History Education Research Journal

History Education Research Journal

UCLPRESS

Research article

Exploring differential effects of an intervention on historical inquiry tasks: a qualitative analysis of 12th-grade students' progress

Marjolein Wilke,[1,*] Fien Depaepe,[2] Karel Van Nieuwenhuyse[1]

[1] Faculty of Arts, KU Leuven, Leuven, Belgium
[2] Center for Instructional Psychology and Technology, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium; ITEC, imec research group at KU Leuven, Leuven, Belgium
* Correspondence: marjolein.wilke@kuleuven.be; marjoleinwilke@gmail.com

Submission date: 8 August 2022; Acceptance date: 14 July 2023; Publication date: 24 August 2023

How to cite
Wilke, M., Depaepe, F. and Van Nieuwenhuyse, K. (2023) 'Exploring differential effects of an intervention on historical inquiry tasks: a qualitative analysis of 12th-grade students' progress'. *History Education Research Journal*, 20 (1), 5. DOI: https://doi.org/10.14324/HERJ.20.1.05.

Peer review
This article has been peer-reviewed through the journal's standard double-anonymous peer-review process, where both the reviewers and authors are anonymised during review.

Copyright
2023, Marjolein Wilke, Fien Depaepe and Karel Van Nieuwenhuyse. This is an open-access article distributed under the terms of the Creative Commons Attribution Licence (CC BY) 4.0 https://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited • DOI: https://doi.org/10.14324/HERJ.20.1.05.

Open access
History Education Research Journal is a peer-reviewed open-access journal.

Abstract

Multiple-documents-based (inquiry) tasks are often used to examine historical thinking, as they require students to apply discipline-specific ways of reasoning and writing. Intervention studies using such tasks have often relied on principles from cognitive apprenticeship to make these discipline-specific heuristics explicit to students. While several studies have found positive results, they offer little insight into how and where exactly students' progress on historical thinking manifests itself, nor into the differential effects of the intervention. Building on essay data gathered during an intervention study on students' historical inquiry skills, this study explores differential effects of the intervention according to students' initial historical inquiry ability. To this end, a purposeful sample of students was selected for whom the intervention was particularly effective. The qualitative analysis of students' essay tasks (pretest and posttest) revealed

remarkable differences between students with high and low pretest scores. Although both groups made progress on all aspects of the essay task, they differed in terms of where and how this progress manifested itself: at posttest, students with a high initial score outperformed others in evaluating sources and rebuttals. This study offers insight into patterns of progress in students' historical inquiry skills which can inform differentiation in instructional practices.

Keywords history education; secondary education; historical thinking; historical inquiry; qualitative analysis; multiple-documents-based tasks; intra-class differentiation

Introduction

History education across several Western countries has undergone significant changes, from a subject focused on the teaching of particular (national) narratives to one with a much broader scope. One purpose, initiated predominantly in German literature, was the development of students' historical consciousness. In this view, which gradually spread across continental Europe, history education can support students in orienting themselves in life and in developing their identity. It allows them to make sense of the past, and to guide their actions in the present, and it offers them perspectives for the future (Rüsen, 2004). In the anglophone community, as well in several European countries, historical thinking gradually became a central goal for history education (Seixas, 2017). Historical thinking focuses on 'the way we go about doing history' (Lee and Ashby, 2000: 199; Lévesque and Clark, 2018). With the introduction of historical thinking into history curricula, students not only acquire substantive knowledge about the past, but also second-order and procedural knowledge that introduces them to history as an academic discipline, allowing them to gain insight into the way that historical knowledge is constructed (Lee, 2004; Van Drie and Van Boxtel, 2008; VanSledright and Limón, 2006).

Within this latter framework, historical thinking is considered to be a complex and 'unnatural' act (Wineburg, 2001), as it is opposite to how students spontaneously tend to look at and engage with the past. Several (intervention) studies have therefore examined effective strategies to foster aspects of students' historical thinking, such as causal reasoning (Stoel et al., 2017), historical contextualisation and perspective-taking (Huijgen et al., 2018), critical source analysis (Nokes et al., 2007; Reisman, 2012) and historical writing (De La Paz, 2005; De La Paz and Felton, 2010; De La Paz et al., 2016; Van Drie et al., 2015). These studies often rely on multiple-documents-based (inquiry) tasks and adopt principles such as explicit teaching, modelling and guided practice supported by scaffolds, in line with the model of cognitive apprenticeship (Collins et al., 1991).

In these studies, multiple-documents-based (inquiry) tasks are frequently used during the intervention, and as part of the measurement instrument, to examine changes in students' historical thinking. These inquiry tasks are usually evaluated using rubrics assessing various aspects of students' historical reasoning and writing (for example, De La Paz et al., 2016; Sendur et al., 2021; Stoel et al., 2017). Aside from a few exceptions (for example, Stoel, 2017), intervention studies typically do not include a detailed analysis of students' progress on the various aspects of these rubrics. In most cases, there is hence no in-depth information on the precise nature of students' progress on the multiple-documents-based (inquiry) tasks. Moreover, these studies usually report on the student sample as a whole, and do not address (possible) differential effects of the intervention among students. Thus, while these studies provide important information about effective instructional practices to foster progress in students' historical thinking, they do not offer as much information about how exactly such progress manifests itself (differently) in students' underlying thinking processes. These insights, however, are equally valuable and necessary for informing effective teaching practices.

To overcome this gap, this study builds on data collected in a previous intervention study (Wilke et al., 2022) to explore the effects of an intervention on students' historical thinking in an in-depth way. Our intervention study was a cluster randomised control intervention study with 628 students (intervention: $n = 402$; control: $n = 226$). The study was based on a lesson series taking up 12–14 history classes aimed specifically at fostering students' historical inquiry competences, that is, their ability to generate a substantiated answer to a historical question, based on a critical analysis of sources. The

https://doi.org/10.14324/HERJ.20.1.05

lesson series was based on inquiry tasks and design principles from cognitive apprenticeship (Collins et al., 1991). We examined the effects of the lesson series on students' historical inquiry skills via a historical essay task. A first remarkable observation was that students differed significantly in their initial starting points for the historical essay tasks. A second was that, although the intervention was in general successful in improving students' historical inquiry skills, students did not progress equally: significant differences remained in place at the posttest level. Affective variables, procedural knowledge and epistemological beliefs did not provide a sufficient explanation for these large differences in students' scores. The data from this study provide an excellent opportunity to obtain more insight into how precisely students' progress manifested itself in these historical inquiry tasks, and to examine whether this progress differed among students according to their initial scores.

Via a qualitative study of a purposeful selection of students from the intervention group, this study therefore aims to explore students' progress on historical skills in a detailed way, considering differences in students' initial scores. That way, the study aspires not only to gain more insight into (the precise shape of) students' progress on historical skills, but also to reflect on practical implications for the classroom context, for instance, with regard to differentiation.

Theoretical framework

Multiple-documents-based (inquiry) tasks

Although there is no uniform operationalisation of historical thinking, this research builds on influential models situated in the anglophone-inspired research tradition (for example, Seixas and Morton, 2013; Van Drie and Van Boxtel, 2008; Wineburg, 2001). In these models, historical thinking aims, through an introduction of students to the methods of historians, to allow students to understand the interpretive and constructed nature of historical knowledge. Core aspects of historical thinking include asking historical questions, constructing and applying a historical frame of reference, critical source analysis, applying historical modes of reasoning (such as agency, causality, continuity and change), and constructing or critically analysing historical representations.

Multiple-documents-based (inquiry) tasks are often used to foster and assess students' historical thinking. These tasks combine various aspects of historical thinking as they require students to critically evaluate and compare different sources, and to provide arguments and evidence in support of a certain claim (De La Paz and Felton, 2010; Monte-Sano, 2010, 2016; Van Drie and Van Boxtel, 2008; Van Nieuwenhuyse, 2020; Voet and De Wever, 2017). These tasks are considered authentic tasks for the discipline of history (Van Boxtel and Van Drie, 2018).

Providing a substantiated historical claim based on multiple documents requires students to do far more than synthesise the information in the different sources. Considering that not all historical sources are equally valuable, students need to apply discipline-specific heuristics or procedural knowledge to evaluate and compare them (Rouet et al., 1996). Wineburg's (1991) influential expert–novice study identified three key heuristics applied by expert historians when dealing with multiple historical documents. Source corroboration refers to comparing (differing accounts in) sources, looking for ways in which the sources confirm or contradict each other (Britt and Aglinskas, 2002; De La Paz and Felton, 2010; Wineburg, 1991). In order to weigh the evidence in (conflicting) sources against each other, experts also applied 'sourcing', that is, examining the author of the source, as well as when and where it was created (Wineburg, 1991). A final heuristic, 'contextualisation', refers to the need to pay attention to when and where the historical document originated (Wineburg, 1991). Over time, several scholars have elaborated upon these heuristics, adding, for instance, 'close reading', referring to the need to pay attention to aspects such as word choice and language in a document (Reisman, 2012). Scholars have also elaborated upon the heuristic of sourcing with additional criteria for assessing a source's value, such as the documents and information on which a source is based, the goal and audience of a source, and the need to evaluate a source's representativeness (for example, Britt and Aglinskas, 2002; De La Paz and Felton, 2010; Hicks et al., 2004; Perfetti et al., 1994; Van Drie and Van Boxtel, 2008; Van Nieuwenhuyse, 2020; Wilschut et al., 2012).

Scholars have argued that discipline-specific strategies are not only at play in the analysis of multiple documents, but also when formulating a substantiated answer, which is often done through writing. Disciplinary writing practices include presenting arguments, using evidence to support a claim, and rebutting opposing evidence or claims (De La Paz and Felton, 2010; Monte-Sano, 2010; Nokes, 2017;

https://doi.org/10.14324/HERJ.20.1.05

Nokes and De La Paz, 2018). Monte-Sano (2010) identifies five disciplinary characteristics in students' historical writing based on a multiple-documents-based task. First, factual and interpretive accuracy of the events and information described in the documents. Second, persuasiveness of evidence, referring to the need to substantiate historical claims with relevant and convincing evidence based on the sources. Third, sourcing of evidence, indicating that the essay contains references to the author of a document and the bias inherent in the source. Fourth, corroboration of evidence, meaning that the essay pays attention to the way that different sources work together to support a certain claim but also address counterevidence. Fifth, contextualisation of evidence, referring to the inclusion of contextual knowledge to evaluate the documents and to the essay's use of sources in a way that respects the contemporary meaning of the source. Monte-Sano (2010) demonstrates that while these characteristics will remain distinct among students with lower historical writing abilities, they will overlap when students become more proficient.

Students' performance on written inquiry tasks is related to their knowledge of historical text structures. Regarding causal historical writing, Coffin (2004, 2006), for instance, makes a distinction between accounts based on narrative (that is, recording genres) and accounts based on analysis (that is, explanatory and arguing genres). The former have a narrative structure, present a factual account and contain little argumentation or author voice. The latter are more analytical and tend to reflect the interpretive nature of history; for instance, by providing arguments, comparing various causes and drawing evaluative conclusions. Building on Coffin's (2004, 2006) distinction in causal writing genres, Stoel (2017) establishes a relationship between students' essay structure and their historical reasoning reflected in an inquiry task. In an intervention study on students' causal historical writing, he demonstrates that students' initial essay structure mediated the kinds of revisions that students made in their essays after the intervention. The intervention, however, had no effects on the structure of students' essays.

Students' historical thinking

Studies using (multiple) documents-based tasks have demonstrated that, spontaneously, students struggle with the application of discipline-specific ways of reasoning and writing. As students tend to approach sources as carriers of information, they are not inclined to corroborate or critically evaluate them (Britt and Aglinskas, 2002; Rouet et al., 1996; Van der Eem et al., 2022; Wineburg, 1991). When students do evaluate the value of a source, they often do so in a superficial or inaccurate way; for instance, by relying on the contents of a source, rather than on the contextual information accompanying it. Harris et al. (2016: 114) found that students frequently relied on the content of a source to establish whether it was reliable; for instance, referring to whether a source connected to students' prior knowledge or experience, whether it contained facts or statistics, or even whether it 'looked official' or 'sounded true'. When confronted with writing tasks based on multiple sources, students tend to focus on summarising the content of sources, rather than on providing a substantiated claim (for example, Nokes, 2017; Stahl et al., 1996; Young and Leinhardt, 1998). Students' historical writing is thus more likely to take the form of a 'factual' historical account (Van Boxtel et al., 2021).

Although these studies have convincingly demonstrated that applying discipline-specific heuristics is difficult for students, they have also shown that students' performance on these multiple-documents-based tasks can differ strongly (for example, Monte-Sano, 2010). Nokes (2017), for instance, described patterns in 8th graders' (12 or 13 years old) sourcing and argumentative writing, demonstrating the wide variety of students' performance in this regard. Considering this variation, he argues that teachers must adapt the design of assessments for students' historical thinking to students' incoming abilities.

Fostering students' historical thinking

In order to foster students' ability to provide a substantiated historical representation, intervention studies have successfully relied on design principles which aim to make disciplinary ways of reasoning visible to students, and to provide them with support while applying them. In particular, studies have adopted design principles such as explicit teaching on discipline-specific heuristics, modelling, guided practice supported by scaffolding, feedback and (whole-class and peer-to-peer) interaction (Nokes and De La Paz, 2018; Van Boxtel et al., 2021), in line with the general educational model of cognitive

https://doi.org/10.14324/HERJ.20.1.05

apprenticeship (Collins et al., 1991). De La Paz and Felton (2010), for instance, found that students who had received instruction on analysing sources, and on writing an argumentative essay, created essays that were longer, were of higher quality, contained more advanced claims and rebuttals, and cited more sources. Reisman (2012) studied the effects of document-based lessons, including explicit teaching on disciplinary strategies and support for students through a process of modelling, guided and independent practice. This approach yielded positive effects on students' evaluation of single sources, yet not on intertextual strategies, such as source corroboration. In explaining these results, Reisman (2012) pointed to the way that these intertextual strategies were approached during the intervention, but also hypothesised that perhaps these strategies are more sophisticated, as they require a connection between several documents. Positive results of the instructional practices described above have been found across various studies (for example, De La Paz, 2005; De La Paz et al., 2016; Huijgen et al., 2018; Monte-Sano and De La Paz, 2012; Nokes et al., 2007; Reisman, 2012; Stoel et al., 2017). Several of them have demonstrated that these positive results apply to students regardless of their incoming abilities (for example, De La Paz and Felton, 2010; De La Paz et al., 2016; Reisman, 2012). These studies, however, provide little insight into how precisely students progress on these inquiry tasks, and how the intervention affected students (differently) according to their incoming abilities.

In our own intervention study (Wilke et al., 2022), we similarly established that a lesson series based on inquiry tasks and design principles derived from the model of cognitive apprenticeship was effective in improving students' ability to provide a substantiated answer to a historical question. However, we noted significant differences in students' initial level of historical inquiry skills, and we found that the intervention did not yield the same effect on all students. Based on existing studies, as well as on our own study, a number of questions remain unanswered. First, how precisely do students progress in terms of historical thinking. Do students advance evenly across various aspects of historical thinking, or are some aspects of historical thinking easier to acquire? Second, what is the role of students' initial historical thinking ability in this regard? For instance, does students' initial ability influence how the progress on historical thinking takes shape? Overall, little is known about how interventions aimed at fostering students' historical thinking affect students on an individual level.

The current study

Building further on the historical essay tasks gathered in our previous (intervention) study (Wilke et al., 2022), this study explores the progress of historical skills of a specific group of students on an individual level. Following the operationalisation of the intervention study on which this follow-up study is based, and in line with curricular requirements in Flanders (Belgium), this study defines historical inquiry skills as 'the ability to construct substantiated historical representations, based on a critical analysis of multiple sources and taking into account multiple perspectives' (Wilke et al., 2022: 103).

In particular, this study examines how precisely students' progress on historical inquiry skills takes shape among those students who progressed substantially, and how this progress differs depending on students' starting position. To this end, a purposeful sample of students is selected from the intervention group; and their essay tasks are analysed qualitatively. For this group of students, the following twofold research question is explored: (1) What does progress on students' historical inquiry skills look like among students who have progressed considerably; and (2) How does it differ depending on students' starting positions? Students' learning progression is thereby defined as the change in score between their pretest and posttest essay task. Those essays are compared in order to describe changes occurring in students' historical inquiry skills, in the context of an intervention.

Method

Research context and participants

This study was conducted in the 12th grade of general secondary education in Flanders. History education is a mandatory subject taking up two hours a week. In the 12th grade, history teachers have both a master's degree (most often in history) and a teaching degree. However, their familiarity with (teaching) historical thinking varies as there is no mandatory professionalisation, and attention to this concept on teacher training courses has only become more prominent in recent years. Moreover, Flemish history standards provide teachers with a lot of freedom and only mention historical thinking in an

implicit way, providing no concrete guidelines on how to address it in practice (Van Nieuwenhuyse, 2020). This particular context has created a large diversity in teachers' practices regarding historical thinking. Overall, however, these practices are characterised by a focus on substantive knowledge about the past and only limited attention to second-order and procedural knowledge. This is most clearly demonstrated by teachers' use of sources in an illustrative way or to obtain substantive knowledge (Van Nieuwenhuyse et al., 2017). Even when teachers evaluate sources, they regularly do so in a superficial way, and they occasionally depart (in line with history textbooks) from a simplistic distinction between objective and subjective sources (Wilke and Depaepe, 2019). In general, teachers are also not inclined to organise full historical inquiries based on multiple sources in their practice, and they have a limited understanding of what these entail (Voet and De Wever, 2016, 2017).

Between 2019 and 2025, a curriculum reform is gradually being implemented in Flemish history education which puts forward historical thinking as a central goal. In this context, 40 teachers in 32 schools were recruited to participate in the intervention study. Schools were assigned randomly to a control (n = 226) or an intervention group (n = 402). The intervention study was conducted in the 12th (last) stage of secondary education, in which the curriculum reform had not yet been implemented. (For a detailed description of the intervention study, see Wilke et al. (2022).) This study only examines students in the intervention group. Students in this group had a mean age of 17.04 (*SD* 0.39); 38.2 per cent identified as male, 61.6 per cent as female and 0.2 per cent as other.

Procedure and design of the intervention study

The experimental condition

Students in the intervention group participated in a lesson series of 12–14 history lessons. The lessons were centred on the theme of decolonisation after 1945, and were aimed at fostering students' historical inquiry skills. Decolonisation was selected as the central theme because it easily allows for the inclusion of multiple perspectives, and it is studied in the last stage of Flemish history education. Design principles for the lesson series were derived from the model of cognitive apprenticeship (Collins et al., 1991), and they were applied to the specific context of history education via existing literature on the promotion of historical thinking. (For a detailed description of the design of the lesson series, see Wilke et al. (2023).)

The lesson series made use of four inquiry-based writing tasks, constituting authentic, discipline-specific tasks (Van Boxtel and Van Drie, 2018). They were centred on an evaluative historical question (Van Drie et al., 2006) to which students had to provide a substantiated answer based on a critical analysis of sources. The provided sources differed in terms of perspective on the historical question, and in terms of value. Each source was accompanied by background information, which students could use to evaluate its value. Students were taught how to provide a substantiated historical representation in response to these historical questions. A high-quality substantiated representation was taught to contain the following elements: a clear response to the historical question (claim); arguments derived from the sources to support the stance; references to the sources in support of the claim; and an evaluation of the value of the provided sources in order to argue why they considered the supporting sources to be valuable, and to rebut counter-arguments based on sources that are less valuable.

Throughout the lesson series, teachers first modelled an inquiry task. They modelled the reasoning process and steps required for the critical analysis of the sources, and for the composition of a (written) substantiated stance. In the subsequent lessons, students worked on other inquiry tasks in groups, supported by scaffolds which faded out gradually. Scaffolds (based on those of the University of Michigan (2021) and Monte-Sano et al. (2014)) were provided for each step in the process: corroborating and evaluating sources, weighing evidence and arguments, and formulating a substantiated answer to the historical question. Teachers coached students during the inquiry tasks by providing feedback and by adjusting the level of scaffolding to their students' needs, among other ways.

Explicit teaching on second-order and procedural knowledge related to how to construct a substantiated answer to a historical question, and related to critical sources analysis, was also an essential element throughout the lesson series. Teachers provided, for instance, explicit teaching on the meaning of the reliability and representativeness of the source, and on how to evaluate it (for example, by examining the author, the goal and audience of the source, and the time when the source was created), and on the essential elements and structure of a substantiated answer.

https://doi.org/10.14324/HERJ.20.1.05

To facilitate historical reasoning, students worked in groups on the inquiry tasks. Furthermore, whole-class discussion was used to discuss students' findings, and to provide additional opportunities for feedback and to advance students' reasoning (Reisman, 2012; Stoel et al., 2017; Van Boxtel and Van Drie, 2013, 2017).

Procedure

The study was conducted between January and April 2021. Students completed a pre- and posttest, each taking up two lessons (2 × 50 minutes). Teachers in the intervention group received training before the start of the study, and they were monitored during the intervention to ensure treatment fidelity. Online questionnaires were collected after each lesson, and two interviews were conducted with each participating teacher to discuss teachers' experiences and possible deviations from the provided lesson plans. Teachers who had skipped one of the inquiry tasks, or who had to delay the administering of the posttest (due to Covid-19 measures in Flemish education), were excluded from the study. Students who had missed more than two classes were also excluded (n = 12).

Measurement instrument: To measure students' historical inquiry skills, the pretest and posttest contained a historical essay task. Each task contained a short introduction, a historical question and a set of sources, containing different arguments and differing in perspective and value. The provided sources were selected in such a way that there was no clear-cut answer to the historical question. Rather, the sources provided conflicting accounts and answers to the historical question, in order to encourage students to critically evaluate and corroborate the various sources. Students were required to write a 15–20 sentence essay providing a substantiated answer to these questions (that is, substantiated with arguments and evidence), based on a critical analysis of the available sources (that is, not based on their own views).

The historical essay tasks centred on the questions: (1) whether the police at the Sharpeville massacre in South Africa (1960) had a legitimate reason to use violence; and (2) whether Dutch violence during the Indonesian Independence War (1945–9) was structural or incidental. These topics were related to the theme of the lesson series, but they were not covered in the materials. Moreover, the topics of the tasks are generally not well known among Flemish secondary school students. Students were provided with excerpts from both primary and secondary sources, presenting different views and arguments regarding the subject, and accompanied with background information needed to evaluate the source's value (for example, author and date of origin). Each task contained five sources, which were selected to be similar in type (primary–secondary, textual–visual) and difficulty level between the pretest and posttest. An overview of the essay tasks is presented in Table 1.

Students' essays were evaluated with a rubric containing six criteria. The initial rubric was based on research by Monte-Sano (2012), Monte-Sano and De La Paz (2012) and Monte-Sano et al. (2014), but adapted to better correspond with the explicit teaching in the lesson series. It was further adjusted after a pilot study in two classes (n = 36), which also evaluated the difficulty and feasibility of the essay tasks. The final rubric contained the following criteria: (1) making a claim in response to the historical question; (2) argumentation in support of the claim – this criterion evaluated the amount of arguments that were used to support the essay's claim; (3) quality of the argumentation, assessing the quality of students' arguments in support of their claim; (4) use of sources as supporting evidence – this criterion assessed the extent to which students had referenced specific sources in support of their claim; (5) evaluation of the value of the supporting sources, indicating to what extent students demonstrated that they had evaluated the value (in terms of reliability and/or representativeness) of the arguments they used in support of their claim; and (6) engagement with counter-arguments and evidence – this criterion indicated to what extent students had acknowledged the existence of arguments and evidence that went against their claim, and whether they had provided suitable rebuttals.

For each criterion, four competency levels were distinguished, ranging from poor (1) to excellent (4), resulting in a total score out of 24. Two raters scored 10 per cent of the pretest essays, and intraclass correlation (two-way, absolute agreement, single measures) was calculated to establish inter-rater reliability. This yielded at least a good agreement (>0.6) for each separate criterion (intra-class correlations (ICC) ranged from 0.64 to 1; M = 0.82, SD = 0.12) and an excellent agreement (>0.75) for the total score (ICC ranged from 0.75 to 0.97; M = 0.87, SD = 0.08) (Cicchetti, 1994; Hallgren, 2012).

https://doi.org/10.14324/HERJ.20.1.05

Table 1. Overview of historical essay tasks at pretest and posttest

Historical essay task			
Pretest: Police violence at the Sharpeville massacre		Posttest: Excessive Dutch violence during the Indonesian Independence War	
Source and provided background information	Content	Source	Content
A. Anti-apartheid journalist's report on the events at Sharpeville.	A. The report states that the police started shooting without any warning and without any provocation. It also reports that only three policemen got hurt, while 200 demonstrators were hurt by police violence. The journalist reports not to have seen any weapons among the demonstrators.	A. Excerpt from a television interview (1969) with a soldier who was active in the war. The soldier had unsuccessfully approached various media outlets immediately after the war to tell his story.	A. In the interview, the soldier states that excessive violence was conducted structurally during the war.
B. Photograph taken at the Sharpeville demonstration by an anti-apartheid photographer.	B. The photograph shows demonstrators running away (unarmed) from the police.	B. Photograph (1947) of Dutch soldiers during the war. The photograph was part of an exhibition of so-called 'unwanted images', as the photograph was deliberately kept from the Dutch general public during the war in an attempt to shape their perception of the war.	B. The photograph shows armed Dutch soldiers alongside a number of Indonesian soldiers who were taken captive.

History Education Research Journal
https://doi.org/10.14324/HERJ.20.1.05

C.	Excerpt from the court trial, interrogating the commanding officer about the events at the demonstration.	C.	In the court trial, the commanding officer reports that the demonstrators were hostile, and that the police were expecting an assault. He states that after a gunshot was heard, the policemen started shooting. He suggests that demonstrators were hurt because of ricocheting bullets.	C.	Letter from a veteran corporal during the Indonesian war in response to the television interview (1969).	C.	In the letter, the veteran corporal states that the television interview was a lie: violence is part of any war, but there was no excessive violence during this particular war.
D.	Official statement issued by the South Africa High Commissioner in London, shortly after the events, and after strong criticism from the international community.	D.	The statement reports that the police were attacked by the demonstrators, and acted in self-defence.	D.	Excerpt from a government report (1969) investigating the nature of the violence during the war. The report had to be completed within three months.	D.	The report states that Dutch soldiers behaved correctly during the war, that there were no cover-ups of excessive violence, and that there even existed accounts of positive acts committed by Dutch soldiers in Indonesia.
E.	Excerpt from a historian's book on the events at Sharpeville.	E.	The historian states that only a few of the demonstrators were armed, and that when a shot was fired from the crowd, the police started shooting. They fired in two rounds, including one round affecting the demonstrators running away from the field.	E.	Statement from the Dutch government in response to the publication of new historical research by Rémy Limpach on the subject of excessive violence during the war.	E.	In the statement, the government announces that a new investigation will be conducted into the existence and nature of excessive violence.

History Education Research Journal
https://doi.org/10.14324/HERJ.20.1.05

Data sampling: In line with the methodology applied by Stoel (2017), this study builds on the essay data collected as part of the intervention study, and uses purposeful sampling to study changes in students' essays in an in-depth manner. In our previous study (Wilke et al., 2022), we established large differences in students' initial essay scores, as well as in the progress they made. These differences are illustrated in the descriptive results in Table 2.

Table 2. Descriptive results on historical essay at pre- and posttest in the intervention group (*n* = 402)

Historical essay task (points out of 24)		
Pretest *M* (*SD*)	Posttest *M* (*SD*)	Change *M* (*SD*)
15.20 (3.58)	18.14 (4.45)	2.95 (4.93)

Table 2 first shows that students' performance on the pretest and posttest differed strongly, as evidenced by the large standard deviations. The standard deviation is even larger in the posttest than in the pretest, suggesting that differences among students increased after the intervention. Moreover, the large standard deviation for the change scores between pretest and posttest illustrates that the intervention did not generate the same level of progress among students. For this study, we opted for a critical case purposeful sampling (Patton, 2015), as we were particularly interested in those students for whom the intervention was very effective, that is, students who had made substantial progress on historical inquiry skills. This was operationalised as an increase in students' historical essay tasks of at least 6 points. Students with a 6-point increase between pre- and posttest progressed at least two times more in points than average, and the increase allows progression to be noticed on each of the rubric's criteria. In light of the research question, we then opted to divide students within this sample into two homogeneous groups according to their pretest score on historical inquiry, as we were interested in differential effects of the intervention according to students' starting positions. We therefore distinguished between students with moderate-to-high versus low initial scores, where a pretest score of 12 points or more out of 24 (that is, half of the maximum score) was considered to be a moderate-to-high starting point. A pretest score of 11 points or less (out of 24) was considered to be a low starting point.

This initially yielded a sample of 112 students who had made at least 6 points progress on the historical essay tasks. Students who had obtained a pretest score of 19 or more on this task (n = 65) had been excluded, as they would have been unable to progress 6 points. Of the 112 students, 45 had a low starting position and 67 had a moderate-to-high starting position. We selected a subsample of both groups for further qualitative analysis: students with a pretest score of 8 or 9 (n = 19), and those with a pretest score of 15 or 16 (n = 26). These subsamples were selected because these students' essays were neither the weakest nor the very best, and they still had room for improvement, yet they differed significantly (6 to 8 points) from each other. (Detailed information on both of the subsamples is provided in Table 4, found in the Results section.)

Data analysis: First, descriptive quantitative results of the obtained samples, based on the evaluation rubric's score, were explored to examine general patterns in students' essays and differences according to their starting positions. These findings were then further examined through a qualitative analysis. The qualitative analysis was conducted using NVivo, and it was an iterative process departing from an a priori set of codes, stemming from the evaluation rubric, supplemented by additional codes emerging from the essay tasks. In the essay tasks, all instances related to one of the coding themes were marked and coded. If applicable, multiple codes were given to one particular instance. The initial coding scheme was developed by the first author, and reviewed and discussed with the second and third authors until a final coding scheme was established, including clear descriptions for each of the codes. Descriptions and examples for each code are shown in Table 3.

https://doi.org/10.14324/HERJ.20.1.05

Table 3. Coding scheme

Code	Description	Example stemming from students' pretest (1) or posttest (2) essays
Position statement (claim)		
No statement	The essay does not contain a statement or claim in response to the historical/topical question.	'There has never been any clarity on who shot first and whether this was a matter of police brutality or pure self-defence.' (1) 'The police violence was both legitimate and illegitimate.' (1)
Explicit statement	The essay explicitly puts forward a clear statement or claim in response to the historical/topical question.	'Based on the analysis of these sources, I conclude that the acts of violence committed by Dutch troops in Indonesia were structural acts.' (2)
Implicit statement	The essay contains a statement or claim in response to the historical/topical question, but is not explicitly stated as such.	'In my opinion, this was a misunderstanding where there were bad apples in both groups. However, the police should have given a warning and stopped shooting once the crowd fled.' (1) 'I can well imagine that the police panicked because of this attack. But shooting at innocent people is never the solution, I think.' (1)
Partial statement	The essay contains a statement or claim that is only partly in response to the historical/topical question (for example, not entirely in line with the question).	'What exactly do I think about it? Well, I haven't been able to do enough research yet, but I think it's easy to conclude that killing people, for whatever cause and certainly in their backs, is something terrible and you should be punished for it.' (1)
Argumentation in support of the claim		
Own argumentation	The student provides an argument in support of their answer to the historical question, but this is not based on the sources.	'In addition, a lot of innocent people were injured. Because these people were fleeing, they were shot in the back. If a demonstration undertaken by a civil rights organisation itself violates the civil rights of others, I personally think that strict intervention is allowed.' (1) 'I think the police felt threatened by the large crowd, so they reacted in an extreme way. They killed everyone because they were afraid.' (1)
Argument derived from the sources		

https://doi.org/10.14324/HERJ.20.1.05

Form _implicit _explicit	Implicit: the argument is mentioned, but not clearly presented in support of the claim. Explicit: the argument is clearly put forward in support of the claim (for example, by use of phrases such as 'a first reason is that' or 'therefore').	*Implicit* 'One may ask: being pelted with some stones and one protester drawing a weapon, is this a good reason to start shooting at the protesters?' (1) 'Looking at the interview with the war veteran Joop Hueting in 1969, this man talks about the atrocities that he has seen happen. Namely: villages that were ruined, horrible interrogations that took place without military necessity.' (2) *Explicit* 'Another argument for systematic violence is the interview with Joop Hueting in 1969. In it he describes that that he and his fellow soldiers did many horrible things. It was normal, he says.' (2) 'A photograph taken in Malang in East Java in 1947 shows us how Dutch soldiers held Indonesian civilians at gunpoint with dead people lying next to them. This terrible image was seen as an undesirable one by the Dutch government and thus was not shown in the Dutch press during the period in which the troops were active in the Dutch East Indies. From this we can conclude that unnecessary violence was used and that the Dutch government tried to hide this from the population.' (2) 'Moreover, no warning shots were given and the police were pelted with stones (see sources A and C).' (1)
Quality _high _mediocre _low	High: the argument is derived from the sources, based on an accurate understanding of the source(s), supports the claim and is accurately/clearly presented/explained. Mediocre: the argument is derived from the sources and supports the claim, but it is not clearly presented/accurately explained.	*High* 'In 2016, the government launched a new investigation in response to a book that had been published. This book relies on Limpach's research. He confirmed that the violence was structural in nature and the new investigation shows that the government itself has its doubts about the real nature of the violence.' (1) 'According to journalist Humphrey Tyler, the crowd was not warned to disperse. The police accused the crowd of being heavily armed, however, the journalist claims that he did not see these heavy weapons.' (1) *Mediocre* 'I think the biggest problem here is with the mentality of the police before the demonstration and the violence. The police were not only prepared for violence, they expected violence (source C).' (1)

History Education Research Journal
https://doi.org/10.14324/HERJ.20.1.05

	Low: the argument is derived from the sources, but it is based on an inaccurate understanding of the source and/or does not support the claim.	*Low* 'It is noteworthy, that source E also came about after an accusation towards the Dutch. A comprehensive investigation, conducted by Limpach, once again states that the Dutch violence was of a structural nature. In 2016, the Dutch government insists on conducting a second investigation, in order to prove their innocence after all' [student misinterpreted the source]. (2)
Use of sources		
Source reference	A specific source is mentioned (for example, by referencing the source's identification letter or author), but not connected to the claim.	'Historian Tom Lodge describes the situation from a historical perspective. He states that the police shot at fleeing people, which can also be proven by the many bullet wounds in the back.' (1)
Source reference in support of the claim	A specific source is mentioned (for example, by referencing the source's identification letter or author), and presented in support of the claim.	'A first reason is that extreme violence was the normal course of events and was therefore a structural fact. This is said in the television interview with Joop Hueting (source A).' (2) 'A second reason [for the students' claim] is that the Dutch government wanted to cover up the incidents. I established this through the "unwanted" photograph that did later get exhibited in the Resistance Museum in Amsterdam. This also showed how the Dutch government manipulated the people by misrepresenting the events and making them look better.' (2)
Vague source reference	Sources are mentioned in general, without reference to one or more specific sources.	'Several sources indicate that the violence was illegitimate and that the police abused their power.' (1)
Evaluation of the supporting sources' value, in terms of their reliability and/or representativeness		
No evaluation of a source's value	No mention of any element or aspect of the source which may influence its value.	'Of these [protestors], many were unarmed. A few, however, had brought firearms (source E), which eventually caused a misunderstanding and thus led to the massacre.' (1) 'The picture of Ian Berry confirms this, we see the crowd running away. Also we see none of the so-called heavy weapons.' (1)
Superficial evaluation of a source's value	One or more relevant aspects of a source is mentioned (for example, author, date, goal of the source), but this is not connected to an evaluation of its value.	'Humphrey Tyler, a reporter, and Ian Berry, a photographer, both worked for Drum Magazine. This magazine expressed their support for the anti-Apartheid movement. The two men were present at the Sharpeville demonstration.' (1) 'The historian states that the demonstrators did start firing first but that the police reacted much too violently.' (1)

History Education Research Journal
https://doi.org/10.14324/HERJ.20.1.05

Evaluation of a source's value	The student evaluates the value of a source/explains why they consider the source to be valuable.	
Quality _High _Mediocre _Low _Nuance (additional/double coding)	High: Reasoning regarding a source's value makes sense/is accurate and clear. Mediocre: Reasoning regarding a source's value touches upon relevant aspects, but is not entirely clear or accurate, for instance in the case of oversimplification. Low: Reasoning regarding a source's value is missing or completely inaccurate. Nuance: Reasoning is presented both in favour of the source's value, and also against.	*High* 'A study by Dr Limpach shows that the violence was indeed structural in nature. Dr Limpach is himself a historian and has done extensive research to reach this conclusion, which leads me to believe that this source is reliable.' (1) *Mediocre* 'One can assume that the statement of the South Africa High Commissioner is subjective. The people in this commission were white and therefore had no idea what it is like to be black, in 1960 there was also very much racism.' (1) 'Because of this we see that the media manipulated the Dutch people by showing them only the pictures they wanted. This source is very reliable as it is an original photo from the war.' (2) 'Because Joop was an eyewitness we can conclude that the source is reliable.' (2) *Low* 'In this research under supervision of Dr Limpach, it was shown that unnecessary violence was indeed used on a structural level. Although Dr Limpach is an educated man we can question the reliability of this source as the research was carried out long after the facts.' (2) 'Source B shows a photo of Dutch soldiers standing next to wounded and deceased Indonesian soldiers. This is a reliable source because photoshop did not exist at that time and it reflects reality very well.' (2) *Nuance* 'I consider this to be a reliable source since he himself was present in the Dutch East Indies as a soldier, and therefore experienced the facts himself. One could say that he perhaps does not tell the facts 100% accurate as there is a 20 year time span between the incidents and the interview.' (2) 'In my opinion, this source is a reliable source to answer this question because his goal is to talk about his experiences, he also tried to tell his story in the 1950s but was always rejected. However, we do not know whether Hueting's experiences are representative of all the other soldiers since no other accounts were given.' (2)

https://doi.org/10.14324/HERJ.20.1.05

Criterion: Specification of which aspect of the source is mentioned in relation to its value _author _goal _audience _societal context _information _time/place _representativeness		*Time, information, goal:* 'Hueting's testimony is indeed reliable. He was present at the moment itself (although this was 20 years ago, so this is not a sufficient condition). In addition, his goal appears to be simply to tell the truth to the Dutch population. He has nothing to gain from this interview, does not defend himself and only tells what he saw.' (2)
Form _Implicit _Explicit	Implicit: the student evaluates the source's value, but does not explicitly mention this by referring to notions such as reliability, value, representativeness, neutrality or objectivity. Explicit: the student makes explicit that they are evaluating the source's value (for example, by referring to reliability, representativeness, neutrality or subjectivity of the source).	*Implicit* 'Joop Hueting talked about the war crimes in an interview and indicated that these were not just a few incidents but were daily occurrences. He was a veteran of the war so an eyewitness.' (2) *Explicit* 'We can assume that Joop Hueting is a reliable source as the man was an eyewitness during the war.' (2) 'One can assume that the statement of the South Africa High Commissioner is subjective. The people in this commission were white and therefore had no idea what it is like to be black, in 1960 there was also very much racism.' (1)
Rebuttal of counter-arguments and evidence		
Mentioned without dialogue	The essay contains a possible counter-argument or source, but this is not presented as such (no confrontation with other sources or arguments), or the student only refutes it based on their own input.	'Investigations have shown that a large number of people were shot in the back. According to Pienaar, this had two reasons: ricocheting bullets may have hit the protesters in the back. Also, some of them may have been hit in the back while fleeing. I find the first reason a bit unlikely.' (1)
Superficial dialogue	The existence of counter-arguments or sources is mentioned, but they are not refuted.	'According to the police, their unit was pelted by a large number of stones, and the crowd of protesters was also armed with heavy weapons. Humphrey on the other hand had not seen any weapons himself.' (1) 'Sources C and D contradict my position and thus state that everything was done correctly and the Dutch are not to blame.' (2)

History Education Research Journal
https://doi.org/10.14324/HERJ.20.1.05

Rebuttal	The existence of a counter-argument or source is mentioned and refuted by comparing the sources' value.	
Quality _High _Mediocre _Low	High: Reasoning regarding differences between the sources' value makes sense/is accurate and clear. Mediocre: Reasoning regarding differences between the sources' value touches upon relevant aspects, but is not entirely clear or accurate; for instance, in the case of oversimplification. Low: Reasoning regarding differences between the sources' value is missing or completely inaccurate.	*High* 'In the reader's letter from a former veteran we can read that he accuses Hueting of telling incorrect facts. To his knowledge no prisoners of war were ever killed, they were always released. At first sight this seems to be a reliable source, he has seen the facts happening just like Hueting and is therefore an eyewitness. But we may assume that after the television interview with Hueting, the authorities from this period will have received much criticism. So there is a possibility that this Corporal just wanted to clear his name. It is also possible that both Hueting and the Corporal were active in a different area. And thus that they did not witness the same acts of violence.' (2) *Mediocre* 'Another source on this subject is the 1969 government report. However, I do not find this reliable … They did not have the time to properly investigate both sides of the story and therefore I do not find the source representative.' (2) *Low* 'Source 3 is a reader's letter. This is probably just an old man wanting to clear his conscience. Thus, not so reliable.' (2)
Criterion: Specification of which aspect of the source is mentioned in relation to its value _author _goal _audience _societal context _information _time _place _representativeness _medium		

https://doi.org/10.14324/HERJ.20.1.05

Form _Implicit _Explicit	Implicit: the student evaluates the source's value to rebut the counter-argument, but does not explicitly mention this by referring to notions such as reliability, value, representativeness, neutrality or objectivity. Explicit: the student makes explicit that they are evaluating the source's value to rebut the counter-argument (for example, by referring to reliability, representativeness, neutrality or subjectivity of the source).	*Implicit* 'A possible counterargument could be that a study was already issued by the Dutch government in 1969. However, this study was only allowed to last three months (which is not enough time to comprehensively investigate everything). In addition, it came from the government itself. They could present the facts in a more positive way in order to present themselves in a more favourable way.' (2) *Explicit* 'J.G. Schuitema, a corporal from 1946 to 1948 in Central Java, issued a reader's letter eight days after the television interview, in which he writes that what Joop Hueting said is a big lie. Personally, I think this is a less reliable source, since he was a corporal and thus in charge of several soldiers. I think that with this letter he mainly wanted to save his own skin.' (2)
Essay type		
Descriptive	The essay's main focus is on describing what happened or summarising the content of the various sources. The essay provides a report of the events, as described in the sources. The essay may contain a claim in response to the historical question, but no arguments or sources are clearly put forward to support the claim.	
Argumentative	The essay's main focus is on arguing why a certain claim has been made. The essay provides a claim, and substantiates it by providing arguments and/or by referring to sources that support the claim. Arguments may be derived from the sources, but can also be based on students' own ideas. There is no attention to the value of sources.	
Evaluative	The essay's main focus is on (systematically) evaluating the different sources that were provided. Although the essay may include a claim, there is no/very little attention to (explicitly) substantiating the claim with arguments or sources (lack of connection between the claim and the sources).	
Argumentative-evaluative	The essay both substantiates a claim with arguments and sources *and* demonstrates attention to the value of sources. The essay contains arguments and source references in support of a claim *and* argues why certain sources may be considered more/less valuable in light of the historical/topical question.	

History Education Research Journal
https://doi.org/10.14324/HERJ.20.1.05

The predefined set of codes was derived from the criteria in the evaluation rubric, and it contained five main themes: claim; argumentation (including quality of argumentation); use of sources as evidence; evaluation of sources' value (that is, reliability and/or representativeness); and rebuttal of counter-arguments and evidence. For each of these main themes, subcodes were created allowing analysis of students' essay tasks in a fine-grained way.

'Claim' contained subcodes to indicate whether the essay contained an answer to the historical question, whether this answer was in line with the question, and whether the claim was explicitly or implicitly put forward. For the main code 'argumentation', subcodes were created to distinguish between students' own arguments and those derived from the sources, to examine the quality of the arguments (high, mediocre, low), and to indicate whether they were used implicitly or explicitly. 'Use of sources as evidence' evaluated whether the essay contained references to specific sources, and whether they were used explicitly in support of the essay's stance. 'Evaluation of sources' value' contained subcodes indicating whether students merely mentioned information about the source, such as who the author was (that is, superficial attention to the value of a source), or whether they also evaluated the source's value. On occasions where students evaluated the value of a source, the quality of the evaluation (high, mediocre, low) was coded, as well as which criteria they had used to do so (for example, author, date, goal) and whether this was done in an implicit or explicit way (explicit mentioning of reliability, value, representativeness and so on). The theme 'engagement with counter-arguments and evidence' contained subcodes to indicate whether students had acknowledged the existence of counter-arguments or evidence, whether they had engaged with them, and how this was done (superficially or by evaluating the sources' value). If the engagement with counter-argumentation included an evaluation of the sources' value, subcodes indicated whether this was done implicitly or explicitly, the quality of the evaluation (high, mediocre, low) and which criteria were used.

The coding process was initially guided by the pre-established coding scheme. Additionally, open and axial coding was used to highlight and label instances that were indicative of students' historical inquiry skills, which were not yet covered by the existing coding scheme (Baarda et al., 2009). Through this process, a new code was added to the coding theme 'evaluation of a source's value'. The code 'nuanced evaluation' was added to highlight instances where students had evaluated a source's value in a nuanced way, paying attention both to aspects that could increase and to aspects that could decrease the source's value. This code was added to the coding scheme as an additional or 'double' code. Students' reasoning regarding the evaluation of a source's value was hence coded as high, medium or low, and – if applicable – was additionally also coded as nuanced.

Besides the five main coding themes stemming from the evaluation rubric, the process of open and axial coding resulted in an additional coding theme. This coding theme was used to describe various types of students' essays as it became clear that these took on distinctive forms, and transformed between pretest and posttest. These essay types were coded inductively, grounded in the data themselves, as they did not correspond to existing typologies from the literature, such as those of Coffin (2004, 2006). Based on students' essays, we distinguished four main codes for this theme, describing the main focus of students' writing: descriptive, evaluative, argumentative and argumentative-evaluative, depending on the main focus of the essay. A descriptive essay's main focus was on describing what happened, and what the various sources mentioned about the event. An argumentative essay's main focus was on arguing a certain answer to the historical question by use of arguments and/or evidence (sources), but without an evaluation of the sources. Evaluative essays were mainly centred on the evaluation of the sources, but lacked a connection to the essay's claim. Finally, an argumentative-evaluative essay did both. It substantiated a certain answer to the historical question with arguments and evidence (sources), and it included an evaluation of the sources.

As the coding for the five main themes derived from the evaluation rubric stayed close to the initial rubric, for which an acceptable inter-rater agreement had been reached, no additional measures for inter-rater agreement were calculated for the coding related to these themes. For the additional coding theme 'essay type', the initial coding by the first author was reviewed and discussed with the third author until consensual agreement was reached.

https://doi.org/10.14324/HERJ.20.1.05

Results

In order to examine changes in the historical inquiry tasks among students who had made a lot of progress, the quantitative results based on the evaluation rubric were first examined and compared for students with a low and high starting position. Table 4 shows for both groups their main total score on the historical essay task at pretest and posttest, and their mean score on each criterion of the evaluation rubric.

Table 4. Mean score on students' historical essay tasks at pretest and posttest, based on the evaluation rubric, for the selected subsamples

	Low pretest score on historical essay task *n* = 19			High pretest score on historical essay task *n* = 26		
Historical essay task	**Pretest *M* (*SD*)**	**Posttest *M* (*SD*)**	**Change *M* (*SD*)**	**Pretest *M* (*SD*)**	**Posttest *M* (*SD*)**	**Change *M* (*SD*)**
Claim	1.79 (1.10)	3.95 (0.23)	2.16 (1.12)	3.23 (0.77)	3.92 (0.39)	0.69 (0.74)
Argumentation	1.05 (0.23)	3.37 (0.76)	2.32 (0.82)	2.73 (0.78)	3.77 (0.43)	1.04 (0.92)
Quality of argumentation	1.00 (0.00)	3.47 (0.61)	2.47 (0.61)	3.50 (0.65)	3.85 (0.37)	0.35 (0.80)
Use of sources as evidence	1.00 (0.00)	3.74 (0.65)	2.74 (0.65)	2.81 (0.94)	4.00 (0.00)	1.19 (0.94)
Evaluation of sources' value	1.84 (0.61)	2.68 (0.95)	0.84 (1.12)	1.73 (0.45)	3.58 (0.50)	1.85 (0.78)
Engagement with counter-arguments and evidence	1.58 (0.51)	2.47 (0.84)	0.89 (0.99)	1.50 (0.51)	3.38 (0.57)	1.88 (0.77)
Total score	8.26 (0.45)	19.68 (2.56)	11.42 (2.61)	15.50 (0.51)	22.50 (1.03)	7.00 (0.94)

Based on Table 4, a number of general trends can be distinguished in students' progress on historical inquiry skills. At pretest, students with a high pretest score achieved remarkably higher scores on the criteria of claim, argumentation, quality of argumentation and evidence. On the criteria of evaluating the value of sources and engaging with counter-arguments and evidence, both groups performed more or less equally low at pretest.

Both groups progressed on all six criteria between pretest and posttest. At posttest, students with a low pretest score seemed to catch up in terms of claim, argumentation, quality of argumentation and evidence, as the differences between both groups decreased. Regarding evaluating a source's value and dialogue, however, differences between both groups increased. Whereas students in both groups obtained more or less equal scores on these criteria at pretest, students with a high initial score clearly outperformed those with a low initial score at posttest.

While both groups progressed on all criteria, this progress was unevenly distributed among the various criteria, and it manifested itself differently in each group. In the group with a low pretest score, progress appears to have happened predominantly in terms of stance, argumentation, quality of argumentations and use of evidence. Progress also occurred in the other criteria, but it was far less apparent. In the high pretest group, strong progress was made with regard to use of evidence, evaluating a source, and engagement with counter-arguments.

The results of the qualitative analysis allow for an exploration of how precisely this progress manifested itself in the essay tasks, and how it differed between the two groups. For both groups, the frequency of codes within each main coding theme is depicted in Table 5. These results will be discussed for each of the main coding themes, first for the group with a low pretest score on historical skills, then for the group with a high pretest score. Throughout the discussion of the results, quotations stemming from students' essay tasks will be used to illustrate the findings. We have selected examples that were both common and clearly illustrative for the observed phenomena.

History Education Research Journal
https://doi.org/10.14324/HERJ.20.1.05

Table 5. Frequency of coding and number of essays in which each code was used

Name	Students with a low initial score on historical skills *n* = 19				Students with a high initial score on historical skills *n* = 26			
	PRETEST		POSTTEST		PRETEST		POSTTEST	
	Number of essays	Frequency	Number of essays	Frequency	Number of essays	Frequency	Number of essays	Frequency
Claim								
None	6	6	0	0	0	0	0	0
Explicit	3	3	18	18	14	14	26	26
Implicit	6	6	1	1	11	11	0	0
Partial	4	4	0	0	1	1	0	0
Argumentation								
Own argument	4	4	0	0	3	4	0	0
Explicit argument derived from sources	5	7	12	21	17	27	21	45
Implicit argument derived from sources	6	8	7	9	14	18	10	18
Quality								
High	6	11	15	24	25	41	26	59
Mediocre	4	4	4	4	3	3	3	3
Low	0	0	2	2	1	1	1	1
Use of sources as evidence								
Vague source reference	2	2	1	1	4	8	1	1
Source reference	11	25	7	10	14	38	8	12
Source reference in support of the claim	6	10	19	36	14	22	25	53

History Education Research Journal
https://doi.org/10.14324/HERJ.20.1.05

Evaluation of sources' value								
No evaluation	7	8	14	16	14	22	12	13
Superficial evaluation	8	12	3	3	15	26	4	4
Evaluation of a source's value								
Form								
Explicit evaluation	4	8	14	20	6	8	23	42
Implicit evaluation	5	7	7	7	2	4	6	6
Quality								
High	2	2	7	8	2	2	16	22
Mediocre	6	10	9	11	5	9	18	23
Low	3	3	8	8	1	1	3	3
Nuance (additional/double coding)	0	0	6	6	0	0	10	11
Criterion								
Author	9	13	6	6	6	13	6	10
Societal context	0	0	1	1	1	1	9	11
Goal and audience	1	1	2	2	3	3	6	6
Information	5	7	11	15	8	9	23	33
Representativeness	0	0	5	8	0	0	10	15
Time and place	2	3	3	3	2	2	11	12
None	0	0	1	1	0	0	0	0
Engagement with counter-arguments and evidence								
Mentioned without dialogue	6	7	1	1	11	12	1	1
Superficial dialogue	4	5	4	5	9	11	1	1
Rebuttal								
Form								
Explicit evaluation	0	0	10	16	1	1	23	35
Implicit evaluation	1	1	9	11	0	0	10	12

History Education Research Journal
https://doi.org/10.14324/HERJ.20.1.05

Quality								
High	0	0	2	2	0	0	2	2
Mediocre	1	1	12	16	1	1	12	14
Low	0	0	7	9	0	0	22	31
Criterion								
Author	1	1	7	9	1	1	18	22
Goal and audience	0	0	6	6	0	0	13	17
Information	0	0	14	17	0	0	25	25
Representativeness	0	0	4	5	0	0	13	16
Time and place	0	0	0	0	0	0	4	4
Societal context	0	0	0	0	0	0	1	1
None	0	0	0	0	0	0	0	0
Essay type								
Descriptive	14	14	0	0	11	11	0	0
Evaluative	3	3	6	6	2	2	8	8
Argumentative	1	1	4	4	13	13	2	2
Argumentative-evaluative	1	1	9	9	0	0	16	16

https://doi.org/10.14324/HERJ.20.1.05

Progress on historical inquiry skills among students with a low initial score

Claim

Students at pretest had great difficulty in providing an answer to the historical question. Only 3 essays included an explicit answer to the historical question. Other students only referred to the historical question implicitly (6), partially (4) or not at all (6). At posttest, all but one of the essays contained an explicit claim in response to the historical question. Students thus exhibited great progress on this aspect of the essay tasks.

In their pretest essay, student B3, for instance, provided an elaborate report on the events at the Sharpeville demonstration, ending the essay with 'we conclude that the Apartheid regime was a very dark chapter in the history of colonialisation', which did not provide an answer to the historical question at hand. At posttest, this student's essay contained a very different ending: 'After a considered evaluation and source corroboration, I came to the following conclusion: The violence of the Dutch military was structural in nature.'

Argumentation

At pretest, several difficulties can be observed with regard to students' use of arguments. First, some of the students' essays (8) did not contain any argument derived from the sources. Second, students did not always base their arguments on the information provided by sources, but provided arguments based on their own ideas. Regarding the police violence at the Sharpeville demonstration, student B10, for instance, explained why the police did not have a legitimate reason to use violence, using arguments that were not derived from the sources:

> But still I stand by the fact that what those police officers did could have been avoided and that the decision to fire dozens of shots and to injure 200 people should not and can never be approved. Those protestors already had fewer benefits and already started the protest from a disadvantaged position. They were therefore frustrated, but an escalated protest should not resemble a scene from the First World War! If I study the sources further, there are lies and falsehoods to be found on sides of the story, but the most can be found on the side of the policemen. It can be said that those policemen were not happy with the [protest's] turnout and because it got out of hand, they were very angry and heard a shot which made them think, subconsciously, that it was best to shoot at those, in their eyes, filthy people. This is all due to the Apartheid regime which made those police officers look at those black people as inferior.

When students supported their answer with their own arguments, this sometimes reflected presentism. For instance, student B18 argued against the police violence by explicitly referring to contemporary regulations for police enforcement:

> According to Belgian law, a policeman must avoid using his gun at all costs. It also states that an officer must first fire a warning shot before firing a bullet. Finally, the force used must always be proportionate to the situation and the crime. So the police should only fire when it is unavoidable. We can apply these standards to the situation in Sharpeville. … By today's standards, the police officer must himself be in danger, in mortal danger even, before he or she shoots. Whether that was the case in South Africa I leave in the open, as I have too few sources to make a judgement. If the indigenous masses actually became aggressive, perhaps a better solution would have been to arrest the offenders with or without force.

Both students hence argued not based on the information provided by the sources, but rather based on contemporary arguments or on their own views on the issue. A third issue regarding students' argumentation at pretest was that in cases where arguments derived from the sources were used, half of them were only mentioned implicitly (8 out of 15), although they were usually of high quality (11 out of 15).

Between pre- and posttest, students made a lot of progress regarding argumentation. Whereas at pretest 10 essays contained not a single argument derived from the sources, all but 3 of the essays at posttest contained at least one source, and students no longer used arguments based on their own views. The number of arguments derived from the sources doubled (from 15 to 30), and students connected

https://doi.org/10.14324/HERJ.20.1.05

the arguments more explicitly to their answer (21 out of 30 were used explicitly). At posttest, student B10, for instance, substantiated his answer explicitly with arguments derived from the sources, rather than providing his own explanation for the events:

> I am convinced that these [violent acts] are rather of a structural nature because of the following arguments. First, these atrocities committed by Dutch soldiers have been confirmed by the study of Dr Limpach (see source E). This is an extensive study on the use of excessive violence by soldiers of both the Dutch and Indonesian side of these acts during the war of independence. In addition, Dr Limpach also mentions that these Dutch atrocities were part of a larger scheme, of a larger structural nature.

The overall increase in the number of arguments, and in the number of arguments used explicitly in support of a claim, show that, after the intervention, students in general appeared to have gained more awareness of the need to (explicitly) substantiate their claim with arguments derived from the sources.

Use of sources

Regarding the use of sources, the majority of students referred to specific sources during their pretest essays. However, in most cases, students merely mentioned sources, and did not explicitly put forward these source as evidence to support their claim (25 out of 35 sources were not connected to the essay's claim). At posttest, students increasingly referenced sources in their essays (from 35 at pretest to 46 at posttest). Significant progress was also made in the way that students used these sources. At posttest, most of the sources that were mentioned were connected explicitly to students' claims and, hence, were used as evidence in support of these claims (36 out of 46 used sources explicitly). Whereas at pretest only 6 essays used specific sources to support their claim, all posttest essays contained at least one source that was explicitly used as such.

This change is apparent in the excerpts from student B10 (see above). At pretest, this student only vaguely referred to the sources ('If I study the sources further, there are lies and falsehoods to be found on both sides of the story, but most of them are found on the side of the policemen'), whereas at posttest, this student connected specific sources to the claims in their essay.

Evaluating the sources' value

At pretest, students' attention to the value of sources was rather poor. For the majority of sources, students either did not mention anything that may have influenced the sources' value (8 times), or merely mentioned something about the source, such as the author, but did not connect it to an evaluation of the source's value (12 times). Student B13, for instance, mentioned important aspects of the sources, but did not relate this to the value of the sources:

> From source C we can infer that civilians were hit in the back, we can corroborate this from source B. Source C, a lieutenant corporal in charge of the police officers present [at the demonstration], states that a police officer was abruptly pushed backwards by the crowd. This, in turn, is confirmed by source E, the historian.

On 15 occasions, students evaluated the value of a source. This evaluation, however, was only done explicitly about half the time (8 out of 15), and the quality of students' reasoning was mostly low (3) or mediocre (10). This was most often due to simplistic reasoning, based on a naive distinction between 'objective' and 'subjective' sources. Other common issues with regard to students' evaluation of sources were unsubstantiated evaluations of a source's value, or reasoning errors. Student B9, for instance, mentioned in the posttest 'The picture in source B shows the harsh reality in the Dutch East Indies. Because it is a photograph, this source is reliable.' Another example of a reasoning error can be found in essay B12, where the student claimed the following:

> In 2011 an excerpt from the book by the historian Tom Lodge was published. In it he tells of a private attack by someone in the crowd, which has never been effectively proven. In addition, this was not published until 2011, which makes this source less reliable.

A similar reasoning error was apparent in essay B13, in which the student evaluated the statement of the Dutch government declaring the start of a new investigation into excessive violence during the war:

https://doi.org/10.14324/HERJ.20.1.05

> In the last source, the Dutch government decides to open another investigation following the publication of a book by Rémy Limpach who also states that the violence was structural in nature. This confirms again that there is sufficient evidence to start an investigation. It is again confirmed that there are indications of structural violence during the decolonisation of the Dutch East Indies. However, it must be noted that this source was created in 2019, 65 years after the facts. This again makes it less reliable due to the large time difference.

While these students clearly tried to evaluate the sources by reflecting on the time frame between the historical event and the source, their reasoning is in this case rather poor.

At posttest, students' attention to the value of sources advanced in parallel with the increased use of sources as evidence. While students still regularly neglected to evaluate a source (16 times), or did so only superficially (3 times), the majority of sources at posttest were subjected to an evaluation of their value, indicating increased attention to the need to evaluate sources. Compared to the pretest, students increasingly evaluated sources explicitly (20 out of 27 times), and this was done in a more accurate way. Whereas at pretest students referred mostly to 'objectivity' or 'subjectivity', students at posttest referred to the more accurate concepts of reliability and representativeness. Here, a clear influence can be distinguished of the intervention, which paid explicit attention to these second-order concepts.

While the quality of students' reasoning about the value of sources remained somewhat the same, varying between low (8), mediocre (11) and high (8), students did exhibit a broader understanding of the concept of reliability. At posttest, 5 students evaluated the representativeness of a source, and 6 students included a nuanced evaluation of a source, with attention both to aspects that increased the source's value and to aspects which may decrease it. Student B11 wrote at posttest about the television interview with Joop Hueting:

> You could say that the source is unreliable because it was written 20 years later. But Joop Hueting was Dutch himself and confessed to the terrible deeds carried out by the Dutch. He also had to go through a lot of trouble to be able to tell his story, because many newspapers had rejected him.

Overall, students at posttest demonstrated increased and more explicit attention to the value of sources. This progress, however, should not be overstated, as students still did not systematically evaluate the sources' value, and the quality of students' evaluations still varied strongly.

Engagement with counter-arguments and evidence

Students' pretest essays were characterised by a lack of thorough attention to counter-arguments and rebuttals. In the rare cases (4 essays) where some attention to counter-arguments was present, this only happened in a superficial way. Student B11, for instance, mentioned that the sources disagreed with regard to the question of whether the demonstrators threw rocks at the police, but offered no explanation as to why these sources' accounts may have differed, nor about which of the sources she deemed more reliable:

> In response, the police fired bullets at the crowd. According to source 4, the police fired in self-defence. I can understand that if everyone had been throwing rocks at them, but that was not the case according to sources A and E.

At posttest, most essays exhibited some attention to the existence of a counter-argument and, in line with students' increased explicit attention to evaluating sources, these rebuttals were more often linked explicitly to the source's value. On 26 occasions, students attempted to rebut a counter-argument by comparing the value of the sources. These evaluations, however, were usually mediocre (16) or poor (2) in quality, although 8 students managed to demonstrate high-quality reasoning in this regard.

Essay type

Pretest essays predominantly took on a descriptive focus (14), with students describing the events based on information that they had extracted from the provided sources. Between pretest and posttest, strong shifts occurred in the nature of students' essays, as these were no longer descriptive at posttest. Rather, students' essays became evaluative (6), argumentative (4) or, in the majority of cases (9), both. This

https://doi.org/10.14324/HERJ.20.1.05

indicates an increased attention to substantiating a claim and evaluating the value of sources, which is in line with changes that were established in the previous coding themes.

Progress on historical inquiry skills among students with a high initial score

Claim

Concerning the statement provided in students' essay tasks, all essays at pretest contained an answer to the historical question, although 11 out of 26 did so implicitly. At posttest, all essays contained an explicit answer to the historical question.

Argumentation

Most students performed relatively well regarding argumentation at pretest. Their essays contained arguments which were mainly of high quality (41 out of 45), and the majority of which were used explicitly in support of the claim (27 out of 45 were used explicitly).

At posttest, students' argumentation improved further in a number of aspects: students provided even more high-quality arguments (on average from 1.7 arguments at pretest to 2.4 at posttest) which were predominantly connected explicitly to students' claims (45 out of 63). Despite these positive results, however, 10 essays still contained at least one argument which was not connected to the essay's claim.

Use of sources

At pretest, students' essays contained a total of 60 references to specific sources. However, 4 essays contained only vague references to the sources, and the majority of source references (38 out of 60) were not connected to students' claims.

At posttest, the number of specific source references remained more or less stable (from 60 to 65), yet the most striking progress was how these were used in the essays. At posttest, the vast majority of sources (53 out of 65) were referenced specifically in support of the answer that was argued. This progress, however, did not occur across all essays. In some cases, students' use of sources became less instead of more explicit. These students had pretest essays which contained several source references in support of their claim, yet at posttest their use of sources became less explicitly connected to their essay's stance. Of the posttest essays, 8 predominantly contained references to sources which were not used in support of the stance.

Evaluating the sources' value

Students' pretest essays overall showed only low levels of attention to the value of sources. Students either did not include any information that may be relevant to the source's value (22 times), or only mentioned it superficially (26 times). In some essays, however, students did already go beyond the superficial level and included an evaluation of a source. This was done mostly in an explicit way (8 out of 12 times) and based on mediocre reasoning.

Between pretest and posttest, significant progress could be observed both in students' attention to the value of sources and in the quality of their reasoning. Regarding the attention to the value of sources, students' at posttest were more inclined to include an evaluation of a source. Whereas at pretest, students only evaluated a source 12 times, this happened 48 times at posttest. In almost all instances (42 out of 48 times) students explicitly referred to the notion of reliability or representativeness when evaluating a source. The quality of students' reasoning also improved, with 22 out of 48 evaluations being of high quality (22). High-quality reasoning is illustrated in the posttest of student A8. This student explicitly evaluated several sources, including the television interview with Hueting, about which she wrote:

> This concerns an interview with a veteran who was at the scene himself and is therefore an eyewitness and even participated in the extreme violence himself. This source is therefore highly reliable. However, this source alone does not give a complete picture because it is only about the experiences of one person in the area where he fought. There is also another war veteran who contradicts this in a reader's letter.

https://doi.org/10.14324/HERJ.20.1.05

Students, however, still struggled with the evaluation of sources, and encountered similar difficulties to students with low pretest scores. Student A18's reasoning, for instance, depicted a naive, rather simplistic evaluation of the television interview with Joop Hueting. She stated:

> After the testimony of an ex-soldier Joop Hueting in 1969 it became clear that the violence occurred on a regular basis. He described how he himself saw that militarily unnecessary actions were carried out, such as the assault on the *kampongs* (local villages) and revenge actions. Because Joop was an eyewitness we can conclude that the source is reliable.

Whether the author was an eyewitness or not is definitely relevant in evaluating the source, but it does not by itself guarantee that a source is reliable, especially considering that the source set contained another eyewitness who provided an entirely opposite view. Another student (A19) inaccurately used the criterion of 'time' to question the reliability of a historian's work. She argued that 'This source is very reliable because the research deals specifically with this issue. However, we must take into account that the book was written sixty years after the date of the event.'

Despite the variance in the quality of students' reasoning about the value of sources, students did apply a more diverse set of criteria to evaluate the sources, and showed increased attention to the representativeness of sources (15 times) and for nuanced evaluations (11 times).

Engagement with counter-arguments and evidence

Pretest essays demonstrated only limited attention to counter-arguments, with counter-arguments being mentioned 11 times, but rarely rebutted (only once at pretest). This changed drastically in the posttest, where students exhibited a strongly increased attention to counter-arguments (47 times). In these instances, students not only mentioned the existence of a counter-argument or evidence, but rebutted it by evaluating the various sources. On 35 occasions, this was done explicitly, and on 12 occasions, it was done implicitly. Not only did students' attention to counter-arguments and evidence increase, their reasoning at posttest was mostly of high quality (31 out of 47 times). For example, in her essay about the Indonesian Independence War, student A4 wrote:

> In contrast to this study, the *Excessennota*, a study commissioned by the government, states: 'The government regrets that excesses occurred, but it maintains its view that the armed forces as a whole behaved correctly in Indonesia. The collected data confirm that systematic brutality did not occur'. However, this does not strike me as a reliable source because it was written at the behest of the government and they may indeed have an interest in covering up the situation. In addition, it was not a thorough study because it had to be completed in a short period of time.

This student explicitly rebutted the counter-argument provided by the government report by accurately arguing why this source was less reliable than the other sources provided.

When rebutting counter-arguments by referring to the value of sources, the same difficulties sometimes occurred as described earlier regarding the evaluation of sources. Students sometimes expressed rather simplistic reasoning, or their reasoning was not well supported. In both cases, their reasoning was categorised as 'mediocre quality'.

Essay type

Pretest essays of students with a high pretest score were most likely to take on a descriptive (11) or an argumentative (13) focus. At posttest, no descriptive essays were identified. Rather, students' essays had shifted predominantly to an evaluative (8) or an argumentative and evaluative focus (16), indicating an increased attention to evaluating sources.

Conclusion and discussion

Building on data collected in a large-scale intervention study, this research aimed to explore differential effects of an intervention on students' historical inquiry skills according to their initial starting positions. The qualitative analysis of essays of a purposeful sample of students who had progressed substantially

https://doi.org/10.14324/HERJ.20.1.05

after the intervention revealed remarkable differences between students with high and low pretest scores. Although both groups had made progress on all criteria evaluated in the rubric, the groups differed in terms of where this progress was most apparent.

At pretest, students with low pretest scores typically struggled with the need to formulate an explicit answer to the historical question, and to substantiate it with arguments and evidence. Their pretest essay either did not contain a claim, or their claim was only weakly substantiated with evidence and arguments from the sources. Progress in this group was situated mostly with regard to these aspects. Between pretest and posttest, these students overall seemed to become more aware of the need to make a claim and to support it, using the available sources. Besides these notable changes, there also appeared to be an emerging recognition among most students of the need to address the sources' value and to engage with counter-arguments. Students, however, did so only sporadically, and when they did, their reasoning was often not of high quality. These changes are reflected in the changing essay structures, where a shift occurred away from descriptive essays which lacked a substantiated stance. Instead, students wrote essays which were centred on substantiating a stance, evaluating sources or both.

Students with a high pretest score predominantly wrote pretest essays in which they took a certain position towards the historical question, and substantiated it with arguments and sources. This, however, did not always happen in an explicit way. At pretest, students' essays exhibited a limited attention to the value of sources, but usually in a superficial way. The same can be said for their engagement with counter-arguments, which occurred only in a limited way. In the posttest, these students made especially strong progress with regard to evaluating sources and engaging with counter-arguments. This was done more consistently and explicitly, and based on high-quality reasoning. They hence demonstrated an increased awareness of the need to argue why they considered some sources to be more valuable than others, as well as an increased ability to do so. With regard to argumentation and using sources as evidence, the change between pretest and posttest among these students was more ambiguous. Although there was an overall increase in the number of arguments and source references, these were in some cases not clearly connected to the position taken. Students hence did not always progress in a straightforward way across all criteria. An unwanted side effect of students' increased attention to the value of sources appeared to be that students failed to use the content of the sources to provide arguments for their answer to the historical question, and overlooked the need make a clear connection between the sources they discussed and the position taken. In some cases, students' increased focus on evaluating sources thus seemed to be at the expense of the argumentative nature of the essay. Rather than adding the evaluation of sources to their already argumentative essays, these students shifted from an argumentative to an evaluative essay type.

The qualitative analysis of the essay tasks allowed for a much more detailed mapping of changes in students' pretest and posttest essays and, in so doing, provided insight into patterns of progress. These analyses show that there seems to be some sort of sequence in students' development, with certain aspects of historical inquiry only developing fully when students are better acquainted with other aspects. In particular, providing a claim, and arguments and evidence in support of the claim, seem to be the basic aspects of historical inquiry that students master first. Only then do students seem to progress in more advanced aspects of historical inquiry, namely critically evaluating sources and engaging with counter-arguments and evidence. This is evidenced by the differences in progress made by students with high and low initial scores. These clearly show that progression on the more advanced aspects only occurred substantially among those students whose pretest essays showed that they had already mastered to a great extent the more basic elements of historical inquiry.

While the analyses show a certain progression in students' historical inquiry skills on the separate criteria of the evaluation rubric and qualitative coding scheme, they also reveal that progress on the various aspects of historical inquiry are not separated from each other. It seems that as students become more proficient in these (written) inquiry tasks, the progress that students make on one aspect of the inquiry task is inextricably linked to progress on other aspects. For instance, students' increased attention to substantiating a claim resulted in improvements with regard to argumentation, which became more elaborate and more extensively based on the sources provided, as well as with regard to the use of evidence. This is in line with Monte-Sano's (2010) claim that characteristics of students' historical writing will overlap at higher performance levels. Yet, this connection did not exist between all aspects of students' writing, but rather among the more basic aspects and the more advanced.

The student essays also make clear that historical thinking remains an 'unnatural act' (Wineburg, 2001), and demonstrate in particular how difficult certain discipline-specific ways of reasoning are

https://doi.org/10.14324/HERJ.20.1.05

for students. Most notably, and in line with earlier studies (for example, Harris et al., 2016; Rouet et al., 1996; Van der Eem et al., 2022), evaluating sources and engaging with counter-arguments and evidence remained a challenge, even among those students for whom the intervention had worked particularly well.

As in Stoel's (2017) research, we observed a relationship between text structure and certain characteristics of students' historical writing. In our analysis, however, four, rather than two, essay types emerged. In contrast to Stoel's (2017) research, we observed strong shifts in these text structures between pretest and posttest as a result of the intervention. After the intervention, students' essays more often adopted a text structure that paid more attention to the evaluation of sources. It appears that the intervention's explicit attention to procedural knowledge on how to formulate a substantiated answer to a historical question based on sources affected students' essays. These evolutions nevertheless demonstrate that progress on historical inquiry is not a strictly linear process. Changes in the essay types showed that, in some cases, students' increased attention to the value of sources decreased their attention to other aspects, such as argumentation and using evidence. In particular, evaluative essays appeared to function in some cases as an intermediate position, when students evolved from a descriptive essay to a truly argumentative-evaluative essay.

An important limitation of this study is that it is based on a purposeful sample of precisely those students who had progressed substantially on historical inquiry. This sample allowed us to closely examine what such progress looked like and what role students' starting position played in this regard, but it provides no insight into the way in which progress manifested itself in students for whom the intervention yielded a more moderate effect. The study is therefore limited in terms of its generalisability. In addition, the study provides no explanation as to why the intervention worked particularly well for these students. The use of written inquiry tasks constituted another limitation. While disciplinary writing in history has distinct characteristics (for example, Monte-Sano, 2010), students' historical writing is also connected to their general literacy skills (for example, Reisman, 2012), and students are not always able to accurately demonstrate their historical reasoning in a written task (Stoel et al., 2015; Van Drie et al., 2013). Also, this study adopts a particular operationalisation of 'progression' in history, measured in a relatively short period of time, rather than as a gradual process taking place over several years. Moreover, this study measured students' progress by means of a change score between their pretest and posttest performance in the context of an intervention study. The study used only these two data points, and it does not provide insight into the precise process of change occurring in between these two measurements. Such data would provide an even better understanding of the development of students' reasoning.

Despite these limitations, the study provides some important implications for history education and future studies, as it reveals patterns in students' progress in historical inquiry and offers possibilities for differentiating according to students' starting positions. Based on our findings, it seems beneficial to focus first on the argumentative elements of written historical inquiry tasks (that is, claim, arguments and evidence), before adding the need to include the evaluative elements (that is, source evaluation and rebuttal of counter-arguments and evidence). In line with Nokes (2017), we also recommend that assessments of students' historical thinking, as well as teachers' instructional practices, should be adapted to these differences in students' initial abilities, which may be present within a single group of students. Students with lower incoming abilities may need to focus predominantly on substantiating their claims, while students who have mastered these aspects may be encouraged to provide more attention to accurately evaluating sources and engaging with counter-arguments. Several intervention studies, including our own, have demonstrated the benefits of using scaffolds when asking students to engage with documents-based tasks (for example, De La Paz and Felton, 2010; De La Paz et al., 2016). These scaffolds can be used as a tool for differentiation in the classroom, when adapted to students' individual needs.

For the reporting of future intervention studies, this study demonstrates the value of including a more in-depth analysis and description of the effects of the designed intervention. It also highlights the importance of paying attention to ingroup differences in intervention studies, especially in light of their practical implications for instructional practice.

https://doi.org/10.14324/HERJ.20.1.05

Funding

This work was supported by the Special Research Funds, KU Leuven (grant number 3H180247). The funding agency was not involved in the research.

Acknowledgments

We would like to thank the teachers and students who participated in this study, as well as the members of the Dutch-Flemish History Research Group (NVOG) for providing feedback on an earlier version of this manuscript.

Declarations and conflicts of interest

Research ethics statement

The authors declare that research ethics approval for this article was provided by the Social and Societal Ethics Committee (SMEC – KU Leuven).

Consent for publication statement

The authors declare that research participants' informed consent to publication of findings – including photos, videos and any personal or identifiable information – was secured prior to publication.

Conflicts of interest statement

The authors declare no conflict of interest with this work. All efforts to sufficiently anonymise the authors during peer review of this article have been made. The authors declare no further conflicts with this article.

References

Baarda, D.B., de Goede, M.P. and Teunissen, J. (2009) *Basisboek kwalitatief onderzoek: Handleiding Voor Het Opzetten en Uitvoeren van Kwalitatief Onderzoek*. Groningen: Noordhoff Uitgevers.

Britt, M.A. and Aglinskas, C. (2002) 'Improving students' ability to identify and use source information'. *Cognition and Instruction*, 20 (4), 485–522. [CrossRef]

Cicchetti, D.V. (1994) 'Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology'. *Psychological Assessment*, 6 (4), 284–90. [CrossRef]

Coffin, C. (2004) 'Learning to write history: The role of causality'. *Written Communication*, 21 (3), 261–89. [CrossRef]

Coffin, C. (2006) *Historical Discourse: The language of time, cause and evaluation*. London: Continuum.

Collins, A., Brown, J.S. and Holum, A. (1991) 'Cognitive apprenticeship: Making thinking visible'. *American Educator*, 6 (11), 38–46. Accessed 23 July 2023. https://www.aft.org/ae/winter1991/collins_brown_holum.

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), 139–56. [CrossRef]

De La Paz, S. and Felton, M. (2010) 'Reading and writing from multiple source documents in history: Effects of strategy instruction with low to average high school writers'. *Contemporary Educational Psychology*, 35 (3), 174–92. [CrossRef]

De La Paz, S., Monte-Sano, C., Felton, M., Croninger, R., Jackson, C. and Piantedosi, K.W. (2016) 'A historical writing apprenticeship for adolescents: Integrating disciplinary learning with cognitive strategies'. *Reading Research Quarterly*, 52 (1), 31–52. [CrossRef]

Hallgren, K.A. (2012) 'Computing inter-rater reliability, for observational data: An overview and tutorial'. *Tutorials in Quantitative Methods for Psychology*, 8 (1), 23–34. [CrossRef]

Harris, L.M., Halvorsen, A.-L. and Aponte-Martínez, G.J. (2016) '"[My] family has gone through that": How high school students determine the trustworthiness of historical documents'. *Journal of Social Studies Research*, 40 (2), 109–21. [CrossRef]

Hicks, D., Doolittle, P.E. and Ewing, T. (2004) 'The SCIM-C strategy: Expert historians, historical inquiry, and multimedia'. *Social Education*, 68 (3), 221–5. Accessed 23 July 2023. https://go.gale.com/ps/i.do?p=AONE&u=googlescholar&id=GALE\TU\textbar{}A116451996&v=2.1&it=r&asid=15689948.

Huijgen, T., Van de Grift, W., Van Boxtel, C. and Holthuis, P. (2018) 'Promoting historical contextualization: The development and testing of a pedagogy'. *Journal of Curriculum Studies*, 50 (3), 410–34. [CrossRef]

Lee, P. (2004) 'Understanding history'. In P. Seixas (ed.), *Theorizing Historical Consciousness*. Toronto: University of Toronto Press, 129–64.

Lee, P. and Ashby, R. (2000) 'Progression in historical understanding among students ages 7–14'. In P. Stearns, P. Seixas and S. Wineburg (eds.), *Knowing, Teaching, and Learning History: National and international perspectives*. New York: New York University Press, 199–222.

Lévesque, S. and Clark, P. (2018) 'Historical thinking: Definitions and educational applications'. In S.A. Metzger and L.M. Harris (eds), *The Wiley International Handbook of History Teaching and Learning*. Hoboken, NJ: Wiley Blackwell, 119–48.

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), 539–68. [CrossRef]

Monte-Sano, C. (2012) 'What makes a good history essay? Assessing historical aspects of argumentative writing'. *Social Education*, 76 (6), 294–8. Accessed 23 July 2023. https://www.socialstudies.org/social-education/76/6/what-makes-good-history-essay-assessing-historical-aspects-argumentative.

Monte-Sano, C. (2016) 'Argumentation in history classrooms: A key path to understanding the discipline and preparing citizens'. *Theory Into Practice*, 55 (4), 311–9. [CrossRef]

Monte-Sano, C. and De La Paz, S. (2012) 'Using writing tasks to elicit adolescents' historical reasoning'. *Journal of Literacy Research*, 44 (3), 273–99. [CrossRef]

Monte-Sano, C., De La Paz, S. and Felton, M. (2014) *Reading, Thinking, and Writing About History: Teaching argument writing to diverse learners in the common core classroom, Grades 6–12*. New York: Teachers College Press.

Nokes, J. (2017) 'Historical reading and writing in secondary school classrooms'. In M. Carretero, S. Berger and M. Grever (eds), *Palgrave Handbook of Research in Historical Culture and Education*. London: Palgrave Macmillan, 553–71. [CrossRef]

Nokes, J.D. and De La Paz, S. (2018) 'Writing and argumentation in history education'. In S.A. Metzger and L.McArthur Harris (eds.), *The Wiley International Handbook of History Teaching and Learning*. Chichester: Wiley Blackwell, 551–78.

Nokes, J.D., Dole, J.A. and Hacker, D.J. (2007) 'Teaching high school students to use heuristics while reading historical texts'. *Journal of Educational Psychology*, 99 (3), 492–504. Accessed 23 July 2023. https://psycnet.apa.org/doi/10.1037/0022-0663.99.3.492. [CrossRef]

Patton, M.Q. (2015) *Qualitative Research & Evaluation Methods: Integrating theory and practice*. Thousand Oaks, CA: Sage.

Perfetti, C.A., Britt, M.A., Rouet, J.-F., Georgi, M.C. and Mason, R.A. (1994) 'How students use texts to learn and reason about historical uncertainty'. In M. Carretero and J.F. Voss (eds.), *Cognitive and Instructional Processes in History and the Social Sciences*. Hillsdale, NJ: Lawrence Erlbaum, 257–83.

Reisman, A. (2012) 'Reading like a historian: A document-based history curriculum intervention in urban high schools'. *Cognition and Instruction*, 30 (1), 86–112. [CrossRef]

Rouet, J.-F., Britt, M.A., Mason, R.A. and Perfetti, C.A. (1996) 'Using multiple sources of evidence to reason about history'. *Journal of Educational Psychology*, 88 (3), 478–93. Accessed 23 July 2023. https://psycnet.apa.org/doi/10.1037/0022-0663.88.3.478. [CrossRef]

Rüsen, J. (2004) 'Historical consciousness: Narrative structure, moral function, and ontogenetic development'. In P. Seixas (ed.), *Theorizing Historical Consciousness*. Toronto: University of Toronto Press, 63–85.

Seixas, P. (2017) 'Historical consciousness and historical thinking'. In M. Carretero, S. Berger and M. Grever, *Palgrave Handbook of Research in Historical Culture and Education*. London: Palgrave Macmillan, 59–69.

Seixas, P. and Morton, T. (2013) *The Big Six Historical Thinking Concepts*. Toronto: Nelson Education.

Sendur, K.A., Van Drie, J. and Van Boxtel, C. (2021) 'Historical contextualization in students' writing'. *Journal of the Learning Sciences*, 30 (4–5), 797–836. [CrossRef]

Stahl, S.A., Hynd, C.R., Britton, B.K., McNish, M.M. and Bosquet, D. (1996) 'What happens when students read multiple source documents in history?'. *Reading Research Quarterly*, 31 (4), 430–56. https://www.jstor.org/stable/748185. [CrossRef]

Stoel, G.L. (2017) 'Teaching Towards Historical Expertise: Developing students' ability to reason causally in history'. PhD thesis, University of Amsterdam, Amsterdam. Accessed 23 July 2023. https://hdl.handle.net/11245.1/1cfaaed8-9929-462e-b2e0-ad1945f1f034.

Stoel, G., Van Drie, J. and Van Boxtel, C. (2015) 'Teaching towards historical expertise: Developing a pedagogy for fostering causal reasoning in history'. *Journal of Curriculum Studies*, 47 (1), 49–76. [CrossRef]

Stoel, G.L., Van Drie, J.P. and Van Boxtel, C.A.M. (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), 321–37. [CrossRef]

University of Michigan. (2021) *Disciplinary Literacy Tools Structure, a Process for Inquiry and Argument Writing*. Accessed 23 July 2023. https://readinquirewrite.umich.edu/tools-structure/.

Van Boxtel, C. and 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.

Van Boxtel, C. and Van Drie, J. (2017) 'Engaging students in historical reasoning: The need for dialogic history education'. In M. Carretero, S. Berger and M. Grever (eds.), *Palgrave Handbook of Research in Historical Culture and Education*. London: Palgrave Macmillan, 573–89. [CrossRef]

Van Boxtel, C. and Van Drie, J. (2018) 'Historical reasoning: Conceptualizations and educational applications?'. In S.A. Metzger and L.M. Harris (eds.), *The Wiley International Handbook of History Teaching and Learning*. New York: Wiley Blackwell, 149–76.

Van Boxtel, C., Voet, M. and Stoel, G. (2021) 'Inquiry learning in history'. In R.G. Duncan and S.A. Chinn, *International Handbook of Inquiry and Learning*. New York: Routledge, 296–310. [CrossRef]

Van der Eem, M., Van Drie, J., Brand-Gruwel, S. and Van Boxtel, C. (2022) 'Students' evaluation of the trustworthiness of historical sources: Procedural knowledge and task value as predictors of student performance'. *Journal of Social Studies Research*. [CrossRef]

Van Drie, J. and Van Boxtel, C. (2008) 'Historical reasoning: Towards a framework for analysing students' reasoning about the past'. *Educational Psychology Review*, 20 (2), 87–110. [CrossRef]

Van Drie, J., Van Boxtel, C. and Van der Linden, J.L. (2006) 'Historical reasoning in a computer-supported collaborative learning environment'. In A.M. O'Donnell, C.E. Hmelo-Silver and G. Erkens, *Collaborative Learning, Reasoning and Technology*. London: Routledge, 266–97. [CrossRef]

Van Drie, J., Van Boxtel, C. and Stam, B. (2013) '"But why is this so important?" Discussing historical significance in the classroom'. *International Journal of Historical Learning, Teaching and Research*, 12 (1), 146–68. [CrossRef]

Van Drie, J., Braaksma, M. and Van Boxtel, C. (2015) 'Writing in history: Effects of writing instruction on historical reasoning and text quality'. *Journal of Writing Research*, 7 (1), 123–56. [CrossRef]

Van Nieuwenhuyse, K. (2020) 'From knowing the national past to doing history: History (teacher) education in Flanders since 1918'. In C. Berg, and T. Christou (eds.), *The Palgrave Handbook of History and Social Studies Education*. London: Palgrave Macmillan, 355–86. [CrossRef]

Van Nieuwenhuyse, K., Roose, H., Wils, K., Depaepe, F. and Verschaffel, L. (2017) 'Reasoning with and/or about sources? The use of primary sources in Flemish secondary school history education'. *Historical Encounters: A journal of historical consciousness, historical cultures, and history education*, 4 (2), 48–70. Accessed 23 July 2023. https://www.researchgate.net/publication/319955608_Reasoning_with_andor_about_sources_The_use_of_primary_sources_in_Flemish_secondary_school_history_education.

VanSledright, B. and Limón, M. (2006) 'Learning and teaching social studies: A review of cognitive research in history and geography'. In P. Alexander and P. Winne (eds.), *Handbook of Educational Psychology*. Mahwah, NJ: Lawrence Erlbaum, 545–70.

Voet, M. and De Wever, B. (2016) 'History teachers' conceptions of inquiry-based learning, beliefs about the nature of history, and their relation to the classroom context'. *Teaching and Teacher Education*, 55, 57–67. [CrossRef]

Voet, M. and De Wever, B. (2017) 'History teachers' knowledge of inquiry methods: An analysis of cognitive processes used during a historical inquiry'. *Journal of Teacher Education*, 68 (3), 312–29. [CrossRef]

Wilke, M. and Depaepe, F. (2019) 'Teachers and historical thinking: An exploration of the relationship between conceptualization, beliefs and instructional practices among Flemish history teachers'. *International Journal for History and Social Sciences Education*, 4, 101–35.

Wilke, M., Depaepe, F. and Van Nieuwenhuyse, K. (2022) 'Fostering historical thinking and democratic citizenship? A cluster randomized controlled intervention study'. *Contemporary Educational Psychology*, 71, 102–15. [CrossRef]

Wilke, M., Depaepe, F. and Van Nieuwenhuyse, K. (2023) 'Fostering secondary students' historical thinking: A design study in Flemish history education'. *Journal of Formative Design in Learning*, 7, 61–81. [CrossRef] [PubMed]

Wilschut, A., Van Straaten, D. and Van Riessen, M. (2012) *Geschiedenisdidactiek: Handboek voor de Vakdocent*. Bussum: Coutinho.

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. [CrossRef]

Wineburg, S. (2001) *Historical Thinking and Other Unnatural Acts: Charting the future of teaching the past*. Philadelphia: Temple University Press.

Young, K.M. and Leinhardt, G. (1998) 'Writing from primary documents: A way of knowing in history'. *Written Communication*, 15 (1), 25–68. [CrossRef]

https://doi.org/10.14324/HERJ.20.1.05