven weeks. 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 is studied the following semester in one of the required history courses. Students were made aware of this link to motivate them 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 modeled 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. 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 to 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 to assess the quality of several sample claims, as well as write their own. Later, 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 responses as models of how a student or historian might be expected to complete such a task. 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, 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" (p. 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 CLIL-based course integrated 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, 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 necessary background knowledge and helping students see the value of using historical context to ground their claims about the past. These activities were also used to advance students' procedural knowledge of contextualization since instructors first modeled contextualization in the discussion-based case study, and later supported students as they made and defended an evidence-based claim grounded in the historical context.

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 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).

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.

To further engage with the historical background knowledge, students in the experimental condition also completed one quote-sorting activity. During this activity, student 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).

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?

Figure 1. Case study for the unit Roman politics in the late Republic and early empire.

This activity took place during a lesson on politics, and was designed to help students consider the differences in the political systems because students in a previous version of this course sometimes confused 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 taught 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 the instructor led the students in selecting which background information would best help the reader understand the argument and decided where within the response to locate it in order to model effective contextualization. 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.

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. 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.

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.	

Figure 2. An excerpt from a text model and guiding questions activity used with students in the experimental and control conditions.

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 an extended version of another common task.

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

Figure 3. Lesson component timing for a partial sample lesson about politics in the late Roman Republic.

Data sources and analysis

Individual interest in history survey

We first measured students' interest in history in order to see if the two groups differed. 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.

Pretest and posttest document-based question

In addition to the two DBQs noted above, students completed a pretest and posttest DBQ. The pretest DBQ was used to assess their initial historical reasoning and writing. The pretest 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 posttest DBQ to assess their historical reasoning and writing. See Table 2 for an overview of the pre and posttest questions. 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. Sources were chosen that could be used as evidence to support multiple different claims. Both tasks used a similar question structure that was designed to solicit a claim and an argumentative response. The pretest was likely to elicit a response centering around criteria for accuracy, likely structured in a compare/contrast format. Students were likely to respond to the posttest with criteria based on societal structures. These responses could take several forms, including that of a compare/contrast structure. In both tests, the response was scored for quality of the claim and the use of evidence, not the use of a particular structure. In choosing sources, we also considered the potential to elicit other aspects of historical reasoning targeted in this study. For example, all primary sources included a head note including background information about the author that could be used in sourcing. Sources included corroborating information to support different claims, and historical contextualization could be used to consider the specific features of the time period. Between the two assessments, students participated in the experimental or control versions of the historical reasoning course.

The assessment 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 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. These aspects were chosen since they are present in models of historical reasoning in the literature and conform to the types of reasoning found in students' source-based historical writing (Monte-Sano, 2010; Van Boxtel & Van Drie, 2018; Van Drie & Van Boxtel, 2008; Wineburg, 1991).

Table 2. Pretest and posttest questions, sources, and word count requirements.

	Pretest	Posttest
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?
	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

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 pretest 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 correlations, 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 posttest. 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 posttest. 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 posttest. 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.

Exploring historical contextualization

We also prepared students' posttests 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, we first compare students' individual interest in history. 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. 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.

Individual interest in history

We used a survey of individual interest in history, as described in the methodology, in order to see if the two groups differed in terms of their interest. 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. The score in the experimental group (Mdn =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.

Historical reasoning in students' DBQ writing

Using Kolmogorov-Smirnov tests, we found that the pre- and posttest 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 pretest. The total score on the pretest in the experimental group ($Mdn = 5.00$) did not differ significantly from the total score on the pretest 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 posttest ($Mdn = 13.0$) was significantly higher than the total score for the pretest ($Mdn = 5.0$), $T = 1$, $p < .05$, $r = -.61$ Also in the control condition, the total score for the posttest ($Mdn = 12.0$) was significantly higher than the total score for the pretest ($Mdn = 4.0$), $T = 1$, $p < .05$, $r = -.61$. 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 posttest compared to the pretest with respect to all five aspects of the rubric (claim, evidence, source evaluation, historical contextualization and corroboration).

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 posttest essay. To test this hypothesis, we conducted a Mann-Whitney test. The historical contextualization score in the experimental group ($Mdn = 3.00$) did, however, not differ significantly from the score of the control group ($Mdn = 3.00$), $U = 2146.50$, $z = -1.11$, *ns,* $r = .09$. We explored

Table 3. Medians and ranges for pre and posttest by condition (experimental $n = 60$, control $n = 80$).

	Experimental		Control	
	Pretest	Posttest	Pretest	Posttest
	Mdn (Range)	*Mdn* (Range)	*Mdn* (Range)	*Mdn* (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)

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 ($Mdn = 3.00$) scored significantly higher on this aspect than students in the control condition ($Mdn = 2.00$), $U = 1704.50$, $z = -3.08$, $p < .01$, $r = .26$. 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.

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, posttest)." 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, posttest)." 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 essay would develop. 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, posttest)" 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, posttest)." 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 alternate explanations

In this section, we explore 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 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:

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

Alternate 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 was non-normal. The scores for individual questions and the total scores were significantly non-normal ($D(138) = 0.46$ to 0.84, $p < .05$). As a result, 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 (Mdn =11.00) did not differ significantly from the total score on the posttest 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 posttest. 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 posttest. We thus reject Alternate 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. 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.

Alternate 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. The results of Kolmogorov-Smirnov tests indicate that the distribution of the instances of historical contextualization 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 ($Mdn = 1.00$) did not differ significantly from the number in the control group ($Mdn = 1.00$), $U = 2350.00$, $z = -.222$, *ns*. We concluded that students in both groups used a similar number of instances of contextualization in their writing, and reject Alternate Hypothesis 2.

Since students in both groups included similar amounts of written contextualization, we continued to investigate whether other aspects of their

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

contextualization were the same. For our next hypothesis, we examined where students incorporated historical contextualization within 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. For this analysis we identified the location of each instance of historical contextualization, excluding geographical and chronological information, as described in the methodology. See Table 5 below. 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.

Alternate Hypothesis 3: Students in the control and experimental group used historical contextualization in different places in their essay.

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*

Table 5. Frequencies and percentages of the different locations of historical contextualization (HC) within student posttests (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%)

> *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 (posttest).

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, posttest).

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 unrelated 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 (posttest).

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.

Alternate Hypothesis 4: Students in the control and experimental group noted a connection between the historical contextualization and/or drew a pertinent conclusion to a similar extent.

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 conclusion or conclusion. For example, one student (Student 55, posttest) 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.

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. Based on the literature, we originally operationalized written historical reasoning as the inclusion of background knowledge with the most proficient historical contextualization used to support the claim. This qualitative analysis,

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

however, leads us to modify our operationalization of written historical contextualization as including background knowledge that is explicitly used to make an interpretation indicative of the characteristics of the time.

The following examples of students' written historical contextualization in this study illustrate the importance of the interpretive element. In some cases, students included background information that situated the argument in time or place. In this case, students included a brief mention (in italics) such as, "In *Roman Republic*, half of the gladiators were volunteers" (Student 4, posttest). Locating the argument within the Republic is useful contextualization because it helps the reader understand the timeframe of the argument, however, it is difficult to understand the importance of the Republic context in the overall argument. In contrast, Student 75 included background information that constructed a more robust historical context:

> *Roman citizens gave importance to military and gladiator games. For a man, virtus and gloria were very important values in Ancient Rome.* According to Cicero, the gladiators didn't fear of injuries of death, they kept fighting (posttest).

This student has started to construct a historical context that would enable the student to make an interpretation along the lines that Carr (1990) advocates for. This contextualization still leaves the reader to guess about conclusions should be drawn on the basis of this historical background. Another group of students created both a robust historical context and explicitly linked the background information to their argument, as in the following example:

> *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 (Dunkle) (Student 6, posttest).

This excerpt is different from the previous excerpts because it explicitly leads the reader toward the writer's interpretation that the possibility of gaining *virtus* and *gloria* make becoming a gladiator desirable for a freeman. This type of contextualization appears to be more in line with the expectations of using background information to make an interpretation.

In order to reflect this modified operationalization in the rubric, we propose 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. 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.

Discussion and conclusions

In this study, we investigated the written historical reasoning of students in a 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, 2012b), 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, 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 explicit instruction. All students, including those in both the experimental and control groups, significantly improved their historical contextualization. Historical contextualization, however, was among the lowest scoring aspects of written historical reasoning in both conditions and was more resistant to improvement. This is in line with the finding by others (Nokes et al., 2007; Reisman, 2012b). 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 of 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

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.

Figure 4. Proposed historical contextualization rubric.

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.

Surprisingly, students in the experimental group performed significantly better when writing claims than those in the control group. 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 and had a similar number of times to practice forming 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, Van Drie 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. 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 a written 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 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 also underscores the importance of connecting the reconstructed context to the argument as part of written historical contextualization and of the procedural knowledge needed to do this. Further research is needed to better understand similarities and differences between historical contextualization in writing and when reading.

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.

Acknowledgments

We would like to thank the four anonymous reviewers for their constructive feedback which enabled us to improve this manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

ORCID

Kristin A. Sendur http://orcid.org/0000-0003-4736-6766

References

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences*, *12* (3), 307–359. https://doi.org/10.1207/S15327809JLS1203_1

Bernier, A. (1994). Diversity's challenge in the classroom: Language and history pedagogy from the student optic. *The History Teacher*, *28*(1), 37–47. https://doi.org/10.2307/494286

Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction*, *20*(4), 485–522. https://doi.org/10.1207/S1532690XCI2004_2

Carr, E. H. (1990). *What is history?* Penguin UK.

Coffin, C. (2006). *Historical discourse: The language of time, cause and evaluation*. Bloomsbury Publishing.

Collins, A., Brown, J. S., & Holum, A. (1991). Cognitive apprenticeship: Making thinking visible. *American Educator*, *15*(3), 6–11. https://www.aft.org/ae/winter1991/collins_brown_holum

Coyle, D., Hood, P., & Marsh, D. (2010). *Content and language integrated learning*. Ernst Klett Sprachen.

Cumming, A., Lai, C., & Cho, H. (2016). Students' writing from sources for academic purposes: A synthesis of recent research. *Journal of English for Academic Purposes*, *23*, 47–58. https://doi.org/https://doi.org/10.1016/j.jeap.2016.06.002

Dallinger, S., Jonkmann, K., Hollm, J., & Fiege, C. (2016). The effect of content and language integrated learning on students' English and history competences–Killing two birds with one stone? *Learning and Instruction*, *41*, 23–31. https://doi.org/10.1016/j.learninstruc.2015.09.003

De Graaff, R., Koopman, G. J., Anikina, Y., & Westhoff, G. (2007). An observation tool for effective L2 pedagogy in content and language integrated learning (CLIL). *International Journal of Bilingual Education and Bilingualism*, *10*(5), 603–624. https://doi.org/10.2167/beb462.0

De la Paz, S., Felton, M., Monte-Sano, C., Croninger, R., Jackson, C., Deogracias, J. S., & Hoffman, B. P. (2014). Developing historical reading and writing with adolescent readers: Effects on student learning. *Theory & Research in Social Education*, *42*(2), 228–274. https://doi.org/10.1080/00933104.2014.908754

De la Paz, S., Ferretti, R., Wissinger, D., Yee, L., & MacArthur, C. (2012). Adolescents' disciplinary use of evidence, argumentative strategies, and organizational structure in writing about historical controversies. *Written Communication*, *29*(4), 412–454. https://doi.org/10.1177/0741088312461591

De la Paz, S., Monte-Sano, C., Felton, M., Croninger, R., Jackson, C., & Piantedosi, K. W. (2017). A historical writing apprenticeship for adolescents:

Integrating disciplinary learning with cognitive strategies. *Reading Research Quarterly*, *52*(1), 31–52. https://doi.org/doi:10.1002/rrq.147

De Oliveira, L. C. (2011). *Knowing and writing school history: The language of students' expository writing and teachers' expectations*. IAP.

Dunkle, R. (2002). *Roman gladiatorial games*. Brooklyn College. http://depthome.brooklyn.cuny.edu/classics/gladiatr/index.htm

Eurydice. (2006) . *Content and language integrated learning at school in Europe*.

Goo, J., Granena, G., Yilmaz, Y., & Novella, M. (2015). Implicit and explicit instruction in L2 learning. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (Vol. 48, pp. 443–482). John Benjamins Publishing Company.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476. https://doi.org/10.1037/0022-0663.99.3.445

Grant, S. (2018). Teaching practices in history education. In S. A. Metzger & L. M. Harris (Eds.), *The Wiley international handbook of history teaching and learning* (pp. 419–448). Wiley-Blackwell.

Greene, S. (1994). The problems of learning to think like a historian: Writing history in the culture of the classroom. *Educational Psychologist*, *29*(2), 89–96. https://doi.org/10.1207/s15326985ep2902_4

Hillocks, J. G. (1986). *Research on written composition: New directions for teaching*. ERIC.

Hirvela, A. (2016). Academic reading into writing. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 127–137). Routledge.

Huijgen, T., Van Boxtel, C., Van de Grift, W., & Holthuis, P. (2017). Toward historical perspective taking: Students' reasoning when contextualizing the actions of people in the past. *Theory & Research in Social Education*, *45*(1), 110–144. https://doi.org/10.1080/00933104.2016.1208597

Huijgen, T., Van de Grift, W., Van Boxtel, C., & Holthuis, P. (2017). Teaching historical contextualization: The construction of a reliable observation instrument [Journal article]. *European Journal of Psychology of Education*, *32*(2), 159–181. https://doi.org/10.1007/s10212-016-0295-8

Huijgen, T., Van de Grift, W., Van Boxtel, C., & Holthuis, P. (2018). Promoting historical contextualization: The development and testing of a pedagogy. *Journal of Curriculum Studies*, *50*(3), 410–434. https://doi.org/10.1080/00220272.2018.1435724

Hyland, K. (2007). Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing*, *16*(3), 148–164. https://doi.org/10.1016/j.jslw.2007.07.005

Kuhn, D., Shaw, V., & Felton, M. (1997). Effects of dyadic interaction on argumentive reasoning. *Cognition and Instruction*, *15*(3), 287–315. https://doi.org/10.1207/s1532690xci1503_1

Lee, P. (2005). Putting principles into practice: Understanding history. In S. Donovan & J. Bransford (Eds.), *How students learn* (pp. 31–77). National Academies Press.

Lee, P. (2011). Historical literacy and transformative history. In L. Perikleous & D. Shemilt (Eds.), *The future of the past: Why history education matters* (pp. 129–167). Association for historical dialogue and research.

Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement*. *70*(4). https://doi.org/10.1177/0013164409355699

Martin, J. (1991). Distilling knowledge and scaffolding text. In E. Ventola (Ed.), *Functional and systemic linguistics: Approaches and uses* (Vol. 55, pp. 307–337). Mouton de Gruyter.

McCarthy Young, K., & Leinhardt, G. (1998). Writing from primary documents. *Written Communication*, *15*(1), 25–68. https://doi.org/doi:10.1177/0741088398015001002

Miller, R. T., Mitchell, T. D., & Pessoa, S. (2016). Impact of source texts and prompts on students' genre uptake. *Journal of Second Language Writing*, *31*, 11–24. https://doi.org/10.1016/j.jslw.2016.01.001

Monte-Sano, C. (2008). Qualities of historical writing instruction: A comparative case study of two teachers' practices. *American Educational Research Journal*, *45*(4), 1045–1079. https://doi.org/10.3102/0002831208319733

Monte-Sano, C. (2010). Disciplinary literacy in history: An exploration of the historical nature of adolescents' writing. *The Journal of the Learning Sciences*, *19*(4), 539–568. https://doi.org/10.1080/10508406.2010.481014

Monte-Sano, C. (2011). Beyond reading comprehension and summary: Learning to read and write in history by focusing on evidence, perspective, and interpretation. *Curriculum Inquiry*, *41*(2), 212–249. https://doi.org/10.1111/j.1467-873X.2011.00547.x

Monte-Sano, C., & De la Paz, S. (2012). Using writing tasks to elicit adolescents' historical reasoning. *Journal of Literacy Research*, *44*(3), 273–299. https://doi.org/doi:10.1177/1086296X12450445

Myskow, G., & Ono, M. (2018). A matter of facts: L2 writers' use of evidence and evaluation in biographical essays. *Journal of Second Language Writing*, *41*, 55–70. https://doi.org/10.1016/j.jslw.2018.08.002

Nokes, J. D. (2017). Exploring patterns of historical thinking through eighth-grade students' argumentative writing. *Journal of Writing Research*, *8*(3), 437–467. https://doi.org/10.17239/jowr-2017.08.03.02

Nokes, J. D., & De la Paz, S. (2018). Writing and argumentation in history education. In S. A. Metzger & L. McArthur Harris (Eds.), *The Wiley international handbook of history teaching and learning* (pp. 551–578). John Wiley & Sons, Inc.

Nokes, J. D., Dole, J. A., & Hacker, D. J. (2007). Teaching high school students to use heuristics while reading historical texts. *Journal of Educational Psychology*, *99*(3), 492–504. https://doi.org/10.1037/0022-0663.99.3.492

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

O'Neill, D. K. (2012). Designs that fly: What the history of aeronautics tells us about the future of design-based research in education. *International Journal of Research & Method in Education*, *35*(2), 119–140. https://doi.org/10.1080/1743727X.2012.683573

Oattes, H., Fukkink, R., Oostdam, R., de Graaff, R., & Wilschut, A. (2020). A showdown between bilingual and mainstream education: The impact of language of instruction on learning subject content knowledge. *International Journal of Bilingual Education and Bilingualism*, 1–14. https://doi.org/10.1080/13670050.2020.1718592

Pérez-Cañado, M. L. (2012). CLIL research in Europe: Past, present, and future. *International Journal of Bilingual Education and Bilingualism*, *15*(3), 315–341. https://doi.org/10.1080/13670050.2011.630064

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire

(MSLQ). *Educational and Psychological Measurement*, *53*(3), 801–813. https://doi.org/10.1177/0013164493053003024

Reisman, A. (2012a). The 'document-based lesson': Bringing disciplinary inquiry into high school history classrooms with adolescent struggling readers. *Journal of Curriculum Studies*, *44*(2), 233–264. https://doi.org/10.1080/00220272.2011.591436

Reisman, A. (2012b). Reading like a historian: A document-based history curriculum intervention in urban high schools. *Cognition and Instruction*, *30*(1), 86–112. https://doi.org/10.1080/07370008.2011.634081

Reisman, A., & McGrew, S. (2018). Reading in history education: Text, sources, and evidence. In S. A. Metzger & L. M. Harris (Eds.), *The Wiley international handbook of history teaching and learning* (pp. 529–550). Wiley-Blackwell.

Reisman, A., & Wineburg, S. (2008). Teaching the skill of contextualizing in history. *The Social Studies*, *99*(5), 202–207. https://doi.org/10.3200/TSSS.99.5.202-207

Reznitskaya, A., Anderson, R. C., McNurlen, B., Nguyen-Jahiel, K., Archodidou, A., & Kim, S.-Y. (2001). Influence of oral discussion on written argument. *Discourse Processes*, *32*(2–3), 155–175. https://doi.org/10.1080/0163853X.2001.9651596

Reznitskaya, A., & Gregory, M. (2013). Student thought and classroom language: Examining the mechanisms of change in dialogic teaching. *Educational Psychologist*, *48*(2), 114–133. https://doi.org/10.1080/00461520.2013.775898

Schleppegrell, M., & de Oliveira, L. C. (2006). An integrated language and content approach for history teachers. *Journal of English for Academic Purposes*, *5*(4), 254–268. https://doi.org/10.1016/j.jeap.2006.08.003

Sendur, K. A., Van Boxtel, C., & Van Drie, J. (2021). Undergraduate L2 students' performance when evaluating historical sources for reliability. *English for Specific Purposes*, *61*, 17–31. https://doi.org/10.1016/j.esp.2020.08.004

Sendur, K. A., Van Drie, J., Van Boxtel, C., & Kan, K.-J. (2020). Historical reasoning in an undergraduate CLIL course: Students' progression and the role of language proficiency. *International Journal of Bilingual Education and Bilingualism*, 1–17. https://doi.org/10.1080/13670050.2020.1844136

Shemilt, D. (1984). Beauty and the philosopher: Empathy in history and classroom. In A. K. Dickinson, P. J. Lee & P. J. Rogers (Eds.), *Learning history* (pp. 39–84). Heinemann.

Stoel, G., Logtenberg, A., Wansink, B., Huijgen, T., Van Boxtel, C., & Van Drie, J. (2017). Measuring epistemological beliefs in history education: An exploration of naïve and nuanced beliefs. *International Journal of Educational Research*, *83*, 120–134. https://doi.org/10.1016/j.ijer.2017.03.003

Stoel, G., Van Drie, J., & Van Boxtel, C. (2017). The effects of explicit teaching of strategies, second-order concepts, and epistemological underpinnings on students' ability to reason causally in history. *Journal of Educational Psychology*, *109*(3), 321–337. https://doi.org/10.1037/edu0000143

Van Boxtel, C., & Van Drie, J. (2012). "That's in the time of the Romans!" Knowledge and strategies students use to contextualize historical images and documents. *Cognition and Instruction*, *30*(2), 113–145. https://doi.org/10.1080/07370008.2012.661813

Van Boxtel, C., & Van Drie, J. (2013). Historical reasoning in the classroom: What does it look like and how can we enhance it? *Teaching History*, (150), 44–52. http://www.jstor.org/stable/43260513

Van Boxtel, C., & Van Drie, J. (2018). Historical reasoning: Conceptualizations and educational applications. In S. A. Metzger & L. McArthur Harris (Eds.), *The Wiley*

international handbook of history teaching and learning (pp. 149). John Wiley & Sons, Inc.

Van Drie, J., Braaksma, M., & Van Boxtel, C. (2015). Writing in history: Effects of writing instruction on historical reasoning and text quality. *Journal of Writing Research*, *7*(1), 123–156. https://doi.org/10.17239/jowr-2015.07.01.06

Van Drie, J., & Van Boxtel, C. (2008). Historical reasoning: Towards a framework for analyzing students' reasoning about the past. *Educational Psychology Review*, *20* (2), 87–110. https://doi.org/10.1007/s10648-007-9056-1

Van Drie, J., & Van de Ven, P.-H. (2017). Moving ideas: An exploration of students' use of dialogue for writing in history. *Language and Education*, *31*(6), 526–542. https://doi.org/10.1080/09500782.2017.1326504

Van Weijen, D., Rijlaarsdam, G., & Van den Bergh, H. (2019). Source use and argumentation behavior in L1 and L2 writing: A within-writer comparison. *Reading and Writing*, *32*(6), 1635–1655. https://doi.org/10.1007/s11145-018-9842-9

Wegerif, R., Mercer, N., & Dawes, L. (1999). From social interaction to individual reasoning: An empirical investigation of a possible socio-cultural model of cognitive development. *Learning and Instruction*, *9*(6), 493–516. https://doi.org/10.1016/S0959-4752(99)00013-4

Westhoff, G. (2004). The art of playing a pinball machine. Characteristics of effective SLA-tasks. *Babylonia*, *12*(3), 58–62.

Wineburg, S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology*, *83*(1), 73–87. https://doi.org/10.1037/0022-0663.83.1.73

Wineburg, S. (1998). Reading Abraham Lincoln: An expert/expert study in the interpretation of historical texts. *Cognitive Science*, *22*(3), 319–346. https://doi.org/10.1207/s15516709cog2203_3

Wineburg, S., & Fournier, J. (1994). Contextualized thinking in history. In M. Carretero & J. F. Voss (Eds.), *Cognitive and instructional processes in history and the social sciences* (pp. 285–308). Routledge.

Wineburg, S., & Martin, D. (2009). Tampering with history: Adapting primary sources for struggling readers. *Social Education*, *73*(5), 212–216.

Wineburg, S., Smith, M., & Breakstone, J. (2012). New directions in assessment: Using Library of Congress sources to assess historical understanding. *Social Education*, *76*(6), 290–293.

Appendix. Historical reasoning rubric

	Claim	Use of Evidence	Source Evaluation	Historical Contextualization	Corroboration
4	Presents a clear and accurate claim that adequately addresses the question.	The evidence is accurate, relevant and sufficient to support the claim & the evidence is accurately explained at least once & explicitly linked to the claim at least once.	Refers to at least 1 author by name or title & notes relevant feature(s) of the primary source (PS). Indicates potential effect of the feature on the information &/or explains the effect &/or uses the feature to further the argument. (at least 2/3)	Provides accurate and relevant historical context (HC) (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.	Uses multiple sources to support the same point at least once & explicitly indicates an appropriate link between the sources & explains the link by noting how they are similar
3	Presents a clear and accurate claim that partially addresses the question.	The evidence is accurate, relevant and sufficient to support the claim. The evidence may be accurately explained at least once or explicitly linked to the claim at least once.	Refers to at least 1 author by name or title & notes relevant feature(s) of the PS. Indicates potential effect of the feature on the information or explains the effect or uses the feature to further the argument. (1/3)	Provides accurate and relevant historical context It may be used to implicitly situate &/or further the argument.	Uses multiple sources to support the same point at least once & explicitly indicates an appropriate link between the sources & notes that they are similar
2	Accurately restates the question or topic without directly stating a claim. May contain minor errors.	The evidence is insufficient and may contain irrelevant or inaccurate information. The evidence is explained &/or explicitly linked to the main idea at least once. The explanation or link may be inaccurate.	Refers to at least 1 author by name or title & notes relevant feature(s) of the PS. There may be an attempt to note the effect or use it to further the argument. If included, the interpretation undermines the argument or has errors.	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.	Uses multiple sources to support the same point at least once & explicitly indicates an inappropriate or unclear link between the sources

(*Continued*)

(Continued).

	Claim	Use of Evidence	Source Evaluation	Historical Contextualization	Corroboration
1	The main idea is difficult to discern, implied or marginally addresses the questions &/or is inconsistent with the evidence in the sources &/or the language makes the intended meaning somewhat unclear	The evidence is insufficient and may contain irrelevant or inaccurate information. The evidence is not explained & not explicitly linked to the claim &/or the evidence is primarily copy-pasted.	Refers to at least 1 author by name or title & notes irrelevant or inaccurate feature(s) of the PS. There may be an attempt to note the effect or use it to further the argument. The interpretation may have errors.	Provides historical context that is historically inaccurate &/or largely irrelevant.	Uses multiple sources to support the same point at least once & treats sources separately without explicit corroboration (can look list-like).
0	There is no main idea or the main idea is copy-pasted from the sources or the language makes the main idea incomprehensible.	There is no evidence &/or the evidence is primarily irrelevant	Notes the author without any attempt to assess reliability or fails to note the author or title.	Does not note historical context.	Uses one source for support when multiple are available.