Undergraduate L2 students' performance when evaluating historical sources for reliability

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

DOI
10.1016/j.esp.2020.08.004

Publication date
2021

Document Version
Final published version

Published in
English for specific purposes

English for Specific Purposes 61 (2021) 17–31

Contents lists available at ScienceDirect

English for Specific Purposes

journal homepage: http://ees.elsevier.com/esp/default.asp

Undergraduate L2 students' performance when evaluating historical sources for reliability

Kristin A. Sendur [a,b,*], Carla van Boxtel [a], Jannet van Drie [a]

[a] University of Amsterdam, Research Institute of Child Development and Education, PO Box 15776, 1001 NG, Amsterdam, Netherlands
[b] Sabanci University, Orhanli-Tuzla, 34956, Istanbul, Turkey

ARTICLE INFO

Article history:
Available online 10 September 2020

Keywords:
Writing
Disciplinary literacy
Higher education
L2

ABSTRACT

Evaluating historical sources for reliability, an aspect of sourcing, is a key feature of historical reasoning. While well-studied among proficient and L1 students, the performance of L2 students and the role of their English proficiency is not as well understood. This study examines the oral and written historical reasoning of undergraduate L2 students when evaluating historical sources for reliability and writing with historical sources. In an analysis of think aloud protocols and written answers, we find that students are able to reason historically, albeit in a quite shallow manner. We identify the use of historical contextualization and forming a complete answer as two areas of difficulty and the co-existing role that language proficiency appears to play in some students' performance. A comparison of students' written and oral answers demonstrates that while most students score similarly in both modes, written answers are generally less rich in detail. Finally, we trace students' use of the same historical sources in document-based question essays. We find that while students consistently use the historical sources as evidence, they rarely consider reliability.

1. Introduction

In sourcing, a key feature of historical reasoning, a reader considers who has created the source and the implications of that creator on the intended meaning (Monte-Sano, 2010; Van Drie & Van Boxtel, 2008; Wineburg, 1991). One of the purposes of sourcing is to make a claim of reliability. Students' evaluation of sources has been studied in many areas, such as climate change, media literacy and citizenship (Barzilai & Zohar, 2012; Bråten, Strømsø, & Britt, 2009; McGrew, Ortega, Breakstone, & Wineburg, 2017; Walraven, Brand-Gruwel, & Boshuizen, 2009), This work focuses specifically on sourcing by undergraduate L2 students within the discipline of history since sourcing is a key component of reasoning in history (Van Drie & Van Boxtel, 2008; Wineburg, 1991).

History is an appropriate discipline for those learning an L2 because there is ample opportunity for speaking (Lo, 2014) and writing (Monte-Sano, 2010). Empirical studies of sourcing in history conducted using oral language and writing have informed our understanding of students' performance and the difficulties they face in each of these areas (Monte-Sano, 2008; Wineburg, 1991). Research that explicitly connects school history writing produced after an internal thought process,

* Corresponding author. Sabanci University, Orhanli-Tuzla, 34956, Istanbul, Turkey.
E-mail addresses: kristin.sendur@gmail.com (K.A. Sendur), C.A.M.vanBoxtel@uva.nl (C. van Boxtel), J.P.vanDrie@uva.nl (J. van Drie).

https://doi.org/10.1016/j.esp.2020.08.004
0889-4906/

however, is needed in order to understand how students translate their thoughts into writing, and what they choose to include or exclude in their final written answers. Such research can have significant implications for instruction, particularly for those teaching English for Academic Purposes (EAP), since it can inform the types of information and language necessary to formulate a successful answer.

Language proficiency, particularly the language needed for EAP, is important when evaluating historical sources, and may pose a significant challenge for those reading and writing in a non-native language. Nokes (2011) identifies text comprehension as a major barrier to historical reasoning for some students. While there is some research that addresses historical reasoning and struggling readers (De La Paz, 2005; Reisman, 2012), much of the research on sourcing involves students with high levels of English proficiency (Wineburg, 1991) or where the proficiency level is not noted (Britt & Aglinskas, 2002). Thus, research that focuses on the growing population of L2 students is important. While recent studies investigate the writing patterns of undergraduate L2 writers and instructional approaches in the discipline of history (e.g. Miller, Mitchell, & Pessoa, 2014; Mitchell & Pessoa, 2017; Myskow & Ono, 2018), studies that closely examine the sourcing performance and difficulties of L2 students are needed to understand how they compare to those with higher levels of English proficiency.

This study aims to provide insight into how undergraduate L2 students reason when evaluating the reliability of English-language historical sources. The study also investigates differences between participants' oral and written answers to a sourcing task as well as the use of the same sources in an essay. Finally, this study looks into the language-related difficulties that L2 students encounter when reading historical sources. This research was conducted with native-Turkish speaking undergraduate students studying in an EAP program at an English-medium university in Istanbul, Turkey.

1.1. Evaluating sources for reliability in history

Sourcing is a complex but well-studied aspect of historical reasoning. In his landmark think aloud study, Wineburg (1991) first identified sourcing as a heuristic that historians, but not students, regularly use when evaluating the reliability of historical sources. Considering the authorship of a source helps a reader place it within its genre and historical context. In contrast, Wineburg (1991) found that a lack of sourcing led to issues in meaning making for at least some students.

Similar to Wineburg's (1991) findings, research focusing on writing also demonstrates that students do not consistently consider sourcing. Monte-Sano (2010) found that when writing a document-based question (DBQ) essay, some high school students acknowledged their sources, however students did not demonstrate more advanced concepts of sourcing, such as acknowledging the potential bias of the author. Nokes (2017) classified eighth grade students for the use and quality of sourcing when writing an argumentative DBQ-style essay and a related reliability task. He found that few used a sourcing heuristic to strengthen the quality of the essay. However, when asked directly to evaluate the usefulness and trustworthiness of sources, more were able to do so. Similarly, Britt and Aglinskas (2002) found that more students in their study used sourcing after instruction.

When they do consider the source, students make many different types of arguments about the extent of reliability, many of which focus on the author's background (Rouet, Britt, Mason, & Perfetti, 1996). First, students note the author's position, including the level of expertise (Barzilai & Zohar, 2012; Britt & Aglinskas, 2002) and job (Britt & Aglinskas, 2002). Second, access to information in view of the nature of the author's participation was also used (Britt & Aglinskas, 2002). A third important factor is the author's purpose, or motivation, in writing the source (Britt & Aglinskas, 2002). The author's perspective, or bias, was also considered by some students in Barzilai and Zohar's (2012) study. Finally, students considered characteristics beyond the author's background, such as when it was produced (Britt & Aglinskas, 2002). Students use this background information as the basis for their argument about the extent of the author's reliability.

These studies demonstrate that, in contrast to professionals, students do not tend to spontaneously make use of a sourcing heuristic when reading and writing about historical sources. When they do reason about the reliability of a historical text, students are capable of making several different types of arguments. They are not, however, always successful in their attempts. One reason for the lack of any, or at least sophisticated, sourcing in students' written work may be accounted for by Felton and Herko's (2004) finding that while students are capable of oral argumentation, they struggle when translating those arguments into writing. Or it may be possible that students have learned the mechanics of sourcing before understanding how and why it is important, similar to the findings of Lee and Ashby (2000). In combination with common student misconceptions of what is expected from history writing tasks, such as those outlined by Greene (1994), students may also not see the value of adding sourcing as a part of their writing. While studies of sourcing in both written and oral language have made great contributions to the literature, there is insufficient research that explores the different types of difficulties that students experience and the role of students' language proficiency in those difficulties.

1.2. Reading and writing about history in an L2

Textbooks are a heavily utilized resource in many history classrooms (Cohen, 2005; Yildirim, 2006). Reading these textbooks, however, is challenging to students because of the grammatical structures common to history textbooks, such as nominalization and reasoning within the clause (Martin, 1991), and the abstract nature of historical writing (Schleppegrell &

de Oliveira, 2006). The diverse genres and language of historical sources, such as those in ancient monuments, biographies and poetry may cause even greater difficulty. Wineburg and Martin (2009) argue that primary sources are difficult for many students to read because of issues such as "archaic phrasing and obscure terminology, and a context foreign at best" (p. 212).

While history as a discipline is built on interpretation, this interpretation is often difficult to discern in history textbooks, which tend to present history as fact (Unsworth, 1999). In a recent study, Myskow (2018) found that both high school and university textbooks rarely include judgement of historical actors, a key factor in sourcing. These features of history textbooks may affect students' writing as well, by masking the need to include such features in their own texts. Teachers may be able to partially overcome this, as Miller, Mitchell, and Pessoa (2016) identified the combination of an argumentative prompt and sources without an overt argument resulted in more argumentative writing. Similarly, instruction and writing tasks focused on interpretation seem to show greater gains in argumentation and historical reasoning than those with a focus on summary and recall of facts (Monte-Sano, 2008).

L2 students' concept of themselves as writers may also play a role in how deeply they engage with sources. Wette (2017) found that most citations of L2 students in her study were used in a manner in which students seemed to defer to the sources as authorities. As students gain confidence and competence in source-based writing, however, their approach to sources may change. In their longitudinal study, Thompson, Morton, and Storch (2013) found that some students started to more critically evaluate their sources towards the end of the academic year and incorporate more of their own ideas.

Several recent studies have investigated writing in history in an L2 context. At the secondary level, Lorenzo (2017) identified different cognitive discourse functions, including evaluation, that students in a Content and Language Integrated Learning (CLIL) history class were able to include in their narrative writing. Others have looked at writing patterns in the discipline of history. Miller et al. (2014) found that undergraduate L2 students with better essays interpreted source texts in ways that contributed to the overall argument, while students who wrote essays with lower scores undermined their argument by, for example, contradicting their own argument with contradictory statements. In a corpus study of students in a university-level CLIL history class, Myskow and Ono (2018) describe patterns of evidence students use to support their evaluations of a historical figure. Two patterns that include the combination of circumstances, activities, and transformations are identified as more successful than those who do not include circumstances.

Llinares and Whittaker (2007) have investigated the differences in oral and written language by L2 students in the domain of history. In this study, younger students showed few differences between oral and written production. A later study compared oral and written production in CLIL and L1 history classes (Llinares & Whittaker, 2010). In both speaking and writing, L2 students incorporated less advanced levels of the category circumstances, such as chronology and location. In contrast, L1 students included more advanced notions, including causation and manner, in their writing.

These studies have examined the complexity and challenges of reading and writing in history. While the studies above have shown that L2 students are able to write using diverse history genres, these studies have not explored in depth how students evaluate the reliability of the historical sources they use. L1 students tend not to consider the reliability of historical sources unless prompted to do so. When prompted, students are able to produce different types of arguments to evaluate the historical source. It is important to know if L2 students reason similarly. It is also important to further explore the difficulties that students encounter in order to help plan instruction that can challenge their misconceptions and provide the appropriate support to students.

This study uses a multiple modality approach by combining an analysis of undergraduate L2 students' oral and written sourcing when evaluating the reliability of a historical source. We later trace students' reasoning from the sourcing task to their DBQ essays. By studying students' sourcing in different modalities, we are able to explore the quality of students' sourcing, difficulties students encounter in a given mode, and how their sourcing differs in oral and written modes.

1.3. Research questions

The present study addresses the following research questions:

1. To what extent do undergraduate L2 students make claims supported by arguments when reasoning about the reliability of a historical source?
2. What difficulties do they encounter that can be attributed to language proficiency?
3. What differences are there between L2 students' reasoning about aspects of reliability during the think aloud sessions and their reasoning in writing
   a. when explicitly asked to make claims about reliability?
   b. when asked to use the sources to answer a historical question?

In this study, we would expect that undergraduate L2 students would be able to make claims of reliability supported by arguments when prompted to do so. However, students may experience language-related difficulties comprehending the vocabulary and structures unique to historical sources at the level necessary to draw appropriate oral and written conclusions about the reliability of the source, especially as instructor support is removed.

2. Method

2.1. Participants

Eleven undergraduate Turkish students at a small private English-medium university in Istanbul participated in this study. All students enrolled in a historical reasoning course were invited to participate. Students who volunteered and could fit interviews into their study schedule participated. (See Table 1 for participants' backgrounds.) At the time of data collection, participants were studying at the B2 level according to Common European Framework of Reference for Languages (CEFR) in an intensive English program. Each student's primary instructor informally assessed the participating student as strong or weak within the B2 level based on the instructor's experience working with students at the B2 level in the university. Students at the university typically spend one to two semesters in the program before beginning undergraduate coursework. Almost half of the students at the university major in engineering. History is not offered as a major for undergraduate students.

2.2. Data collection

Students S1 through S6 participated in the fall 2016 semester and the remaining five students participated in the spring 2017 semester. All participants were simultaneously enrolled in a historical reasoning course as a part of an intensive English program. The course was taught by five different instructors using highly scripted lesson plans over 28 h in the fall semester and 32 h in the spring semester.

In the course, students were introduced to the following concepts of historical reasoning: argumentation, historical contextualization, and source evaluation and corroboration heuristics. During the course, students were explicitly taught how to evaluate a source, including which aspects of the source's background to consider when making an assessment of reliability. The course was structured using a CLIL model (Coyle, Hood, & Marsh, 2010) in which students learned about the history of gladiators in the late Roman Republic and early Empire. Because of the difficulty of reading in history, course instructors supported students' reading comprehension through explicit instruction in text annotation. Specifically, students learned to note main ideas and identify questions. Students' reading comprehension was also supported through the use of graphic organizers when reading historical sources, and vocabulary assistance with discipline-specific and general academic terms.

2.3. Materials and data sources

There are three sources of data for this study. We collected students' written answers to three Source Evaluations (SE) as well as the transcript of their oral answer to the same task. We also collected three DBQ per student.

2.3.1. Source evaluation tasks

As a part of the historical reasoning course, students completed three SE tasks modeled after Wineburg, Smith and Breakstone's (2012) Historical Assessment of Thinking to assess their proficiency in evaluating the reliability of an historical source. This task type has been validated by Smith, Breakstone, and Wineburg (2019). The tasks were developed by the first author and the first SE was validated by two post-doctoral researchers teaching in a first year history survey course. (See Table 2 for an outline of SE timing and sources.) To validate the task we met individually with the two post-doctoral researchers. We asked them to read the task and think out loud while answering the question to determine if the question and answer choices were clear, easy to understand, and historically appropriate. We also presented our answer key and discussed which answers could be considered accurate and reasonable given the information available to students. We did not make any changes to the task based on the think aloud, but we did add an additional correct answer to the answer key based on our

Table 1
Characteristics of Participants.

Student	Gender	Intended Faculty	English level	Instructor
S1	M	Engineering and Science	Strong	A
S2	M	Engineering and Science	Strong	B
S3	F	Engineering and Science	Strong	B
S4	M	Management	Strong	C
S5	F	Arts and Social Sciences	Weak	C
S6	M	Engineering and Science	Strong	C
S7	F	Management	Weak	D
S8	F	Engineering and Science	Strong	E
S9	F	Management	Strong	C
S10	M	Engineering and Science	Strong	C
S11	M	Engineering and Science	Weak	C

Note. Relative English proficiency within the B2 level was assessed by each student's instructor.

Table 2
Overview of Source Evaluation Tasks.

	Source Evaluation 1 (SE1)	Source Evaluation 2 (SE2)	Source Evaluation 3 (SE3)
When the SE was completed	Lesson 3	Lesson 4	Following the course quiz (S1–S6); Lesson 8 (S7–S11)
Roman history lesson focus	Socioeconomics	Politics	Culture and Values
Primary Source	*The Deified Augustus* by Suetonius (biography)	*Deeds of the Divine Augustus* by Augustus (monument)	*Satire* 6 by Juvenal (poem)
	163 words	125 words	118 words
	Flesch-Kincaid Grade Level: 8.8	Flesch-Kincaid Grade Level: 10.1	Flesch-Kincaid Grade Level: 7.3
SE task facts	Fact 1: proximity	Fact 1: historical context and position	Fact 1: proximity
	Fact 2: position	Fact 2: corroboration	Fact 2: purpose
	Fact 3: perspective	Fact 3: purpose	Fact 3: perspective
In-class support before think aloud	Instructor-led class discussion of the text and sourcing questions.	Reading comprehension support provided by the instructor during the lesson.	None
	Sample answer modeled by the instructor	Students complete sourcing questions individually.	

discussion with the postdoctoral researchers. We used the feedback from this validation to inform our design of the other two tasks.

Each SE took place in the middle of one of the following modules in the historical reasoning course: socioeconomics, politics, or culture and values. For example, SE2, which is an account of the accomplishments of the first Roman emperor, Augustus, took place within the context of the Roman politics unit so students had necessary background information to help them complete the SE. The first two tasks were a part of the course during both semesters. The third task was used in the course during the spring semester; during the fall semester it was only used for this study.

Three excerpts of historical sources from the Roman Empire were used for the SEs, and are available in Appendix A. Based on principles from Wineburg and Martin (2009), we modified all primary sources in the historical reasoning course by shortening them, simplifying the presentation by including white space around the primary source, and simplifying the vocabulary and structure. A short biography of the author was placed in a box above the excerpt of the historical source. Three additional facts about the author or source were provided below the excerpt. For the task, students explained in writing "which 1 of the 3 facts above might cause you to question the reliability of the (author's) account" as a source for the unit under study.

The facts for each task were chosen based on the nature of the source and the historical content studied in the course since students were instructed to evaluate the reliability of the source within the context of the unit under study. As noted in Table 2, several types of facts were chosen. For example, Seutonius' social class was included as fact 3 in SE1 because it could have affected his perspective of Augustus, and as a part of the unit students had studied Roman social classes. Proximity denoted when an author lived, position indicated the job held by the author and purpose pointed to why the author may have written the source.

Any of the facts could have been conceivably chosen for a given SE if supported with an appropriate argument, although some facts required less interpretation than others.

2.3.2. Thinking-aloud

Students who participated in this study completed each of the three SEs as a think aloud and provided their written answers. When students were scheduled to complete the SE task as a part of their historical reasoning course, participants were pulled out of class to complete the SE task as a think aloud individually with a trained research assistant or the first author. Prior to the first think aloud, participants completed a training session with the first author or a historical reasoning course instructor during which they practiced thinking aloud with another historical source from the course. Students were instructed to say whatever came to mind as they worked through the SE task. Procedures similar to Van Someren, Barnard and Sandberg (1994) were followed. Students could respond in English or Turkish.

During the think aloud collection, students could ask vocabulary and procedural questions, an analysis of which is presented later. Students individually wrote their answers to each SE during the think aloud session. Our analysis makes use of both students' written answer and think aloud transcript. In the case of S1 and S2, we only used the dictated final answer of SE2 that was included in the transcription of the think aloud protocol because S1 lost his written answer and S2 changed his written answer after leaving the think aloud. The written answer and transcript are used for all other students and SEs.

2.3.3. Document-based questions

After concluding each unit in the historical reasoning course, students wrote a DBQ essay. See Table 3 for essay prompts. For each DBQ, students were given a word count (as a range), a list of sources from the course that could be useful (including the source from the related SE task), and the aspects of historical reasoning they should incorporate into their writing.

Table 3
DBQ Prompts and Word Count Requirements.

	DBQ1	DBQ2	DBQ3
Task prompt	Describe Rome's social hierarchy and explain one way it affected people's lives.	Why did politicians sponsor gladiator shows?	How did Roman society view gladiators? Explain two views.
Word count	120-150 words	250-300 words	250-300 words

2.4. Data analysis

All think aloud sessions were audio-recorded, transcribed into c-units and translated, if necessary. A c-unit has been defined by Loban (1976) as "each individual predication with all of its modifiers" (p. 9). Within each transcript all argument chains, consisting of one or more claims and all related arguments, were identified.

We prepared a coding scheme with two distinct levels to analyze students' responses to the SE tasks: Argument and Claim. See Appendix B for the coding scheme. At the Argument level we developed a coding scheme to identify the reasoning that students used when deciding which fact caused them to question the reliability of the source. The Claim level assesses each argument chain as a whole, including the claim of reliability and all arguments that led to the claim.

In order to develop the Argument level of the coding scheme, we first conducted an extensive literature review focusing on sourcing aspects such as the author's purpose, perspective, position and proximity (Barzilai & Zohar, 2012; Bråten et al., 2009; Britt & Aglinskas, 2002; Nokes, 2017; Nokes, Dole, & Hacker, 2007; Wineburg, 1991). We also considered each task and what type of an answer each fact might trigger. Based on this analysis, we added two aspects of historical reasoning: historical contextualization and corroboration (Van Drie & Van Boxtel, 2008; Wineburg, 1991). When contextualizing, students may consider when and where the source was written as well as the social conditions of the time period. Corroboration entails comparing the source with other sources to identify similarities and differences. In coding the transcripts we encountered arguments that did not fit into the categories derived from the literature, such as Exaggeration and Quantify Truth. Students who note exaggeration typically call attention to the effect it might have on the account whereas those who quantify truth may believe that an event that had not been wholly corroborated by other sources must be fabricated. As a result, we added several categories to the argument level so that each argument could be placed in only one category.

The Claim level was developed to describe the overall quality of the entire argument chain. Building on the work of Nokes (2017), this level of coding was used to account for the variety of quality in student work. For example, this level of coding discriminates between a student who employs historical reasoning at the argument level to develop a sound claim versus a student who twists historical reasoning at the argument level to further an incorrect or ahistorical claim. The categories were created mainly inductively after discussing the coding scheme, model answers and students' answers.

After determining the coding scheme, the first and second authors discussed differences in coding and clarified the coding scheme. Differences were primarily in determining surface versus elaborate claims, which were resolved by more clearly defining the two concepts. Two rounds of 40 and 50 randomly determined claims, 50% and 63% of the sample respectively, were coded. Agreement was established as .80 (Cohen's Kappa) during the second round of coding. The same authors also coded 36 arguments, 28% of the sample. After coding a second round of 65 randomly determined arguments, 51% of the sample, agreement was determined as .75 (Cohen's Kappa), and the remaining argument chains were coded by the first author.

Based on the coding scheme described above, we assessed each student's final answer to each oral SE task. A student's answer received one Claim score to describe the overall quality of the answer. Each different argument a student used to determine their final answer was also noted using the Argument part of the coding scheme. In total, each student's oral SE task received one Claim score and at least one Argument score.

Students' written answers were coded at the Claim level of the coding scheme. The first and second authors independently coded all written answers with agreement at .63 (Cohen's Kappa). Differences were resolved through discussion.

To identify language difficulties, we identified all language-related questions and evidence of language-related difficulties, particularly reading comprehension errors, in students' transcripts.

We analyzed the written and oral answers to each SE task to identify differences in the two modes. First, each students' written answer to the SE task was matched to the corresponding part of the think aloud transcript. Then the answers were compared to identify similarities and differences between the two modes. S1 and S2's SE2 were excluded from this analysis since their written answers were missing.

Students 7–11 DBQs were analyzed to determine if claims or arguments from the SE task also appeared in their writing. Their writing was analyzed because they completed all three SE tasks during the historical reasoning course. The analysis identified if historical sources were used in the DBQs, if reliability was considered in the DBQ and if so, how it compared to the oral and written answers of the SE task.

3. Results

3.1. Reasoning about reliability

To address RQ1 regarding students' ability to reason about reliability, in this section we discuss students' arguments and claims of reliability while thinking aloud based on the students' final answer to each SE. For this analysis we first identified each student's final answer to a given SE by consulting the written answer. Except in two cases noted in the methods section, this determination was used to identify the corresponding part of the think aloud which forms the basis of this analysis. See Table 4 for the claims and arguments made by students, and Appendix B for further examples of each category.

The vast majority of all claims in the think aloud transcripts were based on historical reasoning, but many were what we considered to be surface-level claims. In SE1, S6 questions the reliability of Suetonius because "there was a long time gap between the period he lived in and the period he wrote in, the period he lived in and the period he researched in, I believe that it is unreliable" (oral excerpt 3.1). This can be considered surface level because it does not explain why the time gap leads to a loss of reliability. In contrast, another student added that "because of that (time gap) he (Suetonius) can't see lively and it is maybe some sources can change after he died and he can't reach the success ones" (oral excerpt 3.2, S4, SEI). This additional reasoning about how the time gap potentially affected Suetonius' sources can be considered more elaborate.

When broken down into the stages Coffin (2006) uses to classify history genres, the most proficient answers appear to have three required stages: 1) an *orientation* of the author or source's background relevant to the other stages, 2) an *evaluation* that explains how the source or author's background may affect the reliability of the source, and 3) a *deduction* stage during which the student makes the assessment of reliability. Less proficient answers typically omit the evaluation stage. These stages do not follow a required order and are found in differing orders in students' responses.

With the exception of S1, no student had more than one ahistorical claim as a final answer. Ahistorical claims were based on the students' personal conception of history, such as the belief that an event that had not been wholly corroborated by other sources must therefore be fabricated (AA Quantify truth). These claims were not supported by historical evidence. Students made a few claims coded as Historical Claim Incorrect, two of which occurred during a fundamental language misunderstanding during SE3.

In general, students used arguments triggered by the facts in the SEs. Accordingly, S7 decides to question Augustus' account of his actions because "he was emperor so he couldn't be very reliable" (oral excerpt 3.3, SE2). However, additional arguments, notably Omitting Information and Exaggeration, were used although they were not triggered by the SE. S5 uses this approach in completing the same task: "I think Augustus asked the senate to write this text on the bronze pillars and put it put it up the Rome and the Roman empire because like the Augustus is the emperor and he might say do what I say…I don't think someone writes bad things about himself or her herself and put it all around the country" (oral excerpt 3.4, SE2). In this case, S5 shows a more proficient level of source evaluation by indicating how Augustus' position as emperor might lead him to be able to omit unflattering information in the source (HA Omitting Information). Students who used these arguments to reason about the background information wrote claims scored as Historical Claim Elaborate. This seems to indicate that students with the highest levels of performance developed claims containing interpretation based on relevant arguments.

3.2. Difficulties

This section addresses RQ2, and the role that reading comprehension-related language difficulties may have played in students' difficulties in historical reasoning. To understand why students were not able to formulate more elaborate claims, we conducted a further analysis of students' think aloud transcripts, which points to two potential problems.

Table 4
Claims and Arguments Used by 11 Students in the Oral SE Tasks (Percentages and Frequencies).

	SE1	SE2	SE3
Ahistorical claim	18% (2)	18% (2)	9% (1)
Historical claim incorrect	9% (1)	9% (1)	18% (2)
Historical claim surface	55% (6)	36% (4)	36% (4)
Historical claim elaborate	18% (2)	36% (4)	36% (4)
HA Corroboration	–	–	–
HA Proximity	25% (6)	–	7% (2)
HA Historical context	13% (3)	–	14% (4)
HA Position	8% (2)	15% (3)	3% (1)
HA Perspective	21% (5)	15% (3)	34% (10)
HA Omitting information	8% (2)	5% (1)	3% (1)
HA Exaggeration	4% (1)	15% (3)	–
HA Purpose	8% (2)	30% (6)	31% (9)
AA Personal opinion	13% (3)	10% (2)	3% (1)
AA Quantify truth	–	10% (2)	–
AA Logical fallacy	–	–	3% (1)

Note. Claims and arguments for each SE task based on students' final answers. HA = historical argument, AA = ahistorical argument. Each student made one claim and at least one argument.

3.2.1. Difficulties in contextualization

In all three SEs, some students failed to account for relevant historical context. At its most extreme, as in the case of S1 below, the interpretation is in direct contrast to stated historical context. In this example, S1 has misidentified both the author's audience as gladiators and the purpose as giving gladiators hope. When he eventually realized that his planned answer was illogical because it contradicted the information in the SE task, he seemed to question the veracity of the background information rather than revise his answer to account for the available evidence.

> Just now I have a doubt about the second [Juvenal's satires were probably intended for a rich male audience.] also. I said before Juvenal was intended to… no I will prefer second information. I'm changing because I don't want to be opposed to my previous argument. I said these texts were to give hope to gladiators the low rank people but male. But in second (fact) it says for a rich male audience. It won't be logical to write for rich male audience. And three. Yes, in this text just in this text it's criticizing women. Criticizing is not just you know negative. It can also be positive. So it's a positive criticizing (oral excerpt 3.5, S1, SE3).

In his final answer, S1 claims that Juvenal has written the Satire to "give hope to gladiators." This answer demonstrates a difficulty in historical contextualization because his reasoning is in direct contrast to what he has learned about the relative importance of rich males and gladiators in the Roman social hierarchy.

Other students overlooked historical context, leading to overgeneralizations or an interpretation unlikely given the context. In this case, students may have correctly identified features of the author, such as perspective. When the historical source was not placed into its historical context, however, the student twisted the interpretation to fit a more modern interpretation. In the following example, S8 questions the reliability of Juvenal because the student believes that the author is jealous that women are sexually attracted to gladiators, an unlikely interpretation for someone living in Juvenal's Rome.

> So, I would say third one [Juvenal wrote an entire satire criticizing women and their behavior.] makes me questioning the reliability of Juvenal…Because he wants to be a gladiator or may be that attraction from women. And, because they he criticize them, he couldn't attract them so he criticize women. And, he wrote Satire about women and their behavior (oral excerpt 3.6, S8, SE3).

This answer also demonstrates S8's difficulty in historical contextualization because while the interpretation is reasonable given a modern understanding of relationships between men and women, it is an unlikely interpretation for ancient Romans. In addition, S8 has not taken into account that the author has written a satire criticizing women for other behaviors. The resulting interpretation focusing on the author's jealousy of women and gladiators would not hold up as an explanation for what else the author had written.

3.2.2. Difficulties in formulating a complete answer

Another problem seen across SEs was an incomplete answer. Students in many cases were able to identify relevant features of the author's background, but did not explain why they questioned the reliability or how features of the author's background may have affected what the author wrote. One form of the problem included not directly answering the question. In the following excerpt, for example, S3 accurately notes the similarities in socioeconomic status between the author and the senator referenced in the text, and posits a purpose.

> I think it is the second one [Juvenal's satires were probably intended for a rich male audience.] 'cause I feel he is rich. He has power. He don't like to lose a man who has power lose his wife. It makes him afraid…Questioned reliability of Juvenal's account because he is a rich man and he has um same view with the senator. He may also afraid lose his wife or something for a gladiator (oral excerpt 3.7, SE3).

S3's answer above is incomplete, however, because the student does not explain her reasoning by alluding to why the features cause her to question the reliability nor speculate how the source might be affected.

In contrast, S6 (below) also points out relevant features of the author's background, notably the author's purpose, and explains how Juvenal's purpose might affect what he has omitted from his account.

> I think fact 3 [Juvenal wrote an entire satire criticizing women and their behavior.] cause me to question to the reliability of the Juvenal's account because his purpose is to criticize corruption in the Roman his..society. So he doesn't think the women's the people in the Roman society act what they do. He doesn't agree with what Romans doing what Roman society is doing. And he doesn't he clearly he doesn't agree with the idea that gladiators are desirable. So his purpose is to criticize this and that's why he's mentioning this like it's a very very bad thing and he's saying only the bad parts of it. So I think Juvenal is not reliable because of this (oral excerpt 3.8, SE3).

S6's answer above is complete because it includes an assessment of reliability, relevant background information and goes a critical step forward in also reasoning about how the author's purpose might affect what is written. This reasoning is generally the missing component in claims scored as a Historical Claim Surface.

3.3. The role of language

This section also addresses RQ2 by reporting on the role of language in students' comprehension of the historical sources based on the types of language-related questions and misunderstandings in the think aloud transcripts.

3.3.1. Language questions

All think aloud protocols were analyzed to identify student-specified language difficulties. Seven students made a total of 17 requests for definitions, translations, or a request for a word form during SE2 and SE3. Of these, S1 asked five of the questions during SE3. According to a profile of the requested words, approximately half were from the 2000 most frequently used words in English, such as 'improve,' and the other half were discipline-specific words not found on a word list, such as 'corruption' (Cobb, n.d.; Heatley, Nation, & Coxhead, 2002). One word, 'significance,' was from the academic word list. No word was requested twice, and a definition or translation was sufficient. In addition to these requests, one student commented twice on the difficulty of writing.

3.3.2. Language-related difficulties

During the analysis, we identified six transcripts featuring a fundamental misunderstanding of some major concept in the historical source that appear to be related to language, such as a misidentification of the audience or the author's purpose. In four cases the student was not aware of the misunderstanding and in two further cases the student was unable to successfully repair the misunderstanding, in all cases leading to a flawed interpretation.

Two misunderstandings took place in SE2 and four during SE3. This corresponds to a decrease in the amount of support from instructors, as noted in Table 2. Of the four students who had misunderstandings in SE3, three involved a misunderstanding of the intended audience.

S9 is an example of a case in which the student seems unaware of the misunderstanding. In this case, the student has misunderstood that the author, Juvenal, is praising the female character when in reality Juvenal is criticizing her.

> Okay here the writers says about the gladiator's wife. He praised him her very much. The question said that which of the three facts above might cause you question the reliability of this account. So not this one. I think author's satire were probably intending for a rich male audience is causes me to question the reliability because the writer talk about only the woman Eppia who is the senator's wife. In that time there weren't only rich male audience. There were also normal women. There were also citizens and woman praised her because she thinks her children and country. But every woman, but not only…It looks dumb…But like every woman…I can't seem to do this (oral excerpt 3.9, S9, SE3).

In the final part of the excerpt above, S9 appears confused at Juvenal's praise for this lone aristocratic woman, Eppia. As she does not realize her misunderstanding that Juvenal is actually criticizing Eppia's behavior, S9's answer is illogical. S9's answer shows how an inaccurate understanding of one aspect of the source can lead to a completely flawed interpretation.

Inaccurate comprehension of the historical sources may hinder students' ability to reason historically about the historical sources as these language-related difficulties seem to have been a factor or at least a co-occurrence in the two problems discussed in the previous section: discounting historical context and not formulating a complete answer. All of the students who had major comprehension errors in SE3 also had an error in at least one of the other two categories. Of the students without major comprehension errors, only two had errors in the other two categories.

3.4. Comparison of oral and written SE tasks

This section addresses RQ3 by reporting how students comment on reliability in oral and written modes. In this section, we first present a comparison of students' oral and written answers to the SE task. Next, we trace the use of the sourcing information from the SE task to students' DBQ essays written using the same historical sources.

Similar to the oral answers, students' written answers with a Historical Claim Elaborate score typically included an *orientation* stage with relevant background information about the source or author, a *deduction* stage with an assessment of reliability, and an *evaluation* stage with reasoning that explained a potential effect on the source or a detailed explanation of the assessment, as in the following written answer:

> His [Juvenal's] purpose is to criticize this behaviors and he is clearly not agree with the idea that gladiators were desirable. So he is only talking about the bad sides of this. That is why Juvenal is not reliable (written answer 3.10, S6, SE3).

In this example, S6 indicates Juvenal's purpose, reasoning about the potential effect of omitted information on Juvenal's account, and assesses reliability. In contrast, Historical Claim Surface scores typically omit the reasoning about how the author's background may affect the reliability of the source.

When comparing the oral and written scores, there was a difference in nine of the 31 scores. Five written scores were lower than the oral score. While not a significant number, it is somewhat concerning that four of these scores were of the only ten Historical Claim Elaborate scores given during the think aloud. In these four cases, the oral answer contains an explicit argument outlining how the information might be affected by the author's background, which is missing from the written answer. This omission is the primary reason that these written scores are lower than the oral scores.

One such instance occurs in S5's answer to SE3. In the following transcript, she speculated as to the author's actual motivation and the financial relationship between the author and a possible patron, whereas none of this speculation is included in the written answer.

Think Aloud Transcript (excerpt)

We really don't know who he is. Maybe he might needs some money and he might be writer and some rich men just says write this and I am gonna give you money. Maybe, I don't know. It's like this it's made for maybe for just for rich males and maybe like making woman feel bad about theirselves, like who is thinking about leaving the leaving their husband. So like it says you are leaving your children. You are leaving your country. You are leaving your husband or something. So I think it's for rich males for like staying to make their women stay with them. So it might not be reliable and I don't know. (oral excerpt 3.11, S5, SE3)

Written Answer (complete)

This might be written with a rich man's opinion to make their wives' feel guilty if they were thinking about prefer a gladiator. Saying that they were leaving their children and sisters behind might make them feel guilty. (written answer 3.12, S5, SE3)

The lack of this reasoning step resulted in a lower written than oral score for S5. S3 provides a further such example of a lower written score. In the oral answer, the student makes the point that Augustus might make "himself seems more better" because he wants to be remembered. While the purpose, being remembered, is present in the written answer, the reason she decides not to trust him, because he might make himself seem better, is omitted.

Think Aloud Transcript (excerpt)

But the third one Augustus asked the senate to write this text on bronze pillars and put it up in Rome, and around the Roman Empire. When I read it, I can see Augustus wanted to remembered by his people. He want to remembered by I'm with what he did. And I feel he has a little ego…So I feel it's not reliability enough…Augustus asked the senate to write this so from that I can understand he want to remembered by his people with what he did when he was alive. I think with that he want to say I made this Rome I made this city I and I made this Roman Empire. So he trying to say this so I feel he is talking about himself and his making himself seems more better so I think it is not reliable. (oral excerpt 3.13, S3, SE2)

Written Answer (complete)

Augustus asked the senate to write this so from that I can understand he wants to be remembered by his people with what he did when he was alive. He is trying to say he built the empire with those bronze pillars. So I don't think this text is reliability. (written answer 3.14, S3,SE2)

The other differences between written and oral scores were four written scores that were better than the oral scores. In three of these cases, ahistorical reasoning or inaccurate historical contextualization in the oral answer was omitted from the written answer.

3.5. The use of source evaluation in students' DBQ writing

As a part of RQ3, the DBQs of S7–S11 were analyzed to determine if aspects of their sourcing from the SEs also appeared in their DBQ writing. See Appendix C for a complete DBQ. All five students cited at least one historical source as evidence in each of the three DBQs. In most cases, the historical source from the SE task was used. Although students consistently used historical sources as evidence, there are only four total instances across all fifteen DBQ essays that include an assessment of an SE source. In two of these four cases, the student's assessment of reliability openly questions the reliability of the source used as evidence for the DBQ. For example, S7 questions the reliability of her source Juvenal because "he was not objective because of the exaggeration and he criticize only from one perspective (DBQ excerpt 3.15, SE3)." This negative assessment undermines her argument by calling into question the accuracy of one of her sources.

4. Discussion and conclusions

In this study, we examined L2 students reasoning about the reliability of historical sources during an EAP course. As expected based on RQ1, most students were able to make claims of reliability about historical sources when prompted by the SE task. While studies have been conducted with those proficient in English (Britt & Aglinskas, 2002; Wineburg, 1991), this study adds to the literature on how L2 students evaluate the reliability of historical sources and demonstrates that these undergraduate L2 students can assess the reliability of historical sources in topics outside of their planned area of study. It also shows, however, that their analysis is not well elaborated and is prone to misunderstandings.

Previous studies (Barzilai & Zohar, 2012; Britt & Aglinskas, 2002; Rouet et al., 1996) identified different types of arguments used by students to evaluate reliability. The present study adds the ahistorical reasoning-based arguments, logical fallacy and quantifying truth. Identifying students' ahistorical reasoning is important so that teachers can effectively predict and plan their instruction to challenge such misconceptions.

In response to RQ2, we expected students to have language-related difficulties when reading and assessing the reliability of historical sources. We found that some, but not all, students struggled with reading comprehension, especially as instructor support was removed. SE3, which was completed without support, appears to have presented the most challenge for students and may best illustrate what these students were able to complete independently. As a part of this analysis, we also identified two areas of difficulty that appear to co-occur with language-related difficulties: the misuse of contextualization and forming

a complete answer. The finding regarding contextualization is similar to other studies with L2 students in both written and oral production (Llinares & Whittaker, 2010; Myskow & Ono, 2018). Further studies comparing how students reason about sourcing in an L2 versus an L1 may shed further light on the extent of the effect of L2 proficiency on their reasoning.

RQ3 examines students' oral and written answers to the SEs. A comparison of students' oral and written answers offers some support for Felton and Herko's (2004) conclusion that students have difficulty translating oral language into a written answer. While a think aloud is not a dialogue, the oral nature of the task seems to have led students in our study to benefit in similar ways, resulting in richer claims in the think aloud than in the corresponding written answer, even when the score did not change. However, students' reasoning in this small scale study is generally superficial in both the oral and the written answer. A subsequent study in which more students score highly during the think aloud component may give a better indication of the extent that translation is an issue for these students.

This study also demonstrates that when writing DBQs, students were able to use the sources from the SEs. Few students, however, included an assessment of reliability in their writing, similar to studies by others (Monte-Sano, 2010; Nokes, 2017). Half of the assessments undermined the students' arguments by calling the reliability of a cited source into question. This is similar to that found in the writing patterns of lower scoring essays in the study by Miller et al. (2014). It is possible that since students are required to consider the reliability of the source in their DBQ and the SE focuses on negative aspects of reliability, that students include these assessments without realizing that they have damaged their own argument in the process.

Beyond the evaluation of historical sources, this study demonstrates that students need instruction to help them identify what aspects of their thought process to include in a written answer, as well as how to meaningfully incorporate sourcing into a DBQ. Teachers may, for example, provide specific criteria for what should be included in a successful answer. Based on the results of this study, that may include three stages: 1) an assessment of reliability, 2) pertinent background information about the author or source and 3) an explanation of how the background information has resulted in the assessment of reliability. While this level of explicit instruction may be useful for novices in general, it is particularly the case for those struggling with combining the complexity of historical reasoning, reading comprehension, writing and language in the context of an EAP course.

Teachers will need to carefully consider the goals of a course or activity when deciding on support levels for students, particularly those learning in a non-native language. If historical reasoning is a primary goal, then it will be important to continue to provide language support and check comprehension for students who struggle to read historical sources. Instructors can help students cope with reading comprehension of historical sources by pre-teaching vocabulary, selecting simple sources, and modifying difficult sources for vocabulary and structure (Nokes, 2011; Wineburg & Martin, 2009). If, in contrast, language drives the course, then instructors will need to consider how to ensure a reasonable interpretation of the historical sources. In language-focused EAP courses, teachers may also choose structures to highlight or explicitly teach. In this course, causative language to help explain the reasons for a student's evaluation or graded language to temper the student's claim may be appropriate in helping students produce an answer using the criteria described above and further develop their English for Academic Purposes.

Declaration of competing interest

None.

Appendix A

Source Evaluation Tasks

Source Evaluation 1

Directions: Use the text, source information, and your knowledge of history to answer the questions that follow.

> **Source:** "The Divine Augustus" was written by Suetonius. It is a biography that he wrote about the first Roman emperor, Augustus. He also wrote biographies of other Roman emperors.

Augustus stopped the disorderly and disrespectful way of viewing games by passing special laws. He did this because he was angry that people insulted a senator by not offering him a seat at the games in Puteoli. As a result of this insult, the senate decreed that the first row of seats should be reserved for senators at all public shows. In Rome, Augustus would not allow the representatives of the free and allied nations (parts of the Roman Empire) to sit in the front because he was told that even former slaves could become representatives. He separated soldiers from civilians. He assigned separate seats to the married common men. He assigned a section of seats to boys and assigned the seats nearby to their tutors. And he decreed that no one wearing a dark cloak should sit in the middle (they had to sit in the upper seats). He would not allow women to view the gladiators, except from the upper seats (Suetonius, Augustus 44, trans.1998).

Additional facts related to Suetonius:

1. Suetonius lived between approximately 71-135 CE. The emperor Augustus died in 14 CE.
2. Suetonius was in charge of Roman libraries and archives for Emperor Hadrian, and had access to letters that Emperor Augustus wrote.
3. Suetonius was from a wealthy Roman equestrian family. Equestrians were generally not involved in politics.

Question: Which 1 of the 3 facts above might cause you to question the reliability of Suetonius' account?

Source Evaluation 2

Directions: Use the text, source information, and your knowledge of history to answer the questions that follow.

Source: "Deeds of the Divine Augustus" was written by the first Roman emperor, Augustus. In this text he lists his achievements and the money he spent on Rome and Romans.

Three times I gave shows of gladiators under my name and five times under the name of my sons and grandsons; in these shows about 10,000 men fought. Twice I furnished under my name spectacles of athletes gathered from everywhere, and three times under my grandson's name. I celebrated games under my name four times, and furthermore in the place of other magistrates twenty-three times. As master of the college I celebrated the secular games for the college of the Fifteen, with my colleague Marcus Agrippa, when Gaius Furnius and Gaius Silanus were consuls. Consul for the thirteenth time, I celebrated the first games of Mars. After that the senate made a decree and a law saying that consuls should celebrate the games of Mars (Augustus, Deeds of the Divine Augustus 22, trans. 1998).

Additional facts related to Augustus:

1. Augustus wrote new rules about where people could sit during gladiator shows.
2. Other primary sources attest (also say) that at least 7 of these 8 gladiator shows took place.
3. Augustus asked the senate to write this text on bronze pillars and put it up in Rome and around the Roman Empire.

Question: Which 1 of the 3 facts above might cause you to question the reliability of Augustus' account?

Source Evaluation 3

Directions: Use the text, source information, and your knowledge of history to answer the questions that follow.

Source: "Satire" was written by Juvenal. Juvenal used satire to criticize corruption in Roman society and the behaviors of people he didn't agree with.

What beauty set Eppia (a senator's wife) on fire? What youth captured her? What did she see that made her endure being called a gladiator's woman? For her darling Sergius (the gladiator) had already begun to shave (because he was middle aged), and to hope for retirement soon because of a wounded arm. Moreover, there were many deformities on his face; for instance there was a huge wart on the middle of his nose, which was rubbed by his helmet, and a bitter liquid dripped continually from one eye. But he was a gladiator…She preferred this to her children and her country. That woman preferred this to her sister and her husband. The sword is what they love (Juvenal, Satire 6.102–112, as cited in Futrell, 2006).

Additional facts related to Juvenal:

1. Juvenal lived between approximately 60-127 CE.
2. Juvenal's satires were probably intended for a rich male audience.
3. Juvenal wrote an entire satire criticizing women and their behavior.

Question: Which 1 of the 3 facts above might cause you to question the reliability of Juvenal's account?

Appendix B

Coding Scheme

Claim	Example
Code as **Uncertainty** when the student declines to or cannot make a claim of reliability because of some uncertainty.	"We don't have the time that Satire was written. So first one is…I'm not very sure." (S2, SE3)
Code as **Claim** when the student makes a claim of reliability regarding the source or author, but doesn't support the claim with any arguments. In this context, a claim is 1) any statement in which the student determines whether the given primary source is reliable or not reliable based on one or more 'additional facts' read in the source evaluation task or 2) any statement in which the student determines whether an 'additional fact' in the source evaluation task causes the student to question the reliability of the primary source. The claim must either specifically refer to reliability or can be reasonably inferred to refer to reliability.	"I think first one is not unreliable too much." (S3, SE3)
Code as **Ahistorical Claim (AC)** when the student makes a claim regarding the reliability of the source or author based on ahistorical arguments.	"Other primary sources also say that at least seven these gladiator shows took place. Majority thinks that it have to most of this seven. It's out of eight so it's nearly 95% true." (S2, SE2)
Code as **Historical Claim Incorrect (HCI)** when the student makes a claim of reliability regarding the source or author and bases the claim of reliability on historical reasoning, but at least half of the basis is incorrect or the claim is illogical. A basis is incorrect if available evidence provided to the students contradicts it.	"It's not (fact) one. He (Augustus) wrote new…We don't know his new rules. " (S2, SE2)
Code as **Historical Claim Surface (HCS)** when the student makes a claim of reliability regarding the source or author and the claim is reasonable, but is not elaborated upon, the elaboration is shallow, and/or partly incorrect.	"Augustus asked the senate to write these texts on bronze pillars and put it up in Rome and around the roman empire. I think this is unreliable because he want to show his power to the public. And he wrote his text on bronze pillars because he wanted to show everyone his text. And he also put it up Rome and around the Roman Empire be it is also he want to show everyone because he put it in public place." (S10, SE2)
Code as **Historical Claim Elaborate(HCE)** when the student makes a claim of reliability regarding the source or author and the claim is reasonable, correct and well elaborated. A well-elaborated claim contains substantive details explaining or speculating about the claim, author or source.	"I think fact 3 eh cause me to question the reliability of the Juvenal's account because his purpose is to criticize corruption in the Roman society. So he doesn't think the women's the people in the Roman society act, what they do. He doesn't agree with what Romans doing and clearly he doesn't agree with the idea that gladiators are desirable. So his purpose is to criticize this and that's why he's mentioning this like it's a very very bad thing and he's saying only the bad parts of it. So I think Juvenal is not reliable because of this." (S6, SE3)

Arguments	Example
Code as **Historical Argument Corroboration** when the student justifies a claim based on corroborating information in multiple sources.	"I think this is reliable too because the text and other primary sources say the same thing." (S10, SE2)
Code as **Historical Argument Purpose** when the student justifies a claim based on the author's stated or implied purpose and/or the author's stated or potential audience.	"I think this is unreliable because he want to show his power to the public." (S10, SE2)
Code as **Historical Argument Position** when the student justifies a claim based on the author's "occupation, profession or credentials" or access to information due to the author's position.	"He was in charge of Roma libraries…This is reliable because he had access to his (Augustus') letters."
Code as **Historical Argument Perspective** when the student justifies a claim based on the author's perspective, including the author's socioeconomic class, gender, or view of others.	"So he doesn't think the women's the people in the Roman society act, what they do. He doesn't agree with what Romans doing and clearly he doesn't agree with the idea that gladiators are desirable." (S6, SE3)
Code as **Historical Argument Proximity** when the student justifies a claim based on the author's temporal or geographical proximity to the topic.	"According to it the writer lived in a period much later than Augustus, so these writings were written after an 80–90 year period. So if you ask me this gap in periods might have affected the reliability of the writings." (S3, SE1)
Code as **Historical Argument Historical Context** when the student justifies a claim based on the period's temporal, spatial or social context.	"In that time there weren't only rich male audience, there were also normal women." (S9, SE3)
Code as **Historical Argument Exaggeration** when the student justifies a claim based on exaggeration in the source. The argument may be specified or speculative.	"And in the text he talk about how he is rich and powerful. He praise himself and I am guessing he probably add extra information about himself in text." (S8, SE2)
Code as **Historical Argument Omitting Information** when the student justifies a claim based on intentionally omitted information from the source. The argument may be specified or speculative.	"…and he's saying only the bad parts of it." (S6, SE3)
Code as **Ahistorical Argument Quantify truth** when the student justifies a claim based on an amount of truth.	"It's out of eight so it's nearly 95% true." (S2, SE2)
Code as **Ahistorical Argument Personal Opinion** when the student justifies a claim based on the student's own opinion which is not supported by the available evidence.	"All politicians are liars." (S6, SE2)
Code as **Ahistorical Argument Logical Fallacy** when the student justifies a claim based on a previously determined argument. The claim is not altered when faced with conflicting information.	"I'm changing because I don't want to be opposed to my previous argument. " (S1, SE3)

Appendix C

Sample DBQ3 from S7

In the Roman society, there were various views about gladiators. The society give very attention to shows. They seem as a career of chance to get fame and wealth. However the society seem them as the lowest status. Dunkle argues that being a gladiator could seem as a attractive career. Gladiators would fight 2 or 3 times per a year and also, they would have some opportunities to getting a fame and wealth. Chances that they gain with this career ensure to buy their freedom. In addition, the volunteer gladiators could want to get military glory and achieve the adoration of public (Dunkle). Besides the public adoration, in the Satire of Juvenal describes female adoration to gladiators. Career as a gladiator seem womens in Roman society preferred them to their aristocrat or politician husbands. The reason for this statement, women being love to their swords and wounds. Juvenal's Satire is unreliable because he was not objective because of his exaggerations. He criticizes only from one perspective. In the text graffiti by anonymous gave an example Eppia who was the senator's wife to this.

Second view from Roman society, as a career gladiators seem as slave and prisoners of war (Dunkle). The legal status of them seem as the lowest status in Roman society both the Empire and Republic. They had no citizen rights in society differ from free man. In the text written by Galen, being a gladiator explained as an unhealthy career. Describes that gladiators seem as stupid as animal. Due to their wounds they lost their minds and their motor system. Gladiators have no condition, also they lose their feeling in their body. Because of their deformation, the limbs that they have become dislocated (Galen). In the text written by Galen seem as reliable because of he lived during the late second century CE. In addition to his reliability, he was a doctor who worked for a gladiator school. Therefore he was access to information completely. However he tried to convince people to not become a gladiator. Therefore, this source was not objective.

References

Augustus. (1998). *The deeds of the divine Augustus* (T. Bushnell, Trans.). Retrieved from http://classics.mit.edu/Augustus/deeds.html.

Barzilai, S., & Zohar, A. (2012). Epistemic thinking in action: Evaluating and integrating online sources. *Cognition and Instruction, 30*(1), 39-85.

Bråten, I., Strømsø, H. I., & Britt, M. A. (2009). Trust matters: Examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly, 44*(1), 6-28.

Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction, 20*(4), 485-522. Retrieved from http://www.jstor.org/stable/3233902.

Cobb, T. (n.d.). Web vocabprofile. Retrieved from http://www.lextutor.ca/vp/.

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

Cohen, D. J. (2005). By the book: Assessing the place of textbooks in U.S. survey courses. *Journal of American History, 91*(4), 1405-1415. https://doi.org/10.2307/3660181.

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

De La Paz, S. (2005). Effects of historical reasoning instruction and writing strategy mastery in culturally and academically diverse middle school classrooms. *Journal of Educational Psychology, 97*(2), 139.

Felton, M. K., & Herko, S. (2004). From dialogue to two-sided argument: Scaffolding adolescents' persuasive writing. *Journal of Adolescent & Adult Literacy, 47*(8), 672-683. Retrieved from http://www.jstor.org/stable/40016901.

Futrell, A. (2006). *The roman games: A sourcebook*. Hoboken, New Jersey: 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.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *RANGE and FREQUENCY programs*. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx.

Lee, P., & Ashby, R. (2000). Progression in historical understanding among students ages 7-14. In P. N. Stearns, P. Seixas, & S. Wineburg (Eds.), *Knowing, teaching, and learning history* (pp. 199-222). New York: New York University Press.

Llinares, A., & Whittaker, R. (2007). Talking and writing in a foreign Language In CLIL contexts: A linguistic analysis of secondary school learners of geography and history. *Revista Espanola de Linguistica Aplicada, 20*, 83-91.

Llinares, A., & Whittaker, R. (2010). Writing and speaking in the history class. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms* (pp. 125-144). Amsterdam: John Benjamins Publishing Company.

Lo, Y. Y. (2014). L2 learning opportunities in different academic subjects in content-based instruction – evidence in favour of 'conventional wisdom'. *Language and Education, 28*(2), 141-160. https://doi.org/10.1080/09500782.2013.786086.

Loban, W. (1976). *Language development: Kindergarten through grade twelve* (vol. 18)Urbana, Illinois: National Council of Teachers.

Lorenzo, F. (2017). Historical literacy in bilingual settings: Cognitive academic language in CLIL history narratives. *Linguistics and Education, 37*, 32-41. https://doi.org/10.1016/j.linged.2016.11.002.

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

McGrew, S., Ortega, T., Breakstone, J., & Wineburg, S. (2017). The challenge that's bigger than fake news: Teaching students to engage in civic online reasoning. *American Educator, 41*(3), 4-10.

Miller, R. T., Mitchell, T. D., & Pessoa, S. (2014). Valued voices: Students' use of engagement in argumentative history writing. *Linguistics and Education, 28*, 107-120. https://doi.org/10.1016/j.linged.2014.10.002.

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.

Mitchell, T. D., & Pessoa, S. (2017). Scaffolding the writing development of the argument genre in history: The case of two novice writers. *Journal of English for Academic Purposes, 30*, 26-37. https://doi.org/10.1016/j.jeap.2017.10.002.

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. Retrieved from http://www.jstor.org/stable/27667162.

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.

Myskow, G. (2018). Changes in attitude: Evaluative language in secondary school and university history textbooks. *Linguistics and Education, 43*, 53-63. https://doi.org/10.1016/j.linged.2017.12.001.

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. (2011). Recognizing and addressing the barriers to adolescents' "reading like historians". *The History Teacher, 44*(3), 379-404. Retrieved from http://www.jstor.org/stable/41303991.
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., 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.
Reisman, A. (2012). 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.
Rouet, J.-F., Britt, M. A., Mason, R. A., & Perfetti, C. A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology, 88*(3), 478.
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.
Smith, M., Breakstone, J., & Wineburg, S. (2019). History assessments of thinking: A validity study. *Cognition and Instruction, 37*(1), 118-144. https://doi.org/10.1080/07370008.2018.1499646.
Suetonius. (1998). *The lives of the Caesars* (J. Rolfe, Trans.). Cambridge, MA: Harvard University Press.
Thompson, C., Morton, J., & Storch, N. (2013). Where from, who, why and how? A study of the use of sources by first year L2 university students. *Journal of English for Academic Purposes, 12*(2), 99-109. https://doi.org/10.1016/j.jeap.2012.11.004.
Unsworth, L. (1999). Developing critical understanding of the specialised language of school science and history texts: A functional grammatical perspective. Retrieved from *Journal of Adolescent & Adult Literacy, 42*(7), 508-521 http://www.jstor.org/stable/40015633.
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 Someren, M., Barnard, Y., & Sandberg, J. (1994). *The think aloud method: A practical approach to modelling cognitive.* London: Academic Press.
Walraven, A., Brand-Gruwel, S., & Boshuizen, H. P. (2009). How students evaluate information and sources when searching the World Wide Web for information. *Computers & Education, 52*(1), 234-246.
Wette, R. (2017). Source text use by undergraduate post-novice L2 writers in disciplinary assignments: Progress and ongoing challenges. *Journal of Second Language Writing, 37*, 46-58. https://doi.org/10.1016/j.jslw.2017.05.015.
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., & 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.
Yildirim, A. (2006). High school textbooks in Turkey from teachers' and students' perspectives: The case of history textbooks. *Asia Pacific Education Review, 7*(2), 218-228. https://doi.org/10.1007/bf03031545.

Kristin A. Sendur is a PhD candidate at the University of Amsterdam and a staff member at the Center for Individual and Academic Development at Sabanci University. Her research interests include the learning and teaching of history, with a focus on L2 students.

Carla van Boxtel is professor of History Education at the Research Institute of Child Development and Education and the Amsterdam School of Historical Studies of his of the University of Amsterdam. Her research interests are the learning and teaching of history, particularly how to improve students' historical reasoning.

Jannet van Drie is associate professor at the Research Institute of Child Development and Education of the University of Amsterdam. Her research focuses on the learning and teaching of history, with a particular focus on historical reasoning and writing.