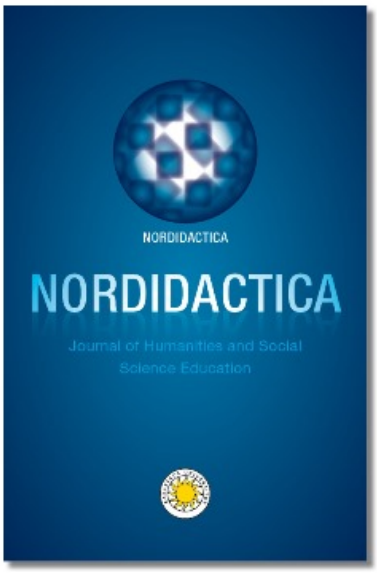
Designing an assessment tool for historical literacy: the case of Copernicus

Amna Khawaja



Nordidactica

- Journal of Humanities and Social Science Education

2018:3

Nordidactica – Journal of Humanities and Social Science Education
Nordidactica 2018:3
ISSN 2000-9879
The online version of this paper can be found at: www.kau.se/nordidactica

NORDIDACTICA – JOURNAL OF HUMANITIES AND SOCIAL SCIENCE EDUCATION
ISSN 2000-9879
2018:3 1-26

Designing an assessment tool for historical literacy: the case of Copernicus

Amna Khawaja

University of Helsinki

Abstract: This design-based research describes the development process of an assessment tool for historical literacy at primary school level where existing assessment materials are scarce. The assessment task was tested thrice during a two-year-period in Finland and Sweden. Sixty pupils participated in a pen-and-paper-test while seven took part in think-aloud interviews. The task included three written documents on the relationship between Nicolaus Copernicus and the Catholic Church. The length of the original documents was reduced and the language simplified age-appropriately. The results revealed a tendency to read the sources as information rather than as evidence. Also, the concept of reliability proved difficult. Alterations during the re-design phases included dividing broad questions into smaller entities and directing pupils´ attention to the characteristics of source types. Many pupils responded to the weighted multiple-choice (WMC) items as if they were traditional multiple-choice questions: they chose the first correct sounding option without pondering over the others. However, one WMC item was particularly successful as all the think-aloud protocols showed that the item met its target construct. As a whole, the artefact did elicit historically literate observations among some pupils.

KEYWORDS: HISTORICAL LITERACY, HISTORICAL THINKING, DESIGN-BASED RESEARCH ASSESSMENT, PRIMARY SCHOOL

About the author: Amna Khawaja is a doctoral student at the University of Helsinki. Her research is focused on historical thinking at a primary school level. Her additional interests are assessment and teaching materials in history education.

Introduction

To follow the international discourse on history education can sometimes be confusing as the same phenomenon might be referred to with several different terms (Seixas & Ercikan, 2015, p. 1; Van Drie & Van Boxtel, 2008, p. 88). There is little consensus on how to use terms such as historical thinking, historical competence, historical skills, historical understanding, historical reading, historical consciousness, or historical literacy in an unequivocal manner.

The key concept of this paper, historical literacy, is as ambiguous as all the other concepts related to the field of history education. According to Seixas (2006, p. 2), historical literacy is achieved through historical thinking thus making the six historical thinking concepts competencies in historical literacy. Maposa and Wasserman (2009) worked on conceptualizing historical literacy and reached a definition in which they include nearly all the concepts associated with history education: content knowledge, second order concepts (e.g. change and continuity, cause and consequence), source work (e.g. sourcing, corroboration, contextualization) as well as historical consciousness and understanding. They also state that among the major theorists there is an agreement that historical literacy is "the embodiment of what a learner acquires through the learning of school history" (Maposa & Wasserman, 2009, p. 59). However, there are others (Haydn, Arthur & Hunt, 2001; Rantala & van den Berg, 2013; Veijola & Rantala, 2018), who see historical literacy as a more specific term than historical thinking, a view that the Finnish Core Curriculum (2014) has adopted as well.

The concept of historical literacy was introduced for the first time in the most recent Finnish National Core Curriculum (revised in 2014). Acquiring the ability to work with documents and to make interpretations based on them is the requirement for historical literacy. A pupil proficient in historical literacy should be able to detect the motives and intentions of those who authored the documents, which means that historical literacy is closely connected with interpretation of sources and historical empathy. In order to form an interpretation a pupil should have sufficient knowledge of the historical period in question. Familiarizing with the attitudes, beliefs and the way of thinking of a certain historical period is a prerequisite for understanding the behaviour and decisions made by those who lived in the past. (Finnish National Core Curriculum for Basic Education 2014.) Keeping this in view, this study aims at producing new, formative assessment material for primary school pupils aged 11–12. The design process itself is at the forefront of this study. The primary focus is on the development of the assessment tool.

In the Finnish context we see a policy shift in the 1990´s when the national curriculum for the first time included some elements of historical thinking. The curriculum of 2004 continued the trend and, among other things, it expected pupils finishing their 6th grade to "understand that historical knowledge is an interpretation of historians and thus can change due to new sources or different perspectives" (Core Curriculum 2004). While the curricular change in Finland from 1994 to 2014 was gradual, the change in Sweden from the national curriculum of 1994 to the most recent one in 2011 was more steep. The narrative approach emphasizing national and local history was replaced by the goal of teaching disciplinary skills (Samuelsson & Wendell

2016, pp.483–484). Despite the changes made in the Swedish national curriculum, the teaching practices did not seem to change significantly (Samuelsson & Wendell 2016; Stolare 2017). Similar concerns were raised in Finland in the evaluation conducted by the Finnish National Board of Education in 2011: pupils finishing their secondary education had poor skills in historical thinking (Ouakrim-Soivio & Kuusela, 2012, pp. 49–50). In view of these results it seems possible, even probable that concentrating on substance knowledge and memorizing isolated facts has continued to thrive in Finnish schools (Rantala, 2012, p. 197). The slow change towards teaching discipline-specific skills in primary school context has been explained in several different ways. According to Stolare (2017) embracing the change means questioning the traditional, national narrative approach to history, something that many teachers are not prepared to do. Another explanation is that as the teachers on primary level are usually class teachers, they might not have a deep understanding of history as a discipline. (Stolare, 2017, p. 37.)

Although the use of textbooks has decreased during the last decade, they are still the most commonly used source in history lessons in the United States according to a report from the 2014 National Assessment of Educational Progress (NAEP 2014). Also Finnish schools rely on textbooks and because a large number of history textbooks used in Finnish schools still focus on content knowledge, the change towards teaching historical thinking has been slow. Assignments employing primary sources or assessment material in keeping with the National Core Curriculum objectives are only rarely found in history textbooks. Although some new emerging digital materials show promise regarding discipline-specific skills, printed textbooks and assessment material still have a strong hold in schools. As Seixas, Gibson and Ercikan (2015, p.103) have pointed out, without proper guidance and means for assessment, teachers are unlikely to adopt new learning objectives defined in curricula. The need for new kinds of assessment materials has been acknowledged and expressed by several researchers (e.g. Breakstone, 2014; Smith, 2018; VanSledright, 2014). To summarize, the challenges in teaching and assessing historical literacy in primary school are not only related to the attitudes and competencies of the teachers as Stolare (2017) points out but also to the availability of appropriate materials. This paper addresses the latter issue and is an attempt to be part of the current international efforts in capturing and assessing complex thinking in history education, especially among primary school pupils.

The main question the present work addresses is:
What kind of assessment tool is appropriate for assessing historical literacy among 11–12-year-old primary school pupils?

As the research question is broad, it will be dealt with in three parts in “Testing and re-designing the assessment task”. The first part discusses open-ended questions, the second part weighted multiple-choice questions and the final part is concerned with reading multiple sources. The aim is to find out whether the items in the task are understood by the pupils as intended, whether they tap into the intended target constructs and whether the sources used are appropriate in length and level of difficulty.

Conceptual framework

Pupils and students need a variety of literacy skills during their education. Shanahan and Shanahan (2008) look at different kinds of literacies as a continuum where there is an increasing specialization over the years: in primary education pupils start with basic literacy, then develop intermediate literacy and finally reach disciplinary literacy in secondary education. They further state that students in secondary education should be given instruction for discipline-specific literacies but generalizable literary strategies such as decoding, fluency and basic comprehension strategies are reasonable goals for primary school pupils (Shanahan & Shanahan, 2008, p. 44). However, using only general literacy strategies when studying history, even in primary school, means that pupils cannot fully engage in historical inquiry, where general reading strategies as Nokes (2010, p. 523) puts it, are ´essential but insufficient´. To become historically literate means reading texts with a certain degree of doubt (Nokes, 2010, p. 521) and as evidence rather than as facts (Wineburg, 1991). Pupils also need to learn specific heuristics, such as those introduced by Wineburg (1991): sourcing, contextualisation and corroboration, all of which are discipline-specific literary skills. The question is, whether or not primary school pupils can be expected to learn these literacy skills.

Several studies have suggested that pupils at primary school level are capable of historical thinking (see Barton, 1994; Foster &Yeager, 1999; Lee &Ashby, 2000; VanSledright, 2002). VanSledright (2002) worked with 11-year-olds, who were able to learn the basic steps for approaching and interpreting sources. As it appears that historical literacy can be taught at primary school level, teachers should have access to age-appropriate assessment material. Yet, most of the new assessment materials and classroom practices are designed for secondary and upper secondary school levels (e.g. Breakstone, 2014; Reisman, 2012a; 2012b; Seixas, Gibson & Ercikan 2015; Veijola & Mikkonen, 2016). However, it is in primary schools that the pupils first encounter the basic concepts of historical thinking and start practicing discipline-specific skills. Therefore the importance of the way historical literacy is taught and assessed in primary schools cannot be overemphasized.

VanSledright (2014) argues that in large-scale history assessment, construct validity is often compromised in order to ensure reliability. If history assessment has low construct validity, it does not capture the true essence of history as a discipline because aspects that are characteristic of history have not in fact been measured (Vansledright, 2014, p.16). For example, if competency in history is assessed through items concerned only with recollection and recognition, it measures only one small aspect of history as a discipline. It has also been noticed in the Nordic context that teachers tend to use assessment material with simple factual questions, which means that pupils are not given the possibility to show their skills in historical thinking (Rosenlund, 2011, p. 141).

In Finnish primary schools, teachers assess competency in history based on 11 target constructs stated in the National Core Curriculum. Four of these constructs are especially relevant to mastering historical literacy. I have designed the task in this paper to tap into these four constructs (Table 1). Each item in the assessment task addresses one or two of the target constructs.

TABLE 1.

The target constructs for historical literacy from the Finnish National Core Curriculum (2014, p. 257).

Target Construct	Description
TC 2	to guide the pupil to recognize different sources of history
TC 3	to guide the pupil to notice that historical information can be interpreted in different ways
TC 5	to guide the pupil to understand the motives behind people´s actions
TC 10	to instruct the pupil to explain how interpretations may change as a consequence of the new sources or new ways of examining them

All the four target constructs in Table 1 are in some way related to source interpretation and thus form an intact entity**.** Although the four target constructs are taken from the Finnish National Core Curriculum, they are in keeping with the key concepts of historical thinking used widely in literature concerning historical thinking (see Seixas, 2006; VanSledright, 2002).

Methods and materials

Design-based research is pragmatic in nature and aims at improving the classroom practices. It is not research about education but rather research for education. An essential element of any design- based research is to create an artefact, which can be utilized widely. (Juuti & Lavonen, 2006, p. 54.) The most essential characteristics of design-based research is its iterative and cyclic nature. The research proceeds by designing, testing, analysing and re-designing. This cycle is repeated several times. (Design-Based Research Collective, 2003; Juuti & Lavonen, 2006.)

The testing of the assessment task was conducted in three phases between November 2015 and October 2017 using two different data gathering approaches (see Table 2). The think-aloud protocols (from now on TAPs) had two objectives. Firstly, it was important to learn about pupils´ thought process as they answered an item. Secondly, TAPs were used to gather information about the way pupils navigated through the sources and questions, something that cannot be done through pen-and-paper tests. Although the think-aloud method stems from psychological research, it has been used for studies in problem-solving (Wineburg, 1991), design-based research (Breakstone, 2014) and specifically to find evidence of cognitive processes in assessment tasks (Ercikan, Arim, Law, Domene & Lacroix, 2010; Smith 2018). The aim is to capture the thinking process of the pupils through verbalization as they read the assessment task. This entails that the thinking is raw, unrefined and reveals any hesitation and self-correction (Martin & Wineburg, 2008, p. 307).

TABLE 2.

The participants of the study according to data-collecting method.

Phase	Year	Place	Pen-and-paper participants	Think-aloud participants
Phase 1	2015	Helsinki	18 (1–18)	2
Phase 2	2016	Helsinki	21 (21–41)	0
Phase 3	2017	Uppsala	21 (42–62)	5
In total			60	7

The participants were 11–12-year old pupils studying in medium-sized Finnish and Swedish schools, which were selected because of their past co-operation with the universities of Helsinki and Uppsala. The schools were normal municipal schools and the classes participating in the study were typical in size and included some pupils with learning difficulties. It was considered advantageous to include Swedish pupils in the study because it would be a test for the wider applicability of the assessment tool.

Although Finland and Sweden have different social and cultural traditions, the two countries have very similar national curricula for history teaching on primary level. In the Swedish National Curriculum for primary school (Lgr11) historical literacy as a concept is not mentioned as such but there are several objectives relevant for historical literacy. For instance pupils should be taught to examine, interpret and evaluate sources. Pupils should learn to evaluate people of the past by the circumstances and belief systems of the time period in question. In addition, pupils are encouraged to examine history from several different perspectives. (Swedish National Agency for Education, Lgr11, p.205.) The most significant difference between the two countries´ curricula is that the Swedish one defines the historical content to be studied in detail, unlike the Finnish equivalent.

Comparison between the Finnish and Swedish pupils´ competency in historical literacy was not my intention and could not be done based on this data since the Finnish and Swedish pupils had different versions of the assessment task. Most of the pupils participated by completing the task in writing (see Table 2). Seven pupils completed the task through think-aloud-interviews, two in 2015 and five in 2017. I asked the respective teachers to select the think-aloud participants based on the pupils´ ability to cope with unfamiliar situations, such as think-aloud interviews. Their school grades or general competency in history was not taken into account. Maximum time reserved for completing the task was 60 minutes.

The assessment task

Because of the central role of the artefact in design-based research, the creation of the assessment task is described in detail. I shall attempt to give my reasons for the use of documents in the assessment task and also discuss the choice to include both open-ended and multiple-choice items in the task.

Working with documents

Choosing to work with historical sources might seem self-evident given the fact that as early as in the 1970´s the Schools Council History Project in the U.K. approached history education through using multiple sources (Booth, 1994). Just a few decades later English and Welsh primary school pupils used material containing both primary and secondary sources (Foster & Yeager, 1999). The use of sources is also prevalent in North American history education. In Finland, on the other hand, the use of sources in assessment tasks has not been a common practice at the primary school level although there should be no obstacles in doing so. Finnish pupils are fairly skilled at reading, which has been shown both by national and international evaluations. The PIRLS 2016 (the Progress in International Reading Literacy Study) evaluated the reading skills of 4th grade pupils in 50 countries or regions. Finland scored fifth highest among all the participants. Sweden had the 12th highest score. Interestingly, in both Finland and Sweden pupils were much more proficient at retrieving information and making simple conclusions than reading in an interpretative and evaluative way. (Mullis, Martin, Foy, & Hooper, 2017.)

The assessment task consisted of two A4-size papers printed on both sides. I titled the assessment task `Copernicus—friend or foe of the Catholic Church?` in order to present a problem for the pupils to solve. There are three written sources (see Table 3). The primary and secondary sources present contradicting views, which is meant to create a cognitive dissonance, thus making the pupils face the interpretative nature of history.

TABLE 3.

Sources used in the assessment task.

Source number	Source type and author	Content
1	Drawing by Copernicus	Heliocentric model
2	Drawing by Velho	Geocentric model
3	A preface by Copernicus	A conciliatory preface dedicated to Pope Paul III. Copernicus acknowledges that there are “some” who would object to his work but does not identify them as people within the Catholic Church.
4	A textbook excerpt	Describes the relationship between Copernicus and the Catholic Church mutually hostile.
5	A letter by Cardinal Schönberg	The Cardinal expresses his admiration and support to Copernicus.

All the written sources used in the task are excerpts, i.e. none are presented in their original form. While the authenticity of sources is thus compromised, Reisman (2012a) thinks that it is necessary to modify sources in a way that they are understandable to pupils both visually and cognitively. The modification process was similar to that of Reisman´s (2012a): focusing, simplification and presentation. I took only the most essential parts of each source and simplified the vocabulary while trying to maintain the tone typical to the era. I shortened the preface from 2086 words to 83, the letter from 303 words to 97 words and I chose a 77-word excerpt from the textbook. While Reisman (2012a), working with high school students, aimed at keeping the sources below 250 words, sources used in the present study were each less than 100 words. In addition to the three written documents, two drawings were used as sources, one drawn by Copernicus himself and one by the Portuguese cartographer and cosmographer Bartolomeu Velho in 1568. These drawings were used in the background questions, where pupils were asked to identify the heliocentric and geocentric models.

Open-ended and multiple-choice items

In order to answer an open-ended item well, pupils need to have a good command of writing and the ability to structure an answer. These qualities are especially important when answering essay questions. Failing to write well can reflect poorly on any assessment performance, thereby restricting the pupil to show his or her skills and knowledge.

As stated earlier, Finnish children are proficient readers but their writing skills are less convincing. In the 2006 assessment conducted by the Finnish National Board for Education, 34 percent of the 7th grade pupils had poor writing skills (Lappalainen, 2007). Also Swedish pupils find writing much more challenging than reading. The national test in Sweden conducted during 2016–2107 showed that up to 41 percent of 6th grade pupils either failed the writing test (14%) or passed it with the lowest of five numerical grades (27%) (Swedish National Agency for Education). The three open-ended questions in the present assessment task require only short answers and therefore minimize the importance of writing skills. On the other hand, open-ended questions, unlike multiple-choice questions, have the potential to reveal the pupil´s reasoning process. Open-ended questions also provide the teachers with valuable information about how each pupil´s historical thinking is progressing (VanSledright, 2014, p. 87). In the present assessment task, I have used three open-ended questions (two compulsory and one extra question) for providing the pupils with means to express their thinking-process.

Multiple-choice questions were not originally designed to give information about the level, let alone the nature, of students´ thought process (Wineburg, 2004, p. 1) and they have been criticized for assessing mainly substance knowledge (Breakstone, 2014; VanSledright, 2014; Rantala, 2012). Recently Smith (2018) studied the cognitive validity of HTT (historical thinking test) multiple-choice items constructed by Reisman (2012b). By conducting think-aloud interviews with 12 high school students he investigated whether or not students engaged in the same cognitive processes that were

intended by Reisman (2012b). The results were promising as the HTT items elicited the intended historical thinking construct (sourcing, contextualization and corroboration). However, Smith (2018) remains cautious for two reasons. First, on each item at least one student answered correctly without using the targeted cognitive process. Secondly, the multiple-choice items under normal test conditions make it possible to draw only binary all-or-none inferences. According to Smith (2018) student proficiency in as many as 13 out of 72 responses could not be described in this binary manner. Smith (2018) therefore concludes that measuring complex historical thinking processes with HTT items might not be possible. (Smith, 2018, p. 22–23)

In an attempt to improve the validity of multiple-choice questions, VanSledright (2011; 2014; 2015) has used weighted multiple-choice questions (from now on WMC), where more than one option generates points. Only one option is entirely incorrect. In the case of four options, several scoring possibilities exist (4,2,1,0 or 4,3,2,1 or 3,2,1,0). There are two advantages of WMC items according to VanSledright (2014). First, they may provide an access to both procedural and substance knowledge and take into consideration the complexity of history as a discipline. The second advantage is the straightforward scoring process. (VanSledright, 2014.) Although multiple-choice questions have their shortcomings, using the weighted version gives the pupils the opportunity to answer without being dependent on their writing skills. This is especially relevant when assessing primary school pupils, as they cannot be expected to write long essays.

Testing and re-designing the assessment task

One of the characteristics of design-based research is that the results are inseparable from the developing process. The results of this study are presented here in three parts. I start by describing how individual items evolve from phase 1 to phase 3. Since it is not possible to present all the changes made in the present assessment task, I have chosen one openended and two multiple-choice questions to explain the re-design decisions. The results concerning one open-ended question have been left out from the present paper. These and analysis of WMC items in phase 1 have been briefly presented earlier in a study about history education in Finland (Rantala & Khawaja, 2018). Neither the design-process nor its development through phases 2 and 3 has been reported earlier. In the present paper, the last part of the results is discussed in ´Reading the sources´, where I explain the way pupils dealt with the documents.

In all the classes participating in the present study, the pupils had studied the history of the Middle Ages before taking part in the test. None of the teachers had seen the assessment task in advance. According to the teachers, they had worked with sources in history lessons in the past. The pupils did not know the structure or the subject of the assessment task in advance.

The open-ended question

The first of three open-ended questions (see Figure 1) addressed the target constructs 2 and 3 (see Table 1), which concern recognizing different historical sources and understanding the interpretative nature of history. The aim of the question was to make pupils realize that all three sources have their shortcomings. Although source 3 (preface by Copernicus) and source 5 (letter by Cardinal Schönberg) are primary sources, they do not cover the overall relationship between Copernicus and the Catholic Church as Schönberg is just one representative of the establishment. Source 4 (the textbook) on the other hand gives a broader view but does not offer us the possibility to see, which sources the authors have used to formulate their interpretation.

Is the information given in the textbook (source 4) as reliable as sources 3 and 5 regarding the relationship between Copernicus and the Catholic Church?

FIGURE 1.

The first open-ended item in its original form in phase 1.

The item revealed that the concept of reliability was too big a challenge for 11-12-year-old pupils. Instead of looking into how, by whom and for which purpose the sources had been created, many thought of reliability as something to do with only the preciseness or the amount of information presented:

> *Yes, because source 4 is more precise (pen-and-paper, pupil 5, phase 1).*
>
> *As reliable....Well, maybe it has more information about this Copernicus, like in a nutshell, but I don't know if it as reliable (think-aloud, pupil 2, phase 1).*

The concept of reliability as such is not mentioned in the assessment criteria for the 6th grade in the current national curriculum. The pupils are assessed by their ability to recognize different kinds of source types and differentiate between fact and interpretation but evaluating the reliability of sources is not expected from them. In his intervention study VanSledright (2002) evaluated the reliability of sources with 11-year-olds. The results were encouraging, which in part led to the incorporation of reliability into the present task. To make the question more comprehensible I replaced the concept of reliability with credibility and divided the item into two parts as seen in Figure 2.

Question 1. Which of the following sources are primary sources? Circle the correct option/options.
A) source 3 *b) source 4* *c) source 5*

Question 2. Imagine that you were asked to write a credible account on the relationship between Copernicus and the Catholic Church. You could use only one source to help you. Which source would you use? Justify your choice.

FIGURE 2.

The open-ended item divided into two items in phase 2.

Each of the two new items addressed only one target construct thus making the items more focused. After identifying the source types in item one, the pupils were supposed to reflect on the qualities of these source types in item two. If the pupils were aware of when, by whom and for which purpose a source was produced, a process Wineburg (1991) calls sourcing, they would see both the strengths and shortcomings of each source. They would be able to notice that second-hand sources are interpretations made by different individuals. In phase 2, the majority (14 of 21 pupils) chose Copernicus´s preface and based their choice on the fact that Copernicus had written it himself. A typical answer in phase 2 was:

> *I would use source three because Copernicus himself is telling how he is feeling. At least that would be reliable information. He also says that some would like to dismiss his thoughts. All the information is coming from him (pen-and-paper, pupil 38, phase 2).*

In phase 2 a third of the class chose a source based on the amount of information the source provided, as was the case in the first phase. These pupils tended merely to repeat the content of the chosen source. In total, 12 pupils out of 21 justified their answer solely or partly by the fact that Copernicus had produced the source. Only two pupils were able to compare the sources and reflect on how the source type affects the credibility. Thus the textbook excerpt was ignored by a large majority and the interpretative nature of a secondary source was left unnoticed. The fact that half of the class justified their choice by stating that Copernicus had written the preface does imply that there was a beginner´s level sourcing. None chose Cardinal Schönberg´s letter.

In order to encourage pupils to consider all the sources and their qualities, I modified the question for phase 3. Instead of choosing one source, the pupils could choose one or two sources. This change would ideally lead pupils to focus more on source types and their qualities and thus comparing the advantages and disadvantages of primary and secondary sources. Comparing sources with one another is what Wineburg (1991) calls corroboration and VanSledright (2002, p. 111) inter-textual reading. This process is essential when evaluating the reliability–or in this case the credibility–of sources.

Phase 3 comprised five pupils´ think-aloud protocols and 21 pupils´ pen-and-paper tests. Only one pupil in the TAPs explicitly and spontaneously compared different source types with each other:

> *I would choose sources 3 and 5 because they are primary sources and one can trust them a lot. Because with source 4 one can never know if everything important has been used for interpretation (think-aloud, pupil 65, phase 3).*

Once again, the most common approach in phase 3 was to concentrate only on the content of the source. Out of 21 written answers in phase 3, six concentrated purely on the source that gave most information on the relationship between Copernicus and the Church. Six more pupils used the content as a partial argument, exactly as in phase 2. One pupil failed to give any argument. In comparison to phase 2 there was however an increased proportion of pupils who used source types and their qualities as justification for their response, thus taking their reasoning to a broader level. More than half chose either the two primary sources or the secondary source, which suggests that the source type became a relevant factor in pupils´ reflections.

One pupil taking part in TAP had difficulties in giving a strong argument for choosing sources 3 and 5 until I asked him to explain why he did not choose the secondary source (source 4). This prompted him to reflect in the following manner:

> *...partly because it´s not a primary source, because it feels, like somehow more credible when one knows what didn´t happen and there are people who tell what actually happened and their view of it. This [the secondary source] on the other hand is something where someone has found facts from different places and we don´t really know if it´s true (think-aloud, pupil 64, phase 3).*

Thus, arguing through negation (why not doing something) might be helpful for some pupils for organizing their thoughts. Dismissing a source means reflecting on its shortcomings, something that pupils tend to overlook. Pupil 64 expressed that he did understand the interpretative nature of history but needed support in arranging his thoughts. The graphic organizer (see Figure 3) is meant to help pupils to address every source one way or another.

Question 2. Imagine that you would have to write a credible account on the relationship between Copernicus and the Catholic Church. You could choose one or two sources to help you. Which source/sources would you use? Give your reasoning for both the source/sources you choose and the one/ones you don´t choose.

	Source 3	Source 4	Source 5
I would use because			
I would not use because			

FIGURE 3.

The final version of item 2.

Monte-Sano (2011, p. 213) states, and Veijola and Rantala (2018, p. 5) agree that using a graphic organizer in tasks for discipline-specific literacy is not productive because the essence of the discipline cannot be captured with such a tool. However, dismissing graphic organizers as incapable of helping to produce discipline-specific thinking might be too hasty. The pupils taking part in all testing phases clearly needed support in organizing their thoughts and a graphic organizer can act as a visual tool, which helps them to construct their ideas. For 11–12-year old pupils a graphic organizer might function as a kind of scaffolding they need before they are able to write fluently on the credibility of sources. Whether a graphic organizer can produce discipline-specific writing surely depends on *how* the tool is used, not on the tool itself.

Weighted multiple-choice items

VanSledright (2015) developed the WMC items but has omitted to examine the validity and reliability of WMCs. He has used peer-review and pilot testing among prospective teachers but has not conducted think-aloud interviews with pupils or students, something that according to him is needed (VanSledright, 2015, p. 82). Smith (2018) acknowledges that WMC items might be a more effective approach to capture historical thinking than traditional multiple-choice items but takes no position until further research has been done on WMC items. One of the aims of this paper is to examine whether or not WMC items are suitable for assessing historical literacy on primary school level.

Question 4. *Use source 3 and the information box. Circle the most suitable alternative. Copernicus dedicated the preface of his book to the Pope because he*	scoring
A. respected the Catholic Church.	1
B. *thought that it was the only way to get the Pope to be interested in his book.*	0
C. *understood that his book might upset the Catholic Church and wanted to be in good terms with the Church.*	4
D. *was afraid that the Pope would condemn his heliocentric model.*	2

FIGURE 4.

Weighted multiple-choice question 4 in phase 1.

The target construct of the first weighted multiple-choice item (see Figure 4) was to assess whether pupils are able to reflect on the motives of those who lived in the past. The pupils were expected to use their general knowledge about the position the Catholic Church had in the 16th century Europe. Additional information was provided in an information box. Option C carried the highest points as it takes into consideration the complexity of the situation and the motives for dedicating the preface to the Pope. It recognizes the power and the influence the Catholic Church had on people. The second-best was option D, which simplifies the matter. Although, as option A states, it is true that Copernicus respected the Catholic Church, this option does not show that Copernicus wanted to achieve something through dedicating the preface.

In the first phase, 4 pupils out of 18 scored full 4 points, while the majority chose the one-point option. Only one pupil was convinced by the 0-point alternative, which was intended to be an example of presentism in thinking. Because as many as four pupils chose two options instead of one, the prompt was clarified further before the next phase.

One of the TAPs shows, how the pupil went through different options, discarding the completely wrong option immediately and then weighed the remaining options:

> *Pupil 20: [Reading option A] Well, it was clear that he respected the Church although he attacked it, at least somewhere it said that he respected it, don´t remember which source it was (going through the sources and information box) Yeah here it says that this Copernicus respected the church.*
>
> *[Reading option B] This probably is not nonsense I mean it probably is nonsense, because he would hardly want the Pope to just buy the book. Sure, the Pope is a significant person, but that just wouldn´t make sense. [Reading option C] ...Yeah, this is a good option. [reading option D]. This is a good option as well. I don´t know which one is the right one, but I won´t choose alternative B. A, C and D, these are all a bit more true, that all of these could have been right....*
>
> *Researcher: Which one do you think is the best option?*
>
> *Pupil: Well this D is like...it´s pretty tough. I mean because he was afraid that the Pope would condemn his heliocentric model. Because if he did, he couldn´t*

> *show it to anyone and he ´d be in trouble, that fear is a powerful thing in people (think-aloud, pupil 20, phase 1).*

Pupil 20 did try to understand Copernicus´s motives but did not fully take a historical perspective. Instead, he referred to fear as a general factor in people´s actions. The TAPs in phase 3 show that the item was able to bring out reflection on Copernicus´s motives, but not always in a historical context. One pupil pointed out the inner conflict of Copernicus because he was a devoted Catholic but could not accept the geocentric model. The information from the information box was also utilized:

> *I would say C because, this [information box] says that the Church had tribunals and could give death sentences, so maybe there were many who wanted to see him hanged, because his… because his, because his book came out and was opposite to what they believed so they could get angry (think-aloud, pupil 65, phase 3).*

The advantages of WMC items become apparent in the think-aloud protocol of pupil 20. The pupil noticed that three options all have elements, which are correct. Were he to face a traditional multiple-choice-question the pupil most probably would have been able to eliminate the three wrong answers without being forced to go through a rationalizing process.

However, the TAPs in phase 3 revealed that the WMC items were not able to engage all the pupils in complex thinking. In fact, four out of five think-aloud protocols show that the pupils chose immediately only one option. One pupil considered all three options. When asked about why they did not choose another option, two pupils only then realized the noteworthiness of other options. The second WMC item (Figure 5) produced a similar outcome: three pupils reflected over two options while two straight away chose one option. Those pupils who failed to notice the nuances in the options may have done so because of the prompting. The scoring system (0,1,2,4 points) was not visible to the pupils. The only clue the pupils had was in the prompt—*Circle the most suitable option. Use source three and the information box*—suggesting that the most suitable option was not the only suitable option. In view of the encouraging results of phase 1 TAPs, the scoring was not shown to pupils, as it seemed that they would be able to weigh the options and notice the nuances. Phase 3 showed that the scoring should be visible to the pupils so that they can engage in more complex thinking. The revised prompting for question 4 is as follows:

- *Read source three and the information box. Circle the most suitable option. Only one option scores 0 points and the other three score either 1, 2 or 4 points.*

Question 5. *The history textbook (source 4) and the Cardinal´s letter (source 5) give two different kinds of impressions on how the Catholic Church reacted to Copernicus´s ideas. How can one explain this?*	scoring
A. The textbook authors have used different sources that tell that the Church condemned the ideas of Copernicus, and they left out the sources that told the opposite story	4
B. The cardinal was lying.	0
C. The textbook authors did not know that some churchmen supported the ideas of Copernicus	1
D. The textbook authors did not find the cardinal's letter an important source.	2

FIGURE 5.

Question 5 remained unaltered throughout the phases.

The aim of question 5 was to assess whether or not pupils understand the typical processes of history as a discipline and the way interpretations might change over time and through new evidence (see Table 1). The think-aloud protocols in both phases 1 and 3 brought out pupils´ thinking on how historical knowledge is produced. Pupil 64 used a total of 11 minutes to reach a conclusion and expressed several times that the question was challenging. The two excerpts are from the middle of his reasoning process and show that the pupil understands that some sources are more known than others:

> *The textbook maybe hasn´t found out about that this letter exists, but it depends on how big this letter is within, like, this history.*
>
> *---ehmm…I´m a bit stuck here…I think that, I think the reason they [sources 4 and 5] are different is because….this is difficult…[reading again through the options]. So, anyway, I think that in case, what are they called, the authors, knew that the letter existed, then I don´t think that they would have thought it unimportant because if you write a text book you should know what you are talking about because pupils are going to learn from the book so it should have proper facts. So that´s why I don´t think it´s D. Option C, it´s kind of the same thing, in case they didn´t, in case they knew, in case the authors knew that the letter existed, they knew that there were people supporting him. So it´s basically about whether the authors ever saw the letter or not--(think-aloud, pupil 64, phase 3)*

The challenge with the WMC items is whether a pupil can be expected to differentiate between the best, the second-best and the third-best option. The most correct and comprehensive option tends to be relatively lengthy, which is the case in both of the WMC items used in the present task. The length of the option might draw attention and make the option compelling solely based on its appearance. This in turn increases the risk of an item producing false positives, i.e. responses that are correct but result from undesired thinking processes (see Shemilt, 2015; Smith, 2018). The results from phase 3, where 19 out of 26 pupils chose the best scoring option, seem to confirm this. All four options should be of approximately the same length despite the weighted scoring.

All the seven TAPs used in this paper suggest that question 5 did engage the pupils to think about how sources are used to make interpretations such as textbooks. Some gave a more sophisticated reasoning than the others, but all seven reflected on the use of sources in making interpretations, thus tapping into the intended construct target.

Reading the sources

One of the most challenging things for teachers to know is how pupils process the information at hand. I have employed think-aloud protocols to find out more about the strategies pupils use when encountering sources, what catches their attention and what is left out unnoticed.

The title of the assessment task asked whether Copernicus was a friend or an enemy of the Catholic Church. I assumed it would be enough to direct the pupils´ thoughts while reading the sources. The TAPs show, that the title did not catch pupils´ attention and they failed to consider the relationship between Copernicus and the Church. Instead, most of their comments while reading the sources were concerned with astronomical issues. The following are examples related to the textbook excerpt (source 4):

> *I think this sounds good, although there is something that sounds a bit strange, it´s like this was in 1543 and that was in the 17th century, so it took a pretty long time before, what´s his name, Kepler and Galileo, reached the same conclusions-- (think-aloud, pupil 64, phase 3)*

> *Yeah, it seems that these Catholics wanted to have their worldview as it was and not anyone see the truth (think-aloud, pupil 65, phase 3).*

The pupils read the sources without knowing what they wanted to know from the sources; in other words, they lacked a point of view. As Seixas (2006) puts it, one can read sources either in search of information or for evidence. Telephone directories are read for information and a footprint on a crime scene for evidence (Seixas, 2006). The pupils in the present task used the first approach. The assessment task was constructed on the presumption that the pupils would detect the contradiction between the sources and go on to resolve it. The TAPs revealed that the contradiction remained unnoticed, which is common even with secondary students (Stahl, Hynd, Britton, McNish & Britton 1996). Reading three sources one after another seemed to direct the pupils to form a single narrative, similar to the one that the secondary source (source 4, a textbook excerpt) presented. This might be due to overloading the working memory by asking the pupils to process all three sources simultaneously. The textbook excerpt with its authoritative tone convinced the pupils, comparable to Paxton´s study (1997), where high school students found an anonymous textbook author more trustworthy than a visible author writing in the first person.

In order to direct pupils´ attention to the relationship between Copernicus and the

Catholic Church, more specific instructions on how to read the sources were given (see Figure 6). In phase 3 the pupils were asked to read all the sources and then answer the questions. The final version, on the other hand, advises pupils to reflect on the relationship after reading each source individually. Pupils are made to focus on each source by asking them to circle an image (emoticon) as seen in Figure 6. Yet, the amount

of writing is not increased. In addition, in attempt to emphasize the contradiction between the sources, one sentence, where Copernicus names Schönberg and Bishop Giese as his friends, was added to the excerpt from the original source.

Read sources 3, 4 and 5. Stop after each source. What kind of a relationship did Copernicus and the Catholic Church have <u>according to the source</u>? Circle an appropriate image.

 *Good/friendly relationship*

 *Neutral relationship*

 *Bad/hostile relationship*

FIGURE 6.

The revised instructions given for reading the sources.

In conclusion, the testing and the developing processes showed that an assessment tool for historical literacy employed at primary school level should possibly entail the following qualities. First, the questions presented should be as specific as possible. Thus, rather than including multiple phases in one question e.g. identifying source types and evaluating their reliability, the question should be devided into smaller components. Scaffolding for structuring the answer might be useful as well because pupils at this level might not yet have the ability to construct cohesive and consistent answers. Second, WMC questions are able to elicit historically literate thinking among some pupils but the scoring system should be visible to pupils. Third, pupils tend to read sources as neutral information. Therefore, if a contradiction of sources is relevant to the task, this contradiction should be emphasized. Reading of the sources could be done in steps. An assessment task at primary school level should not expect pupils to master large entities neither while reading the sources nor answering the questions.

Limitations

Both reliability and validity are relevant concepts when discussing assessment. However, in this paper I shall confine myself with the issues concerning validity. According to Messick (1980) there are two essential questions to be asked when investigating validity of any measurement: "First, is the test any good as a measure of the characteristics it is interpreted to assess? Second, should the test be used for the proposed purpose in the proposed way" (Messick, 1980, p. 1012). I address the first question through the concept of construct validity, as suggested by Messick. This means evaluating the evidence about the properties of the assessment task. (Messick, 1980.)

According to modern validity theory, it is the test, inferences from the results and the use of the test that comprise validity (Messick, 1980, 1989, 1995). The assessment task in this paper is to assess only historical literacy and is not meant to determine a pupil´s competency in historical thinking, which according to the definition used in the present paper entails several other target constructs. Similar to the task by Seixas, Gibson and Ercikan (2015), this task is recommended to be used as a module, part of a bigger entity. The aim has not been to create a task for large-scale testing, but for classroom assessment.

I designed the assessment task to measure pupils´ competencies in historical literacy. Whether or not the task designed was able to do so is a matter of construct validity. In order to give a strong validity argument, the four target constructs should have been reflected in pupils´ responses. However, not every item met its target construct. Even after two phases of alterations, the open-ended item failed to make all the pupils concentrate on the qualities of source types. Although there was an increase in the proportion of those who were able to produce a relevant answer, many remained focused on irrelevant source content.

The think-aloud protocols showed that the first WMC item produced thorough thinking with some pupils but the target construct was met only partially. The majority focused on Copernicus´s motives but failed to take into account the historical perspective of Copernicus´s actions. The second WMC item did tap into its target construct. All the seven think-aloud protocols showed that pupils reflected on how sources are used for making interpretations such as textbooks.

The construct validity of the assessment task presented in this paper is somewhat limited. Invalidity usually stems either from construct underrepresentation or from construct-irrelevant variance (Messick, 1989, 1995). The former means, that the assessment is too narrow and lacks crucial dimensions of the construct. The current Finnish National Curriculum includes 11 target constructs for teaching history out of which 4 are specifically concerned with historical literacy. I have used all target constructs in the assessment task and thus tried to ensure that pupils were given the opportunity to show their competency.

Construct irrelevant variance on the other hand refers to assessment, which measures traits that are irrelevant to the construct, like reading comprehension in mathematical problem solving. To prevent construct irrelevant variance the task provided context knowledge, so that pupils would not be expected to know detailed facts about Copernicus and the Catholic Church. It may be argued that the assessment task required pupils to use their working memory on too demanding a level when asked to read three sources consecutively and to draw inferences based on them. Alterations were made in phase 3 to de-emphasize the role of working memory. Thus neither construct underrepresentation nor construct irrelevance variance seem to be the main cause for limitations in the validity of the assessment task.

If pupils in the participating classes had no experience of working with sources, it would affect the validity. However, both Finnish and Swedish national curricula expect that skills for historical literacy are taught in primary schools. In addition, the teachers confirmed that pupils had worked with documents in their history lessons. As Nokes

(2010) points out, there are several components to reading texts in a historically literate way: in addition to providing pupils with primary sources they should be given explicit instructions on reading strategies as well as opportunities to express their interpretations on the texts. Despite being exposed to sources in the classrooms it is not known whether pupils had been taught the way to approach sources.

The think-aloud interviews of this study were more interactive than recommended by some researchers (Ercikan et al., 2010). I did not restrict myself to a few predetermined phrases such as "go on" or "please continue" but instead occasionally asked further questions about pupils´ choices and arguments. This was meant to relax the only 11–12-year-old interviewees. Some pupils found it easier than others to express their thoughts spontaneously. Despite the initial training with Donald Duck comics—which are popular free-time reading in both Finland and Sweden—most of the interviewees failed to continuously think aloud. A longer pre-interview practice using additional material relevant to the assessment task at hand might have been beneficial.

Discussion

This study is an attempt to produce assessment material for historical literacy through testing and re-designing. The challenge has been to design the material—sources and questions—in a way that the level of historical literacy required is suitable for pupils aged 11 to 12 years.

The results of the present study suggest that both Finnish and Swedish pupils perceived sources as neutral information rather than as evidence to be interpreted. The uncritical approach that pupils had can be explained by the PIRLS 2016 results, which show that evaluative and interpretative reading was difficult for Finnish and Swedish pupils (Mullis et al., 2017). To overcome the problem of reading the sources uncritically I highlighted the contradiction between the primary and secondary sources. It would be of interest to test the assessment task described in the present paper in countries such as the United Kingdom or the United States, where according to the PIRLS-evaluation pupils are more skilled at interpretative reading than at retrieving factual knowledge (Mullis et al., 2017).

Majority of the pupils in the present study had difficulties in constructing consistent and cohesive responses. Pupils were able to make observations, which nevertheless remained isolated and did not lead to clear conclusions. Barton (1994) had similar results with 10–11-year-old pupils learning to work with historical sources. At the beginning of Barton´s one-year research period the pupils could barely name any written sources. At the end of the study period they were able to identify several source types, understand that some sources were more trustworthy than others and recognize the biased nature of some sources. Even then, drawing conclusions from the sources remained difficult, just as in the present study. It may be argued that pupils at primary school level are not to be expected to draw conclusions from historical sources. My approach has been to encourage pupils to draw conclusions by dividing larger entities

into smaller units and by making all the steps leading to a conclusion as visible as possible.

The TAPs showed that not every pupil could engage in disciplinary thinking. Pupils tended to use more general reading comprehension strategies and concentrate on the content of sources. However, by phase 3 several pupils did consider the relevance of source types when assessing their credibility and the interpretative nature of history. The present study supports the contention that working with sources needs experience and systematic guidance. Merely providing pupils with sources does not lead to the development of skills in historical enquiry (Martin & Wineburg, 2008; Nokes 2010; Reisman, 2012; Stahl et al., 1996).

Whether weighted multiple-choice questions have potential use in history assessment needs more research. The think-aloud protocols of this paper reveal that while some pupils were able to engage in thinking that is typical for historical literacy, many treated these nuanced options as traditional multiple-choice options. They chose whichever option sounded plausible. Two of the seven think-aloud protocols show that pupils went back and forth with options as well as sources. In doing so they spent more time with the item, which resulted in thorough thinking. Making the scoring of WMC items visible to pupils could have an impact on how pupils answer the items. An intervention study, where a group of pupils is introduced to the logic of weighted multiple-choice items, would help to decide whether or not being acquainted with the question type would elicit more complex thinking than the traditional multiple-choice items.

For classroom activities, the WMC items could provide an excellent way for creating discussion and debates. As for assessment, the situation is somewhat different. VanSledright (2014) suggests that the scoring can be mutually discussed and if a student is able to make a convincing argument, the scoring can be changed. Discussing the scoring can be a learning experience in itself (VanSledright, 2014, p. 85). The question is whether the WMC items are accurate enough for assessment and whether assessment can be based on items that need discussions to clarify the scoring.

The challenge faced by many of those producing assessment materials in history education is of large-scale testing, i.e. how to construct tests for the masses. In countries that do have large-scale testing, cost and efficiency are important factors when creating assessment materials (Wineburg, 2004; VanSledright, 2014). No such restrictions were considered while designing the present task as it is meant for classroom assessment. This fact provides an excellent opportunity for designing and using assessment material that tap into discipline-specific skills and complex thinking. It follows that open-ended questions, where pupils can express their thinking in their own words, can and should be used. Open-ended questions have the potential of providing teachers valuable information about their pupils´ thinking processes.

When scoring open-ended questions some degree of subjectivity cannot be avoided. To minimize this subjectivity, assessment criteria should be as unambiguous as possible. The criteria I have provided for the teachers include instructions for each scoring option. The assessment criteria were tested tentatively with one teacher (data not shown). The reliability of the assessment criteria could be improved by using a

larger number of teachers. The nature of design-based research is cyclic and continual, which entails more work to improve the task. The most important outcome of the present study is the artefact itself. The most recent version developed after phase 3 remains to be tested. In addition to the present qualitative approach, the assessment tool described and developed here should be tested using a larger data and quantitative methods.

The present paper has pointed out the difficulties in designing new and valid assessment materials. In a primary school context, limitations arise as to how much pupils can be expected to read, process and write when giving their responses. In addition, the maximum duration of 60 minutes limits the number of items in the assessment task. In all the three classes participating in the present study, the average time taken for completing the task was approximately 30 minutes, which would make it possible to include an additional item into the task.

The rather large proportion of those pupils who did focus purely on source content indicates that in the classes participating in this study, historical literacy might not have been taught as prescribed at the national curricula. Assessment material that would focus less on interpretative skills and argumentation might well be more practical for those primary school teachers who carry on emphasizing the importance of content knowledge but it would compromise the aim of embedding historical literacy and historical thinking in primary schools.

References

Barton, K. (1994). “I just kinda know”: Elementary student´s ideas about historical evidence. *Theory and Research in Social Education* 25(4), 407–430. doi.org/10.1080/00933104.1997.10505821

Booth, M. (1994). Cognition in history. A British perspective. *Educational Psychologist 29*(2). 61–69. doi.org/10.1207/s15326985ep2902_1

Breakstone, J. (2014). Try, try, try again: the process of designing new history assessments. *Theory & Research in Social Education, 42*(4), 453–485. doi.org/10.1080/00933104.2014.965860

Design-Based Research Collective (2003). Design-based research: An emerging paradigm for educational enquiry. *Educational Researcher 32*(1), 5–8. Retrieved from: https://search.proquest.com/docview/216901476?accountid=11365

Ercikan, K., Arim, R., Law, D., Domene, J. & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by experts. *Educational Measurement: Issues and Practice 29*(2), 24–35. doi.org/10.1111/j.1745-3992.2010.00173.x

Finnish National Board of Education. (2004). National core curriculum for basic education 2004. Helsinki. Retrieved from: https://www.oph.fi/saadokset_ja_ohjeet/opetussuunnitelmien_ja_tutkintojen_perusteet/perusopetus

Finnish National Board of Education. (2014). National core curriculum for basic education 2014. Helsinki. Retrieved from: https://www.oph.fi/saadokset_ja_ohjeet/opetussuunnitelmien_ja_tutkintojen_perusteet/perusopetus

Foster, S. & Yeager, E. (1999). "You´ve got to put together the pieces": English 12yearolds encounter and learn from historical evidence. *Journal of Curriculum and Supervision 14*(4), 286–317. Retrieved from https://search.proquest.com/docview/196371949?accountid=11365

Haydn, T., Arthur, J. & Hunt, M. (2001). *Learning to teach history in the secondary school.* London: Routledge.

Juuti, K. & Lavonen, J. (2006). Design-based research in science education: one step towards methodology. *Nordic Studies in Science Education, 4*, 54–68. doi.org/10.5617/nordina.424

Lappalainen, O-P. (2007). *On annettu hyviä numeroita: Perusopetuksen 6. vuosiluokan suorittaneiden äidinkielen ja kirjallisuuden oppimistulosten arviointi 2007.* [Good grades have been awarded: Evaluation of Finnish language and literature proficiency among 7th grade pupils]

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

Maposa, M. & Wasserman, J. (2009). Conceptualising historical literacy—a review of the literature. *Yesterday & Today, no 4*, 41–66. Retrieved from http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S2223038620090001000 06&lng=en&tlng=en.

Martin, D. & Wineburg, S. (2008). Seeing thinking on the web. *The History Teacher 41*(3), 305–319. Retrieved from http://www.jstor.org/stable/30036914

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012–1027.

Messick. S. (1989). Meaning and values in test validation. The science and ethics of assessment. *Educational Researcher 18*(2), 5–11.

Messick, S. 1995. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

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

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 International Results in Reading*. Retrieved from Boston College, TIMSS & PIRLS International Study Center. Retrieved from: http://timssandpirls.bc.edu/pirls2016/international-results/

National Assessment of Educational Progress (2014). The Nations Report Card, Retreived from www.nationsreportcard.gov/hgc_2014/#history/contexts

Nokes, J. (2010). Observing literary practices in history classrooms. *Theory and Research in Social Education 38*(4), 525–544. doi.org/10.1080/00933104.2010.10473438

Ouakrim-Soivio, N. & Kuusela, J. (2012). *Historian ja yhteiskuntaopin oppimistulokset perusopetuksen päättövaiheessa 2011* [Learning outcomes in history and social studies at the end of basic education 2011]. Koulutuksen seurantaraportit 2012: 3. Helsinki: Opetushallitus.

Paxton, R. J. (1997). "Someone with like a life wrote it": The effects of a visible author on high school history students. *Journal of Educational Psychology, 89*, 235–250.

Rantala, J. (2012). How Finnish adolescent understand history: Disciplinary thinking in history and its assessment among 16-year-old Finns. *Education Sciences 2*(4), 193–207. doi:10.3390/educsci2040193

Rantala, J. & van den Berg, M. (2013). Lukiolaisten historian tekstitaidot arvioitavina. [Assessing historical literacy among Finnish high-school students] *Kasvatus 44*(4), 394–407.

Rantala, J. & Khawaja, A. (2018). Assessing historical literacy among 12-year-old Finns. *Curriculum Journal*. DOI: 10.1080/09585176.2018.1460273.

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

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

Rosenlund, D. (2011). *Att hantera historia med ett öga stängt*. [To do history with one eye closed]: Lund: Lund universitet.

Samuelsson, J. & Wendell, J. (2016). Historical thinking about sources in the context of a standard-based curriculum: a Swedish case. *The Curriculum Journal 27*(4), 479–499.

Seixas, P. (2006). *Benchmarks of historical thinking: A Framework for assessment in Canada.* Centre for the study of historical consciousness. Retreived from http://archive.historybenchmarks.ca/documents/benchmarks-historicalthinkingframework- assessment-canada.

Seixas, P. & Ercikan, K. (2015). The New shape of history assessment. In K. Ercikan & P. Seixas (Eds.), *New Directions in Assessing Historical Thinking* (pp.1–13). New York: Routledge.

Seixas, P. & Gibson, L. and Ercikan K. (2015). A Design process for assessing historical thinking: The case of a one-hour test. In K. Ercikan, & P. Seixas (Eds.), *New Directions in Assessing Historical Thinking* (pp.102–113). New York: Routledge.

Shanahan, T. & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: rethinking content-area literacy. *Harvard Education Review, 78*(1), 40–59. doi.org/10.17763/haer.78.1.v62444321p602101

Shemilt, D. (2015). The Validity of historical thinking assessments. In K. Ercikan, & P. Seixas (Eds.), *New Directions in Assessing Historical Thinking* (pp.246–256). New York: Routledge.

Smith, M. (2018). New multiple-choice measures of historical thinking: An investigation of cognitive validity. *Theory & Research in Social Education 46*(1), 1–34. doi:10.1080/00933104.2017.1351412

Stahl, S., Hynd, C., Britton, B., McNish, M. & Bosquet, D. (1996). What happens when students read multiple source documents in history*? Reading Research Quaterly 31*(4), 430–450.

Stolare, M. (2017). Did the Viking really have helmets with horns? Sources and narrative content in Swedish upper primary school history teaching. *Education 3–13, 45*(1), 36–50. doi: 10.1080/03004279.2015.1033439

Swedish National Agency for Education. Lgr11 (National Curriculum for primary school 2011, revised in 2018). Retrieved from: https://www.skolverket.se/undervisning/grundskolan/laroplan-och-kursplaner-for-grundskolan

Swedish National Agency for Education. National test results 2016-2017. Retrieved from; https://www.skolverket.se/om-skolverket/publikationer/prov-och-resultat.

Veijola, A. & Mikkonen S. (2016). Historical literacy and contradictory evidence in a Finnish high school setting: The Bronze Soldier of Tallinn. *Historical encounters 3*(1), 1–16.

Veijola, A. & Rantala, J. (2018). Assessing Finnish and Californian High School Students´historical literacy through a document based task. *Nordidactica – Journal of Humanities and Social Science Education. 2018:*1, 1–21. Retrieved from: http://kau.diva-portal.org/smash/get/diva2:1191139/FULLTEXT01.pdf

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. doi:http://dx.doi.org/10.1007/s10648-007-9056-1

VanSledright, B. (2002). *In search of Americas past.* New York:Teachers College Press.

VanSledright, B. (2011). The challenge of rethinking history education: on practices, theories, and policy. New York: Routledge.

VanSledright, B. (2014). Assessing Historical Thinking and Understanding. New York: Routledge.

VanSledright, B. (2015). Assessing for learning in the history classroom. In K. Ercikan & P. Seixas (eds.), *New Directions in Assessing Historical Thinking* (pp.75–88). New York: Routledge.

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. doi:10.1037/0022-0663.83.1.73

Wineburg, S. (2004). Crazy for history. *Journal of American History 90*(4), 1401–1414. https://doi.org/10.2307/3660360