Students' evaluation of the trustworthiness of historical sources: Procedural knowledge and task value as predictors of student performance

van der Eem, M.; van Drie, J.; Brand-Gruwel, S.; van Boxtel, C.

DOI
10.1016/j.jssr.2022.05.003

Publication date
2023

Document Version
Final published version

Published in
The Journal of Social Studies Research

Article

Journal of Social Studies Research
2023, Vol. 47(1) 64–76

DOI: 10.1016/j.jssr.2022.05.003
journals.sagepub.com/home/ssr
Sage

Students' evaluation of the trustworthiness of historical sources: Procedural knowledge and task value as predictors of student performance*

Maartje van der Eem[1], Jannet van Drie[1], Saskia Brand-Gruwel[2], and Carla van Boxtel[1]

Abstract

Evaluating the trustworthiness of sources is important in today's society. However, research has shown that students struggle when applying this skill. This study in history education aims to gain insight into students' procedural knowledge about evaluating the trustworthiness of sources and into the value students attach to learning this skill. Grade 9 students (N = 132) performed tasks and filled out a questionnaire. Students applied more correct criteria of trustworthiness than they reported knowing. They considered this skill somewhat important and useful, but less interesting. Procedural knowledge and task value were significant predictors of students' task performance. Therefore, it is important to make students aware of the knowledge they implicitly possess and to make learning this skill more interesting for students.

Keywords

Historical reasoning, Trustworthiness, Student performance, Procedural knowledge, Task value, Secondary education

1. Introduction

Critically evaluating information is considered to be an important skill in our digitalized and democratic society (e.g., De La Paz & Felton, 2010; Nokes, 2013). The Council of the European Union (2018), for example, has distinguished eight competences for lifelong learning, and the critical evaluation of information is connected to three of these eight competences. Therefore, appeals are made for education to teach students this skill, so they will be able to better recognize fake news and disinformation (European Commission, 2018; McGrew et al., 2017). Although McGrew et al. (2019) have argued that evaluating the trustworthiness of online information requires a different approach than historical sources - online a student actively has to search for the information needed to decide whether to trust a source - the criteria of trustworthiness McGrew et al. mentioned are similar to criteria students are taught in history class: who is behind the information, what is the evidence, and what do other sources say (e.g., De La Paz & Felton, 2010; Reisman, 2012)? Since teaching how to evaluate the trustworthiness of historical sources to answer a specific question is part of the history curriculum in secondary education in many countries, history classrooms are thus an appropriate place to teach students this skill.

However, despite successful interventions which showed improvement in the application of the skill of evaluating the trustworthiness of historical sources (e.g., Britt & Aglinskas, 2002; Nokes et al., 2007; Reisman, 2012), today's students in secondary education still seem to struggle when they have to apply this skill in school tasks in which different types of historical sources such as firsthand accounts and statistical information are used (Harris et al., 2016; Jacobsen et al., 2018; Nokes, 2017). This raises the question of why students still struggle. Is it simply a matter of spending more time practicing this skill in the classrooms? Or are there are other reasons that explain why students perform poorly on questions in which they have to apply this skill?

[1]Research Institute of Child Development and Education, University of Amsterdam, the Netherlands
[2]Zuyd, University of Applied Sciences, the Netherlands

*This work was funded by Dudoc Alfa (grant number DA2-2018-13). Dudoc Alfa had no involvement in the research and in the article.

Received 17 December 2021; revised May 24 2022; revised manuscript accepted 26 May 2022

Corresponding Author:
Maartje van der Eem, Research Institute of Child Development and Education, University of Amsterdam, the Netherlands.
Email: m.vandereem@uva.nl

In the present study, we researched the extent to which students' procedural knowledge and the value students attach to learning this historical skill predict student performance on tasks in which they evaluate the trustworthiness of historical sources.

Insight into students' procedural knowledge may add to our understanding of why students struggle with the skill. Being able to apply a skill like this requires procedural knowledge: knowledge about the way history "is done" (Levesque & Clark, 2018). An element of this procedural knowledge is knowing which criteria of trustworthiness must be applied, for example the criterion of the goal of the maker of the source. This part of students' procedural knowledge has been studied among students evaluating the trustworthiness of internet sources (e.g., Walraven et al., 2009) but not when applying this skill on historical sources.

Research in several other domains has shown that task value (interest in and the importance and usefulness of a specific task) is positively correlated with students' effort when performing tasks (e.g., Kiuru et al., 2020). This raises the question of how useful, important, and interesting students consider the skill of evaluating the trustworthiness of historical sources and whether this predicts their task performance.

In the present study, we first analyzed grade 9 students' task performance: their answers to tasks in which they evaluated the trustworthiness of historical sources (the quality of their answers and the extent to which they use correct criteria of trustworthiness), their procedural knowledge, and the value they attach to learning this historical thinking skill. Students from six different schools participated in the research, so it is possible to control for school-specific effects. The second part of the article focuses on the question to what extent procedural knowledge and task value predict task performance. The results of this study lead to suggestions for improving the teaching of this skill in secondary history education.

2. Theoretical background

2.1. Evaluating the trustworthiness of historical sources

In secondary history education, students are not only taught about the events in the past, but they are also taught how to reason historically. In their framework of historical reasoning, van Drie and van Boxtel (2008) described the six components of historical reasoning; one of those is the use of sources. Evaluating the trustworthiness of sources in relation to a question about the past is part of that specific component of historical reasoning in order to determine whether a source can be used as evidence to answer that question about the past (van Boxtel & van Drie, 2018; Nokes, 2013). This entails that students not only have to mention, for example, that a source about an event in the Second World War was made by an eyewitness, but also have to explain why that could make the source more or less trustworthy.

When evaluating the trustworthiness of sources, historians use three heuristics, which Wineburg (1991) described in his landmark study: sourcing, contextualization, and corroboration. Since then, these three heuristics have been used in many studies on evaluating the trustworthiness of historical sources (e.g., Breakstone, 2014; Britt & Aglinskas, 2002; Nokes et al., 2007; Reisman, 2012; Seixas & Morton, 2013).

The heuristic of sourcing consists of several criteria. The following four criteria are mentioned in most studies, although sometimes phrased slightly different: the position and background of the maker (and/or publisher) of the source, the goal of the maker, how the maker came to know about the events described in the source, and the time in which the source was made (Harris et al., 2016; Jacobsen et al., 2018; Nokes, 2013; Perez et al., 2018; Perfetti et al., 1994; Reisman, 2012; Seixas & Morton, 2013). Corroboration refers to the comparison of different sources (what do other pieces of evidence tell you), and contextualization refers to the historical context of the time in which the source was made (Reisman, 2012; Wineburg, 1991).

Wineburg (1991) showed that historians and high school students differed greatly in their approach when asked to answer a historical question for which they had to evaluate the trustworthiness of historical sources. The students mainly looked at the content of a source, while the historians used the three heuristics. Subsequent research in secondary education has confirmed these results. Some of this research mainly focused on whether students evaluated the trustworthiness of historical sources using the three heuristics; other research focused on all the criteria students used, including the incorrect ones, when asked to evaluate the trustworthiness of sources.

Examples of the first type of research, focused on whether students used the heuristics, are Britt and Aglinskas (2002), Halvorsen et al. (2016), Nokes (2017), and Rantala and van den Berg (2015). Britt and Aglinskas found that grade 11 students scored low on spontaneous use of source information. Halvorsen et al. found that sourcing was the weakest element in grade 11 students' essays. Although students sometimes mentioned information about the source in their essay, they did not discuss the influence of a source characteristic on the trustworthiness of the information in the source. Rantala and van den Berg concluded that the sourcing skills of Finnish students in grade 12 were comparable to those of the students in the Wineburg (1991) study.

Nokes (2017) also focused on the question whether students used the heuristic of sourcing, but performed his research amongst younger students (grade 8). The answers of the students in his research showed considerable differences in both the quantity of the sourcing they applied and

the quality of the sourcing arguments students used. Over 40% of the students did not write about the source of the document at all, and thus performed at the lowest level of sourcing. At the same time, almost 40% of the students performed at level 4 (the second highest of the levels Nokes distinguished), which means that they were able to evaluate the trustworthiness of the sources but did not use that information in their written arguments.

While the above-mentioned studies focused on whether students applied the historical reasoning skill, other studies were more concerned with the specific criteria students used when explaining why they considered a source trustworthy or not. Examples of this second type of research are two studies performed in grade 11 and 12 in the United States. Harris et al. (2016) found that students often used the content of the source to explain whether a source was trustworthy or not. Jacobsen et al. (2018) found that students often used arguments of skepticism (such as: "people are often biased") or blind faith (such as: "experts have the most knowledge") when asked to explain their answer whether they considered a source trustworthy. These criteria are, in fact, related to sourcing, to the criterion of who made the source. However, the students who use criteria of skepticism or blind faith do not take into account the specific characteristics of the source they are evaluating and use generalized statements in their argumentation instead. Both studies found that students also used their own emotions (for example, considering the content of a source to be "not fair") as a criterion for (dis)trusting a source.

In addition to the abovementioned descriptive research, there are also intervention studies on this topic, which showed that although high school students are able to improve the application of the historical skill of evaluating the trustworthiness of sources, it still remains a difficult skill to master for students in this age group. Nokes et al. (2007) concluded after an intervention in which students were taught to use the three heuristics when reading historical sources that students rarely used the heuristic of contextu-alization, while a significant intervention effect was found on the use of sourcing and corroboration. Reisman (2012) also found a significant intervention effect on sourcing, but not on corroboration and contextualization. The authors offered possible explanations for their results: perhaps students lack the background knowledge that is necessary for contextualization (Nokes et al., 2007) or perhaps sourcing is a less sophisticated heuristic and therefore easier to learn (Reisman, 2012).

The results of these intervention studies thus show that if students are explicitly taught the skill of evaluating the trustworthiness of historical sources, they improve their procedural knowledge of sourcing. This suggests that students, once they know the correct criteria, improve on applying the skill when asked to do so. However, research in the domain of internet sources has shown that students are aware of the criteria they must apply, but they just do not apply them enough. In two studies on internet sources, grade 9 students were interviewed about the criteria of trustworthiness they knew. In the first study, Dutch students used only a small number of the criteria when evaluating the trustworthiness of information from the internet, but afterwards, they were able to mention more criteria than they had used in the tasks (Walraven et al., 2009). The second study, focused on German and French students, also concluded that the students did possess knowledge on how to use the sourcing heuristic, but did not always apply that knowledge, for several reasons (Paul et al., 2017).

2.2. Task value

Paul et al. (2017) asked grade 9 students why they had not used the heuristic of sourcing in tasks involving internet sources In 50% of the interviews, their answer was related to motivation. Students said, for example, that they thought it was not important and thought it required a great deal of effort. These students' comments refer to task value, which is concerned with the question "what do I think of this task" (Pintrich et al., 1991). Pintrich et al. distinguished three aspects of task value: interest, importance and usefulness. Interest refers to the extent to which a student enjoys performing a certain task, importance refers to performing well on a task, and usefulness refers to the relationship between the task and future goals of a student (Wigfield & Eccles, 1992).

A student' s belief about the importance and usefulness of a task and their interest in the task might predict a student' s learning intentions, the way a student is cognitively engaged in the task, and the effort put into the task (Ainley, 2006; Eccles & Wigfield, 1995; Pintrich & Schrauben, 1992; Wigfield & Cambria, 2010a). Multiple studies have confirmed this assumption. Velayutham and Aldridge (2013), for example, performed their research in grade 8 to 10 in the domain of science and found that task value was one of the predictors of student's self-regulation. Kiuru et al. (2020) found that a high score on task value was positively correlated with grade 6 students' effort during the tasks. Cole et al. (2008) found a similar correlation between scores on usefulness and importance and undergraduate students' test-taking effort on a low stakes test. The correlation between interest and test effort in social studies was lower compared to importance and usefulness, but it was still a medium-sized correlation. Wigfield and Cambria (2010b) stated that interest has an influence on deep-level learning, more than on surface-level learning, which might explain the findings of Cole et al.

The above-mentioned research thus indicates that task value can influence students' learning behavior and, indirectly, also student performance (Kiuru et al., 2020). There is some information on how students value applying the

skill of evaluating trustworthiness of internet sources (Paul et al., 2017) but when it comes to evaluating the trustworthiness of historical sources, this has not yet been studied. Within the domain of history, Halvorsen et al. (2016) did ask students about their interest, but focused on the historical content of the tasks, and not on the historical reasoning skills needed for the task.

3. Research questions

In the present study, the following research questions will be addressed:

1. How do students perform on tasks in which evaluating the trustworthiness of historical sources is incorporated?
2. To what extent are students able to mention the criteria for evaluating the trustworthiness of historical sources (procedural knowledge)?
3. How do students value learning the skill of evaluating the trustworthiness of sources (task value)?
4. To what extent do procedural knowledge and task value predict the quality of students' task performance?

4. Method

4.1. Participants

This study was conducted in six secondary schools in the Netherlands. Three schools are located in a big city, one in a smaller city, and two in a village. From each school, one grade 9 class participated (total $N = 156$). Grade 9 is the final year of lower secondary education (grades 7-9) in the Netherlands, during which all students are obliged to follow the history curriculum. Students must comply with the final attainment levels for lower secondary education, which includes being able to evaluate the trustworthiness of historical sources. All students attended higher general secondary education, which prepares them for the university of applied sciences. Twenty-four students were excluded due to absence during class or a lack of parental consent; the sample for this study therefore consisted of 132 students. The average number of participating students per class was 22 (SD: 1.79). The average student age was 14.5 years (SD: 0.60) and 51.5% of the students were girls. Their history teachers' teaching experience varied between four and thirty-eight years (M: 15.17, SD: 13.08).

4.2. Context

In the Netherlands, the history of the 20th century is taught in grade 9. The schoolbooks not only pay attention to the important events of the 20th century, but also to the components of historical reasoning. Each schoolbook chapter contains several tasks focused on practicing these skills, such as evaluating the trustworthiness of historical sources. Schools can make their own decisions in how they use these schoolbooks and can differ from each other in, for example, their emphasis on specific parts of the history in the 20th century and the ratio between time spent on the historical events described in the schoolbooks and on the historical reasoning skills, since there is no standardized test at the end of lower secondary education.

When asked about the extent to which the participating teachers practiced this specific historical reasoning skill with their students, all teachers indicated that they used tasks in which students had to evaluate the trustworthiness of historical sources, both during lessons and in summative assessments. However, there were differences between the schools in the amount of time spent on teaching this skill and differences in the way the teacher paid attention to this skill. The teacher of school 4 indicated that the history curriculum in her school focused primarily on teaching students the historical reasoning skills. Students of this school had received more explicit instruction and had had more practice in applying these skills, such as evaluating the trustworthiness of sources, than the students in the other five schools. The teacher of school 1 said that he had given explicit instruction about sourcing and the students were supposed to take notes during this instruction, which took place during one lesson. In other lessons, he regularly referred to this instruction. The other four teachers (school 2, 3, 5 and 6) said that they did mention the basic sourcing criteria such as the time in which the source was made and the maker of the source, but apart from that, they did not provide explicit instruction to their students about this skill. Therefore, it is important to control for school effects when analyzing the results.

4.3. Instruments

4.3.1. Task performance: Historical Assessments of Thinking. Tasks were designed to measure two aspects of students' task performance: the quality of their answers (measured as the score on the tasks) and the extent to which the correct criteria were applied when students evaluated the trustworthiness of sources in relation to a specific historical question. These tasks were based on the Historical Assessments of Thinking (HAT's) designed by the Stanford History Education Group (Breakstone, 2014; Smith et al., 2018). HAT's are short, open-ended tasks using historical sources. The HAT-format was chosen because these assignments directly target the skill under study instead of being part of a larger assignment such as a document-based question, which means that the results are a better representation of students' capabilities (Breakstone, 2014).

In this study, five different historical sources about the First and the Second World War were used, with tasks based

on the HAT formats. We replaced the sources used in the original HAT' s by sources suitable for the Dutch context in grade 9. The phrasing of the tasks was adjusted according to the new sources (in Appendix A, an example of a task can be found). With the exception of one source, the students had to perform several tasks per source, for example explaining why a certain source contained both elements that could be considered trustworthy and elements that might not be trustworthy, leading to a total of fifteen tasks.

The tasks were designed to elicit the heuristics of corroboration and sourcing. The heuristic of corroboration means that students had to compare sources. For the heuristic of sourcing, four criteria were used: (1) the position of the maker of the source, (2) the goal of the maker, (3) the information the maker used to create the source, and (4) the time in which the source was made. Each criterion was the ideal criterion of a task at least twice.

The tasks were piloted in three rounds in three different classes. These classes did not participate in this study. In each pilot round, students read five sources and performed the fifteen tasks. In round 1, two alternative sources were tested but were rejected because they did not elicit students' thinking well enough. After each test round, some small adjustments were made, for example in the phrasing of a task. In the third round of the pilot, the final version of the tasks was used. After this last testing round, a researcher from a different university provided final feedback on the tasks.

4.3.2. Procedural knowledge: open-ended question. To measure students' procedural knowledge of the skill, students were asked to write down all the criteria of evaluating the trustworthiness of a historical source they knew.

4.3.3. Task value: questionnaire. To measure the value students attach to learning and applying the historical reasoning skill of evaluating the trustworthiness of historical sources, a part of the Motivated Strategies for Learning Questionnaire (Pintrich et al., 1993) was used. This Task Value Questionnaire (TVQ) consists of six statements on a seven-point Likert scale, with statements such as "it is important for me to learn this skill". The statements were translated into Dutch. The questionnaire was tested in one class (not participating in the study); no adjustments were necessary. The six statements form one scale; Cronbach's alpha was .86, which is considered a good value (Field, 2018).

4.4. Procedure

Since this research focused on students' abilities within the present curriculum, the participating teachers were not asked to prepare their students, for example by giving extra lessons on the skill under study. The students performed the tasks with paper and pencil, except for fifteen students, who performed the tasks while thinking aloud. These students, three per school (from the schools 2 through 6), were selected by their teachers based on their history grades (good, average, poor). The think-aloud method was used to uncover students' procedural thinking while performing the tasks (Ericsson & Simon, 1980). However, during analysis it became clear that the transcriptions did not provide additional information to the written answers of the other students, since the think-aloud students answered the questions in the same manner as the paper-and-pencil students, that is, by *not* mentioning the procedures they applied to reach their answers, even though they were specifically asked to do so.

The tasks were presented to the students in three different orders to prevent a possible influence of the formulation of the tasks on the students' answers. A Kruskal-Wallis test was performed, to check whether the order in which the tasks were presented influenced student performance. The results showed no significant differences between the average scores of the three different orders in which the tasks were presented to the students, to which the think-aloud students were added as a fourth group, $H(3) = 0.28$, $p = .96$.

The students were allowed a maximum of 40 min to complete the tasks, which was enough time. After handing in the tasks, the students answered the open-ended question measuring their procedural knowledge and they filled out the TVQ. The think-aloud students were asked to explain their answers on the TVQ. The information from the transcriptions could be used to provide more in-depth information on the outcomes of the TVQ.

4.5. Data analysis

4.5.1. Task performance. Students' answers to the tasks were analyzed on two levels: the quality of their answers (task score) and the criteria students used to support their answers. To determine the quality of their answers, a rubric was developed (see appendix B for an example of a rubric). A score was awarded for each answer. For an incorrect answer, a student received zero points; for a sufficient answer, one point was awarded, and for a good answer, two points were awarded. The difference between a sufficient answer and a good answer was (a) the way in which a student elaborated on the criterion used, or (b) the choice and explanation of the best possible criterion if more than one criterion was applicable. For each student, a final score was calculated by adding up the scores of the fifteen tasks, which could lead to a maximum score of 30 points. A total of 450 answers from 30 students (22.7% of the sample) were independently scored by the first author and a research assistant. From each school, the first five students In alphabetical order were scored. There was 89.6% agreement between the two raters; Cohen's kappa was .81. A score of > 0.80 is considered to be an almost perfect agreement

(Landis & Koch, 1977). The first author scored the remaining answers.

For each answer, the criterion used by the student was identified. Although the tasks required the students to name and explain one criterion per answer, students sometimes used more than one criterion. When a student had applied more than one criterion correctly, the most difficult one in the context of that specific task was chosen because that best showed the student's abilities. For example, one student wrote: "[it is trustworthy] because she is an eyewitness, and she wrote it down just for herself". This student used two criteria: the information the author used (eyewitness) and the goal of the author (the author wrote it in her diary and therefore the information was not meant to be shared). The second criterion the student used was scored, as it was more difficult because it required one more thinking step than the first. The first author and a research assistant independently categorized the criteria used in 450 answers from 30 students (22.7% of the sample), with 86% agreement. Cohen's kappa for interrater reliability was 0.84. The first author categorized the criteria in the remaining answers.

4.5.2. Procedural knowledge. The criteria the students mentioned in the open-ended question to measure their procedural knowledge were assigned to the five correct criteria that could be used (the four criteria of sourcing and the one criterion for corroboration). For example, when a student had written down "who made the source", it was categorized under "position of the author". The maximum score was five points. The first author and a research assistant independently categorized the answers of 30 of the 122 students who had answered this question (24.60% of the sample). There was 92% agreement; Cohen's kappa was .90.

4.5.3. Statistical analysis. First, descriptive statistics were used to assess task performance, students' procedural knowledge, and the value they attached to working with this skill (research questions 1-3). Second, a hierarchical multiple regression analysis was used to predict the influence of the independent variables procedural knowledge and task value on the outcome variable task performance, measured as students' task score (research question 4). For the participating schools, dummy variables were created. School 4 was used as the baseline group because student performance in terms of average score on the tasks was significantly different from the other five schools in a regression analysis in which only the dummy variables for the schools were entered.

In the descriptive part of the results, the results of all 132 participating students were used in the analyses. Ten of the think-aloud students were not asked to answer the open-ended question used to measure students' procedural knowledge, so they were removed from the sample in the regression analysis.

Table 1. Task performance per school (scale: 0-30).

	M	SD
All students (N = 132)	8.54	4.97
School 1 (N = 20)	9.70	4.18
School 2 (N = 22)	6.23	5.41
School 3 (N = 20)	8.55	4.59
School 4 (N = 24)	13.08	4.20
School 5 (N = 22)	5.68	3.64
School 6 (N = 24)	7.75	4.04

5. Results

5.1. Task performance: quality of the answers and the use of correct criteria

A total of 86.4% of the students earned less than half of the maximum 30 points for the tasks. The median score was 8.00 and the scores ranged from 0 to 24. In Table 1, the average scores, in total and per school, are presented.

The results in Table 1 show that students did not score well on the tasks. The students of school 4 clearly outperformed the students from the other five schools. The large standard deviations in Table 1 illustrate the wide variation in student performance.

In 995 of the 1980 student answers (50.26%), a correct criterion for trustworthiness was applied. When looking at student level, 4.5% of the students used only one of the five criteria of trustworthiness in their answers, 12.1% used two criteria, 28.8% used three criteria, 36.4% used four criteria, and 18.2% of the students used all five criteria at least once in their answers. On average, the students used 3.52 of the correct criteria (SD: 1.07).

5.2. Procedural knowledge

When asked to write down the criteria of trustworthiness they knew, students mentioned on average 2.71 criteria (SD: 1.41); the number of criteria ranging between 0 and 7. However, not all criteria mentioned by the students were correct. Some students, for example, wrote down "pictures" as a criterion, or "the language in a source must not be too difficult". When looking only at the correct criteria, 15.6% of the students did not write down any of these five criteria, 24.6% wrote down one criterion, 38.5% two criteria, 19.7% three criteria, and 1.6% four criteria. None of the students wrote down all five criteria. In Table 2, the average number of correct criteria mentioned on the open-ended question measuring procedural knowledge is presented and compared with the average number of correct criteria the students used in their answers to the tasks.

The students of school 4 had the best quality of answers on the tasks in terms of score and applied on average the most correct criteria, but the students of school 1 reported the highest number of correct criteria on the open-ended

Table 2. Procedural knowledge: average number of correct criteria mentioned versus average number of applied correct criteria in the tasks (scale: 1–5).

	Procedural knowledge		Use of criteria in tasks	
	M	SD	M	SD
All students	1.67	1.02	3.52	1.07
School 1	2.70	.57	3.85	.67
School 2	1.10	.85	2.68	1.09
School 3	1.44	.86	3.45	1.05
School 4	2.27	.99	4.17	.92
School 5	1.10	.85	3.41	1.05
School 6	1.36	.79	3.50	1.02

Table 3. Scores on the TVQ (Likert scale 1-7), N = 132.

	M	SD
1. I think I can use this skill in other subjects as well.	4.45	1.45
2. It is important for me to learn this skill.	4.60	1.55
3. I am very interested in this skill.	3.78	1.62
4. I think it is useful for me to learn this skill.	4.83	1.51
5. I like working with this skill in history class.	3.51	1.57
6. Understanding how to apply this skill is very important to me.	5.01	1.43
Scale average all students	4.37	1.16
Scale average school 1	4.71	1.00
Scale average school 2	4.16	1.29
Scale average school 3	3.70	1.13
Scale average school 4	4.81	.92
Scale average school 5	4.18	1.22
Scale average school 6	4.61	1.14

question. The students of school 2 and school 5, who had the lowest scores on the tasks, also wrote down the lowest number of correct criteria on the open-ended question.

5.3. Task value

In the TVQ, students were asked to indicate the value they attached to learning the skill of evaluating the trustworthiness of historical sources. The results per statement and the scale average (overall and per school) are presented in Table 3.

On average, students rated the task value moderately positive (4.37 out of 7). The students of the two schools with the highest average scores (school 4 and 1) also answered most positive on the TVQ. The students of school 3, the school whose average task score was practically equal with the overall average task score, were the least positive on the TVQ.

The answers on importance (statements 2 and 6) and usefulness (statements 1 and 4) were more positive than students' answers on the statements about interest in the historical reasoning skill (statement 3 and 5). On average, these last two statements were answered negatively.

5.3.1. Think-aloud students' comments on importance, usefulness, and interest. In order to gain more insight into students' thinking about the importance, usefulness, and interest of learning to evaluate the trustworthiness of historical sources, the fifteen think-aloud students were asked to explain their answers on the TVQ. These fifteen students can be considered a representative part of the total sample, since there were no significant differences between these think-aloud students and the other students on both average task performance (t (130) = -0.161, $p > .05$) and on the TVQ (t (130) = -1,673, $p > .05$).

Whereas most of these students explained their positive score on statement 1 by the fact that they also have to use this skill in subjects such as geography, one student rated this statement with a two and explained that "other subjects, for example geography, are concerned with the present, so everything is written down and is certain. Only in history, you can have doubts because it happened in the past, and in the past, not everything was written down".

Regarding statements 2, 4 and 6 (importance and usefulness), some students referred to school tests to explain why it is important to learn this skill, for example: "[this skill] is important, but other than using it for school tests, I do not think it is important". Other students did not explicitly refer to the use of this skill in the context of school: "[it is important] because, well, if you are able to evaluate sources correctly, then you can also judge people" and "you have to be able to draw a conclusion about whether something is fake or real". Only one student referred explicitly to a situation outside the school context: "after I will have finished school, it will also be useful not to trust everything blindly. If something is written in a newspaper, you still are not sure whether it is true. You have to find other sources as well".

Students were less positive about statement 3 and 5. With regard to statement 3, the students mentioned different reasons to explain why they were not interested in this skill. One of the students said that he did not like to read, referring to the amount of reading applying this skill usually entails. Another student said that she was not interested in this skill because she always receives bad grades on history tests with questions involving sources. A third student explained that he was not interested because he could not think of other situations than a school test where he would have to use this skill. The students who were more interested in applying the skill referred to the process they liked ("the better you think, the more you think") or the fun of being able to distinguish real from fake.

Table 4. Pearson's correlation (two-tailed) among key variables (N = 122).

	Task performance	Procedural knowledge	Task value
Task performance	1.00		
Procedural knowledge	.397***	1.00	
Task value	.344***	.236**	1.00

*$p < .05$, **$p < .01$, ***$p < .001$.

On average, students' answers were negative when asked whether they liked working with this skill in history class (statement 5). One of the students said: "I do not like it very much. We must read the sources over and over again, and then I have to explain why. And giving arguments, those kinds of things". A second student: "no, I do not like it. I think it is difficult and hard to evaluate the trustworthiness". Students who answered more positively provided different reasons. One student said that he liked to investigate whether something was true or not, while one of his classmates said that he answered positive because it is a different activity than the usual learning activities in history class. One student hesitated, saying that it depended on the content of the source. If that was a topic she found interesting, then she liked working with this skill in history class. But if the content of the sources would not appeal to her, it would be boring to evaluate the trustworthiness, she said.

5.4. Predictors of task performance

The fourth research question is concerned with the extent to which the quality of a student's answer (task performance) can be predicted by two variables on student level, namely procedural knowledge and task value, controlled for the school a student attends. In Table 4, the correlations between the key variables used in the regression analysis are presented.

The correlation analysis shows that there is a significant positive relationship between task performance and (1) students' procedural knowledge and (2) the value students attach to learning the skill. According to the guidelines of Cohen (1988), these are both medium-sized correlations.

A multiple linear regression analysis was conducted to predict students' task performance. The results are presented in Table 5. In model 1, procedural knowledge and task value were entered as independent variables (F (2,129) = 17.162, p = .000). In model 2, the school dummies were added. School 4, the highest scoring school, was used as the reference category in model 2. This second model is an improvement compared to the first model: F (7,124) = 8.350, p = .000.

Procedural knowledge and task value are significant positive predictors for the quality of a student's answer in both models, although the level of significance for procedural knowledge decreases in the second model. The B-coefficient of task value is rather similar in both models, whereas the B-coefficient of procedural knowledge decreases with about one-third in the second model. The coefficients for the dummy's school are all negative because school 4 was used as the baseline group and this school was, by far, the best performing school. The indicator school is a significant predictor of students' task performance. Procedural knowledge, task value, and the school of the student together explained 33.9% of the variance in the quality of the answers the students had given on the tasks.

Table 5. Multiple regression analysis on task performance (N = 122).

	B	*SE B*	β	
Model 1				
Constant	.757	1.603		
Procedural knowledge	1.653	.411	.335	***
Task value	1.145	.360	.265	**
Model 2				
Constant	5.695	2.134		
Procedural knowledge	1.045	.474	.212	*
Task value	1.045	.357	.241	**
School 1	3.737	1.316	-.277	**
School 2	4.590	1.431	-.340	**
School 3	2.703	1.443	-.192	
School 5	5.269	1.430	-.390	***
School 6	4.749	1.338	-.365	**

*$p < .05$, **$p < .01$, ***$p < .001$.
Note: R^2 for step 1 = 0.224, AR2 = 0.115 for step 2 ($p < .01$).

6. Discussion

In this study, we examined the performance of grade 9 students, who are 14 or 15 years old, on tasks in which they had to evaluate the trustworthiness of historical sources (research question one), their procedural knowledge of the criteria of trustworthiness (research question two), the value students attach to learning this skill (research question three), and the predictive value of the variables procedural knowledge and task value for students' task performance (research question four). The first research question has been described before in the field of history in secondary education; the other three research questions have - to the best of our knowledge - not yet been researched.

6.1. Task performance

Based on the results of the present study and previous research (Britt & Aglinskas, 2002; Halvorsen et al., 2016; Harris et al., 2016; Jacobsen et al., 2018; Nokes, 2017; Wineburg, 1991) it seems that within a standard history curriculum - thus without intensive and more explicit teaching of the skill, as was done in the intervention studies - students are not taught to evaluate the trustworthiness of historical sources in such a manner that they can apply the skill well enough. Students from five of the six participating schools scored on average less than one third of the points that could be earned on the tasks. The students from the best performing school had spent more time practicing this skill than the students in the other five schools, but even this best scoring school still scored less than half the points.

At the same time, it is important to note that, when comparing the results of this study with the results of other studies in order to draw conclusions about student performance, different methods and age groups are compared with each other. Most studies were performed in grade 11 and 12 (Britt & Aglinskas, 2002; Halvorsen et al., 2016; Harris et al., 2016; Rantala & van den Berg, 2015), so the participating students were older than the grade 9 students in this study. Only the study performed by Nokes (2017) also included students in middle school (grade 8).

The tasks which were given to the students also differed: a document-based question (Halvorsen et al., 2016), a short document-based question combined with follow-up questions (Nokes, 2017), and a ranking task (Harris et al., 2016; Jacobsen et al., 2018). The tasks used in the present study, the HAT's, are more directly aimed at testing the historical reasoning skill (Breakstone, 2014) than, for example, document-based questions, in which students have to use many skills at the same time (Smith et al., 2018). Despite these different task approaches and the age differences, a rather constant finding seems to be the great variation in student performance when evaluating the trustworthiness of historical sources (Jacobsen et al., 2018; Nokes, 2017). In the present study, the standard deviation of 4.97 on the average student score of 8.54 also illustrates the variation in student performance.

Although studies that focused on the arguments that students used to support their answers did not provide exact numbers on the criteria students applied, they describe that often students used the content of the source to justify their answers (Harris et al., 2016; Nokes, 2017; Rantala & van den Berg, 2015). In the present study, the different criteria students used were categorized and counted. And although the average student score is low, the students did use correct criteria of trustworthiness in approximately 50% of their answers. This seems to be a rather high percentage, compared to the other research mentioned above, even though the scores on the tasks seem to suggest otherwise. However, this student performance regarding the criteria applied in the tasks could also be a result of using the HAT format in the present study, instead of more complicated tasks. It thus seems that this HAT-format should not be used only to award a score to an answer and then categorize a student, for example, at basic or proficient level, but also to analyze which criteria students used to obtain a more complete picture of students' capabilities.

6.2. Procedural knowledge

By comparing students' procedural knowledge (which criteria do students report knowing) with the actual application of criteria, this study contributes to the existing body of literature on evaluating the trustworthiness of historical sources in which the focus has only been on students' actual performance on tasks. The students in the present study applied on average more correct criteria than they had acknowledged knowing, which is in contrast with the results of the research on internet sources (Paul et al., 2017; Walraven et al., 2009). A possible explanation for this difference might be the context of the tasks: in the present study, students were asked to perform tasks with which they were more familiar in an educational context, which perhaps made it easier for them to apply correct criteria.

The results of the present study suggest that students implicitly know which criteria are correct, but are not explicitly aware of their knowledge. This was also observed during the think-aloud sessions: none of the fifteen students ever mentioned the procedure they were applying, even though they were specifically asked to do so at the start of the session. They did not seem to realize that there *are* procedures to be followed, such as looking at the author and the date of the source.

The results of the regression analysis showed that procedural knowledge is a significant positive predictor of task performance. However, the results also showed that once the school dummies were added as a variable in the regression, the B-coefficient decreased with about one-third. Thus, a student's procedural knowledge is quite largely effected by the school a student attends.

6.3. Task value

Besides students' procedural knowledge, this research added a second variable to the existing literature on the historical reasoning skill under study: task value. The results of this study are in line with the findings in the studies of Paul et al. (2017) and Walraven (2009), in which students, who were also in grade 9, were interviewed about evaluating the trustworthiness of internet sources. The students in the present study considered the skill of evaluating the trustworthiness of sources more important and useful than interesting, although the scores on usefulness and importance were at

best moderately positive. The negative scores on the two statements about interest are worrisome, since interest has an influence on students' learning (Wigfield & Cambria, 2010b).

The regression analysis showed that task value is a significant positive predictor of task performance, which is in line with previous research performed on task value in other domains (e.g., Cole et al., 2008). In contrast with procedural knowledge, task value seems less dependent on the school a student attends, since adding the school dummies did not change the level of significance and only slightly changed the B-coefficient. So task value is more student dependent. This was illustrated by the diverse explanations these think-aloud students gave as a clarification of their answers on the TVQ, such as an aversion to reading, failures in the past when having to apply this skill, thinking it is only needed on history tests or liking the process of investigating sources.

6.4. Limitations

The findings of this study must be seen in the light of some limitations. First, the think-aloud sessions did not provide extra information in comparison to the written answers on the tasks of the other students. We might have received more useful information if we had interviewed some students after they performed the tasks on paper and then asked them about the choices they had made and the procedures they had followed, a method used by, for example, Smith et al. (2018). Employing retrospective think-aloud sessions would have added more valuable qualitative data to our analyses.

Secondly, the students filled out the TVQ after they had finished the tasks. As Ainley (2006) has pointed out, this might have affected students' answers because they might have been influenced by their ideas about the difficulty of the tasks they had just performed. On the other hand, if students had filled out the TVQ before they started working on the tasks, they might not have had a clear image of the skill under study.

Thirdly, the variables procedural knowledge, task value, and school accounted for about one-third of the variance in task score in the regression analysis. So, about two thirds of the variance of the task score is still unaccounted for. Therefore, future research could focus on other variables that might influence the quality of students' answers, such as students' prior content knowledge and their epistemic beliefs. These beliefs have been researched as predictors of students' answers when evaluating the trustworthiness of internet sources (e.g., Strømsø et al., 2011). For this study, and earlier in the testing phase, we asked students to fill out a questionnaire on epistemic beliefs, using a questionnaire developed by Stoel et al. (2017). However, in our sample the Cronbach's alphas on the three scales of that questionnaire were too low and too unstable to use in our analysis.

Students' prior content knowledge and their epistemic beliefs are variables related to the subject of history. However, it is also possible that students' performance when evaluating the trustworthiness of sources is influenced by other variables that are more related to a students' personality, such as being able to deal with ambiguity and having perseverance, because when applying this historical skill, there usually is not one simple answer and it requires several thinking steps. It would be interesting to further research whether this kind of personal variables influence student performance. Furthermore, the sources used in this research were not about controversial topics or on topics students might have had strong (personal) feelings about. Further research could focus on more controversial topics, such as slavery, to find out whether this influences students' evaluation of trustworthiness.

Finally, the question can be raised whether the results of this study can be seen as representative, since we only had six participating schools. However, because these participating schools were situated in urban, suburban, and rural areas, and because there were differences in the way the participating teachers had paid attention to this historical reasoning skill, we believe that the results can be considered quite representative for Dutch students in grade 9. Since the results of this study only reflect a particular point in time, it would be interesting to conduct a longitudinal study in which a group of students is followed during their years in secondary education to obtain more insight into how a student's ability to apply this skill develops over the years. More insight in this development can help to design tasks in such a way that these tasks are suited for the different age groups in terms of complexity.

7. Conclusion and implications

How to evaluate the trustworthiness of sources is a complicated historical reasoning skill because it requires more than just learning the criteria of trustworthiness. It requires students to make well-founded decisions about whether to use a source as evidence to answer a historical question. The outcomes of this study inform us in a detailed way about, on the one hand, how students evaluate the trustworthiness of historical sources, their procedural knowledge and the value they attach to learning this skill and, on the other hand, the predictive value of these variables on the quality of students' answers. Students did not perform well on the tasks, and they were unable to write down many of the correct criteria of trustworthiness on an open-ended question measuring their procedural knowledge. Results indicated that students were moderately positive about the usefulness and importance of the skill, but they were not particularly interested in learning this skill. Both procedural knowledge and task value were significant predictors of task performance.

These results can be used to further improve the teaching of this historical reasoning skill to students in secondary education. This study has shown that more explicit attention is necessary to help students make their implicit knowing of the criteria for sourcing more explicit. Students are not always aware of the knowledge they possess and apply. By using explicit instruction and modeling, teachers might help their students recognize the procedures and knowledge necessary to evaluate the trustworthiness of sources. By starting with concrete tasks, students can be made aware of the criteria they already unknowingly apply, which might also increase students' interest in the skill. Once the students have realized which knowledge they already possess, the criteria can be explained in a more abstract manner.

When teaching students the skill of evaluating the trustworthiness of historical sources, explicit attention should be paid to why this is an important and useful skill, since the scores on task value were positively related to student performance. Finally, special attention should be paid to making tasks more interesting for students, for example by using authentic tasks with a link to students' own life, to enhance interest in learning how to apply the historical reasoning skill under study. Thus, teaching students how to evaluate the trustworthiness of historical sources requires more interventions rather than simply spending more time practicing the skill in the classrooms.

Appendix A

Example of a Task

The Battle of Stalingrad.

You are writing an essay about the Battle of Stalingrad. This battle took place in the Second World War. In 1942-1943, the Germans fought the Russians in the city of Stalingrad. You have found the following source:

In 2003, a Soviet soldier describes the Battle of Stalingrad:

I remember that it was a clear, beautiful morning. That day, the Germans started an enormous airstrike. I saw more than 2000 planes that bombed the city. Forty thousand people died in Stalingrad because of those bombs. "The barbarian attack", we still call those bombings because they were meant to kill our civilians.

Question: This source helps us understand how the Battle of Stalingrad was fought.

I agree/I do not agree/I both agree and disagree.

Explain your answer:

Appendix B

Rubric for the Battle of Stalingrad task

Score	Description	Example of student' s answer
0	The student's answer is wrong OR the student only uses the content of the source.	- I agree. He explains what happens and even though it is not very detailed, it is clear information.
1	The student recognizes that the author is an eyewitness (information) OR that this is only the story from the Russian side (position).	- I agree because he has seen it himself, because he says he saw many planes and so on. So, I agree because if you hear it from someone who has not seen it himself it is less trustworthy.
2	The student notices that the story was told 60 years after the Battle of Stalingrad and explains why that could make the source less trustworthy (time).	- I disagree because the soldier does not tell this story until 2003, and therefore there is a great chance that he has forgotten details or tells it incorrectly.

References

Ainley, M. (2006). Connecting with learning: Motivation, affect and cognition in interest processes. *Educational Psychology Review*, *18*(4), 391-405.

Breakstone, J. (2014). Try, try, try again: The process of designing new history assessments. *Theory & Research in Social Education*, *42*(4), 453-485.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed. ed.). Routledge.

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*(4), 609-624.

Council of the European Union. (2018). *Council recommendation of 22 May 2018 on key competences for lifelong learning*. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018H0604(01).

De, La, Paz, S., & Felton, M. K. (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), 174-192.

Eccles, J. S., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin*, *21*(3), 215-225.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215-251.

European Commission. (2018). *Tackling online disinformation: A European approach*. Retrieved from https://cor.europa.eu?EN/our-work/Pages/OpinionTimeline.aspx?olpd=CDR-3908-2018.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. Sage publications.

Halvorsen, A.-L., Harris, L. M., Aponte Martinez, G., & Frasier, A. S. (2016). Does students' heritage matter in their performance on and perceptions of historical reasoning tasks? *Journal of Curriculum Studies*, *48*(4), 457-478.

Harris, L. M., Halvorsen, A.-L., & Aponte-Martinez, 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), 109-121.

Jacobsen, R., Halvorsen, A.-L., Frasier, A. S., Schmitt, A., Crocco, M., & Segall, A. (2018). Thinking deeply, thinking emotionally: How high school students make sense of evidence. *Theory & Research in Social Education*, *46*(2), 232-276.

Kiuru, N., Spinath, B., Clem, A.-L., Eklund, K., Ahonen, T., & Hirvonen, R. (2020). The dynamics of motivation, emotion, and task performance in simulated achievement situations. *Learning and Individual Differences*, *80*, 1-11.

Landis, J. R., & Koch, G. G. (1977). *The measurement of observer agreement for categorical data* (pp. 159-174). biometrics.

Levesque, S., & Clark, P. (2018). Historical thinking: Definitions and educational applications. In S. A. Metzger & L. McArthur Harris (Eds.), *The Wiley international handbook of history teaching and learning* (pp. 119-148). Wiley-Blackwell.

McGrew, S., Ortega, T., Breakstone, J., & Wineburg, S. (2017). The challenge that's bigger than fake news: Civic reasoning in a social media environment. *American Educator*, *41*(3), 4-11.

McGrew, S., Smith, M., Breakstone, J., Ortega, T., & Wineburg, S. (2019). Improving university students' web savvy: An intervention study. *British Journal of Educational Psychology*, *89*(3), 485-500.

Nokes, J. D. (2013). *Building students' historical literacies. Learning to read and reason with historical texts and evidence*. Taylor & Francis.

Nokes, J. D. (2017). Exploring patterns of historical thinking through eighth-grade students' argumentative writing. *Journal of Writing Research*, *8*(3), 437-467.

Nokes, J. D., Dole, J. A., & Hacker, D. J. (2007). Teaching high school students to use heuristics while reading historical texts. *Journal of Educational Psychology*, *99*(3), 492-504.

Paul, J., Macedo-Rouet, M., Rouet, J.-F., & Stadtler, M. (2017). Why attend to source information when reading online? The perspective of ninth grade students from two different countries. *Computers & Education*, *113*, 339-354.

Perez, A., Potocki, A., Stadtler, M., Macedo-Rouet, M., Paul, J., Salmeron, L., & Rouet, J.-F. (2018). Fostering teenagers' assessment of information reliability: Effects of a classroom intervention focused on critical source dimensions. *Learning and Instruction*, *58*, 53-64.

Perfetti, C. A., Britt, M. A., Rouet, J.-F., Georgi, M. C., & Mason, R. A. (1994). How students use texts to learn and reason about historical uncertainty. In M. Carretero & J. F. Voss (Eds.), *Cognitive and instructional processes in history and the social sciences* (pp. 257-283). Lawrence Erlbaum Associates Inc.

Pintrich, P. R., & Schrauben, B. (1992). Students' motivational beliefs and their cognitive engagement in classroom academic tasks. *Student perceptions in the classroom*, *7*, 149-183.

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the motivated Strategies for learning questionnaire (MSLQ)*. National Center for Research to Improve Postsecondary Teaching and Learning.

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated Strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement*, *53*(3), 801-813.

Rantala, J., & van den Berg, M. (2015). Finnish high school students' and university students' ability to handle multiple source documents in History. *Historical Encounters*, *2*(1), 70-88.

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

Seixas, P., & Morton, T. (2013). *Six big historical thinking concepts*. Nelson Education Ltd.

Smith, M., Breakstone, J., & Wineburg, S. (2018). History assessments of thinking: A validity study. *Cognition and Instruction*, 1-27.

Stoel, G., Logtenberg, A., Wansink, B., Huijgen, T., van Boxtel, C., & van Drie, J. (2017). Measuring epistemological beliefs in history education: An exploration of naive and nuanced beliefs. *International Journal of Educational Research*, *83*, 120-134.

Strømsø, H. I., Bråten, I., & Britt, M. A. (2011). Do students' beliefs about knowledge and knowing predict their judgement of texts' trustworthiness? *Educational Psychology*, *31*(2), 177-206.

van Boxtel, C., & van Drie, J. (2018). Historical Reasoning: Conceptualizations and Educational Applications. In S. A. Metzger & L. McArthur Harris (Eds.), *The Wiley international handbook of history teaching and learning* (pp. 149-176). Wiley-Blackwell.

van Drie, J., & van Boxtel, C. (2008). Historical reasoning: Towards a framework for analyzing students' reasoning about the past. *Educational Psychology Review*, *20*(2), 87-110.

Velayutham, S., & Aldridge, J. M. (2013). Influence of psychosocial classroom environment on students' motivation and self-regulation in science learning: A structural equation modeling approach. *Research in Science Education, 43(2)*, 507-527.

Walraven, A., Brand-Gruwel, S., & Boshuizen, H. P. (2009). How students evaluate information and sources when searching the World Wide Web for information. *Computers & Education*, *52*(1), 234-246.

Wigfield, A., & Cambria, J. (2010a). Expectancy-value theory: Retrospective and prospective. In S. Karabeninck & T. C.

Urdan (Eds.), *The decade ahead: Theoretical perspectives on motivation and achievement* (pp. 35-70). Emerald Group Publishing Limited.

Wigfield, A., & Cambria, J. (2010b). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, *30*(1), 1-35.

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, *12*(3), 265-310.

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.