UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Observing history teaching

Historical thinking and reasoning in the upper secondary classroom

Gestsdóttir, S.M.

Publication date
2021

Document Version
Other version

License
Other

Link to publication

Citation for published version (APA):
Gestsdóttir, S. M. (2021). *Observing history teaching: Historical thinking and reasoning in the upper secondary classroom*.

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:30 Nov 2021

2 Teaching historical thinking and reasoning: construction of an observation instrument[1]

1. Introduction

History teachers differ a great deal in their goals and teaching practices. Several researchers of history teaching have described at length the different approaches of dyads of (imaginary or real) teachers who at first sight seem direct opposites (Barton & Levstik, 2004; VanSledright, 2011; Wineburg & Wilson, 2001). Some of these teachers can be described as using a 'doing history' approach, which encourages historical thinking and reasoning (HTR). Active teaching approaches that foster historical thinking and reasoning skills, such as the ability to understand that history is a construct in which many perspectives play a role, have been recommended in the literature for some years. Nevertheless, many scholars also state that few history teachers adopt the 'doing history' approach, despite acknowledging its value and showing a willingness to implement it (e.g. Reisman, 2012). Among the variables that may be influencing teachers are their understanding of the construction of historical knowledge, the goals teachers aim for and curriculum requirements. Many teachers find it difficult to operationalise ideas from the literature and envisage 'what it looks like in the classroom'. Recent research in Belgium shows that although history teachers may be favourably inclined towards inquiry-based learning, they may misunderstand what this entails in regard to students' work (Voet & De Wever, 2016; 2017). A study by Wansink et al. (2016) showed that student teachers focus more on teaching historical facts and less on teaching interpretational history than the authors would have preferred.

Reflecting upon one's own teaching practices is a valuable activity in the context of professional development, and observation instruments can facilitate this. Classroom observation instruments can be useful tools to facilitate the transition from theory to practice, supporting professional development, either as part of initial teacher training or in-service training. A domain-specific observation instrument may help teachers answer the question of to what extent and in which ways they teach historical thinking and reasoning. Teacher trainers can use it when observing trainees and the trainees themselves can use the instrument

[1] This chapter has been published as: Gestsdóttir, S. M., Van Boxtel, C. & Van Drie, J. (2018). Teaching historical thinking and reasoning: Construction of an observation instrument. *British Educational Research Journal, 44*(6), 960-981. https://doi.org/10.1002/berj.3471

when observing history teachers. Experienced teachers can use the instrument as the basis for discussing their practices and developing them in order to enhance learning outcomes. Observation instruments can play a major role in teacher-led research and are recommended as one of the elements that can strengthen the links between external and internal research to maximise the transformation of evidence for practice (Nelson & O'Beirne, 2014).

This article is based on the following research question: Which teaching behaviours are characteristic of a teaching approach that stimulates historical thinking and reasoning, and how can it be observed in the classroom? We describe the development of Teach-HTR, an instrument that is specifically made for observing history lessons at the secondary level, focusing on historical thinking and reasoning. It is meant as a tool for the further professional development of experienced history teachers who wish to foster HTR, as well as to assist those who are doing their initial teacher training.

1.1 Teaching historical thinking and reasoning

What does historical thinking and reasoning consist of? In Canada, Seixas and coworkers ran the Benchmarks of Historical Thinking Project from 2006 to 2014, which was designed to foster a new way to conduct history education that is in line with recent international research on history learning. The project is based on six closely interrelated historical thinking concepts. To think historically, students need to be able to establish historical significance, use primary source evidence, identify continuity and change, analyse cause and consequence, take historical perspectives and understand the ethical dimension of historical interpretations (Seixas, 2008; Seixas & Morton, 2013). In the Netherlands, Van Drie and Van Boxtel (2008) presented a framework for the analysis of historical reasoning in the classroom in which they distinguished types of reasoning (about continuity and change, causes and consequences, differences and similarities) and several components that can be described in terms of concrete activities expressed in speech or writing (Van Drie & Van Boxtel, 2008; Van Boxtel & Van Drie, 2018): asking historical questions, using historical sources, contextualisation, argumentation, using substantive concepts and using meta-concepts (second-order concepts). In Britain, a similar shift from substantive history 'to a concern with students' second-order ideas' has also taken place since the early 1990s (Lee & Ashby, 2000), with its origins, in fact, stretching as far back as the beginning of the Schools Council History Project of the

early 1970s (Dawson, 1989). Several other models have been developed to facilitate an understanding of what historical thinking incorporates, for example in Germany (Körber, 2015). In Germany, the FUER ” (Förderung und Entwicklung reflektierten Geschichtsbewusstseins) group developed a model of historical thinking competencies, such as the ability to identify and analyse historical questions and the ability to reconstruct and deconstruct historical narratives, conceptual competencies and competencies of historical orientation (Trautwein et al., 2017). The aforementioned ‘doing history’ approach should not be conceived of as the opposite of ‘knowing history’. The two approaches seem to thrive in close collaboration, as has been supported by research in cognitive psychology (Kirschner et al., 2006). Reasoning about change and continuity, or causes and consequences, requires the use of substantive historical concepts (Van Boxtel & Van Drie, 2018). The distinction between historical knowledge and skills has been referred to as ‘a distracting dichotomy’ in Britain (Counsell, 2000; 2011) and a balanced combination of the two approaches is recommended (Havekes et al., 2012; Lee, 2005; VanSledright, 2011). The strong links between the elements put forward by all of the above make it possible to envisage what a specific teacher behaviour looks like in the classroom, whether the teacher engages students in specific tasks or not.

1.2 Domain specific observation instruments

Observation-based research on history teaching in upper secondary schools is scarce. However, there is a considerable amount of observation-based research focusing on general classroom practices, mainly at primary level. In addition, for upper secondary education, the focus has mainly been on student teachers or teachers who are relative novices, and less on the professional development of experienced teachers. An observation instrument could be helpful for experienced teachers especially, as it helps them to reflect on their regular teaching practices.

Thus, although many classroom observation tools are available, few are suited to the upper secondary level and only two are specifically made for the observation of history lessons: Protocol for Assessing the Teaching of History (PATH; Van Hover et al., 2012) and Framework for Analysing the Teaching of Historical Contextualization (FAT-HC; Huijgen et al., 2017). Van Hover et al. (2012) used the basis of CLASS (Classroom Assessment Scoring System) to develop PATH. CLASS is used to observe and assess the qualities of interactions among

teachers and children, focusing on interactional processes (CLASS, 2014). The authors of PATH call for a validated research-based observation instrument specially developed for history teaching at the secondary level. Although they stress the fact that their instrument is still being developed, they have defined six separate dimensions of history teaching they wish to take a closer look at: lesson components, narrative, interpretation, sources, historical practices and comprehension. Within these dimensions, there are individual items that apply to historical thinking and reasoning, but overall, this is beyond the scope of the instrument. Each of them is broken down into several items in an attempt to identify as precisely as possible the activities of both teachers and students. The quality of those items is then evaluated as low, middle/moderate or high on a seven-point scale, where 1–2 are low, 3–5 are middle and 6–7 are high (La Paro et al., 2004). Furthermore, an instrument for the observation of a special component of history teaching is being developed — that is FAT-HC (Huijgen et al., 2017). Contextualisation is an important part of historical thinking and reasoning, and the instrument aims to increase the understanding of history teachers' subject-specific competencies so that teacher education can better be tailored to teachers' specific needs. FAT-HC utilises a four-point Likert scale to score the items. Although these domain-specific instruments are important, neither of them is suited to providing an overview of teacher behaviour that strengthens students' historical thinking and reasoning. Thus, the need for a broader instrument that still focuses on domain-specific components of history education is evident.

2. Research questions and method

The research question is: How can the teaching of historical thinking and reasoning be operationalised and observed in upper secondary education? At this school level, students are usually older than 15 years of age, and most teachers hold a Master's degree in history before adding a teaching diploma. Both conditions enhance the likelihood of students being taught to think and to reason historically. We developed and evaluated an observation instrument in four phases: (1) literature review, (2) consultation of experts, (3) first pilot of the instrument and (4) second pilot of the instrument. The content validity was evaluated by experts using a content validity rating form. Inter-rater reliability was evaluated using intraclass correlation coefficients (ICCs) and percentage of agreement. Internal consistency was evaluated using Cronbach's alpha. Furthermore, we looked at lessons with a high and a low score in order to

explore the potential for the instrument to give teachers feedback about what they are already doing and the points at which there is room for development. In this section, we discuss the methods used; in the results section, we elaborate upon the outcomes of each of the steps taken.

2.1 Literature review

First, we conducted a literature review to identify concrete teaching behaviours and activities that students are engaged in when learning historical thinking and reasoning. The cases we found in the literature were used to define items and to organise these items into meaningful categories.

2.2 Consultation of experts

Second, the categories and items were first discussed with a group of 13 Dutch history educators and PhD students in history education, and three videotaped lessons of history teachers in upper secondary education in Iceland were used to improve the first version of the instrument. From these lessons, we derived examples for each item that were included in the user instructions for the instrument. Then, the instrument (including the instructions with examples for each item) was validated by asking 11 experts to assess the clarity and importance of the instrument's categories, items and examples (see also Hyrkäs et al., 2003). The experts were researchers, teacher educators and teachers within the domain of history, and some of them led the field of research on historical thinking and reasoning. The experts came from eight different countries: Belgium, Canada, Finland, Germany, Iceland, the Netherlands, the UK and the USA. They were asked about the importance of each category and item of the instrument (on a four-point scale from 'not at all important' to 'very important'), as well as the clarity of each item (from 'not at all clear' to 'very clear') and the accompanying examples. The experts could also make other remarks, such as suggestions for other aspects to be included in the instrument or the reformulation of the items and the accompanying examples. These steps resulted in some revisions of the formulation of the items in the observation instrument. Moreover, new examples were included in the instruction.

2.3 First pilot of the instrument

Third, in this first pilot, the inter-rater reliability and internal consistency of the instrument were assessed using a set of 10 videotaped history lessons in Iceland. We compared the observed lessons to investigate whether the instrument was helpful in discerning, among the lessons, those in which teachers show behaviour that is believed to enhance the historical thinking and reasoning of students and those that include less of these kinds of teaching behaviours. Data were collected from September to November 2014 in 12 Icelandic upper secondary schools. The 10 videotaped lessons were of four male and four female history teachers, randomly chosen from a larger sample of videotaped lessons of 27 teachers in Iceland. Their ages range between 32 and 60, with an average of 45.5. Teaching experience ranged from 4 to 34 years, with an average of 17 years. The length of the lessons varies from 40 to 50 minutes. The video data enable coding by a second coder (Waldis & Wyss, 2012). Two coders (the first author and a student who was participating in a Master's programme for history teaching at a university in Iceland) met for training. Five lessons (also chosen from the larger sample) were then coded separately and discussed carefully before proceeding towards the 10 that were used to assess inter-rater reliability. Each coder scored each of the seven categories of the observation instrument on a Likert scale of 1–4. To support this evaluation, the coder checked the behaviours observed. For example, the behavioural indicators of 'demonstrating historical thinking or reasoning' (category 2) are asking historical questions, providing historical context, making clear that contemporary standards should be avoided, explaining, discerning change and continuity, comparing and assigning historical significance (subitems 3–9, see Appendix 1). A category is scored 1 (not at all) when none of the behavioural indicators is observed. For each category, the instructions provided guidelines for scoring. For example, for the categories 'the use of historical sources' and 'the provision of explicit instruction of HTR strategies', a score of 2 (a little) means that only one of the behavioural indicators is observed. A score of 3 (to some extent) is chosen when the coder observes more indicators. A score of 4 (to a large extent) is chosen when the coder observes a high number of (different) behavioural indicators or considerable attention is paid to some indicators. In the case of category 7, it is difficult to determine the score (1–4) on the basis of how many different indicators are present. For this category, we focused on the amount of attention being paid to the indicator(s). For example, a score of 4 is given for the category 'engaging students in whole-class discussion' (category 7), when the teacher engages students

in a comprehensive whole-class discussion in which students are very much provoked to think/reason historically.

We used percentage agreement to calculate inter-rater agreement for the seven categories. Because a low incidence of the behaviour of interest can result in an artificially inflated percentage agreement, we also calculated ICCs for each category. We applied a two-way random model with absolute agreement and looked at the ICCs for single measures (Hallgren, 2012). We wanted the inter-rater reliability to be characterised by absolute agreement in the ratings, instead of coders providing scores that are similar in rank order. We looked at the ICC for 'single measures' because this ICC reflects the reliability of the ratings provided by a single coder, instead of the average of multiple coders. We considered <0.40 to indicate poor agreement, 0.41– 0.59 to indicate fair agreement, 0.60–0.74 to indicate good agreement and 0.75–1.00 to indicate excellent agreement (Cicchetti, 1994). To assess the internal consistency for the seven categories, we used Cronbach's alpha. We considered 0.70 to be an acceptable reliability coefficient. Although, when training or coaching teachers, a mean score for the seven categories of the instrument is not very useful, internal consistency informs us about the suitability of the instrument for comparing lessons or teachers or for relating the teaching of HTR to teacher characteristics and learning outcomes.

2.4 Second pilot of the instrument

Fourth, the instrument was used to analyse 10 history lessons of eight history teachers in upper secondary education in the Netherlands. The observations in the first pilot were used to add some examples to the instruction document. We decided to add the rating of an extra sample of lessons in another country because it appeared from the first pilot that a substantial number of items of the instrument were not observed in the 10 previous ones in Iceland, and the interrater reliability was rather low for some categories. The 10 lessons that were used in the second pilot were given in April to June 2016 by experienced history teachers in the Netherlands. They were recruited using the network of the second and third authors. These teachers participated in previous studies on historical thinking and reasoning and/or in professionalisation workshops on historical thinking and reasoning. They were expected to show a larger variety of behaviours included in the observation instrument. The teachers involved were seven male history teachers and one female history teacher from seven

different schools spread over the Netherlands in both rural and urban areas. One of the observed teachers also works as a teacher trainer. The age of the teachers ranged from 39 to 64 (mean 49.9) and they had 3–37 years of teaching experience (mean 19.4).

The length of the lessons was 45–50 minutes. For this second set of videotaped lessons, we decided to follow the rating of two teacher trainers, as they will probably use the instrument more than any other group of teachers. In addition to student teachers (who can observe teachers during their internship), the instrument is meant to be used by teacher trainers who can discuss the results with student teachers or experienced teachers in the context of training or further professional development. The coders in this pilot were experienced teacher trainers with a range from some to extensive research experience. They used one videotaped lesson to discuss their coding. This lesson was not part of the sample of lessons that were used to determine internal consistency and inter-rater reliability. We calculated ICCs for the seven main categories and Cronbach's alpha following the same procedure as in the first pilot. In this pilot, we also calculated the ICC for the scale as a whole.

3. Results

3.1 Literature review: Definition of categories and items

The literature review produced the main elements of HTR as perceived by experts in the field of history teaching. This led to an instrument consisting of six categories and an initial set of 32 items. The categories of the instrument are in line with common lesson components, such as a specification of lesson goals, the presentation of new material through instruction or explanations and the activation of students through individual seatwork, group work or whole-class discussions. Some lessons will be more teacher-centred, while others might be more student-centred. In both types of lessons, teachers can aim at the development of historical thinking and reasoning skills. Our initial instrument included one category for 'actively engaging students in historical thinking and reasoning'. After the expert meeting, we split this category in two: one focusing on individual and group tasks and one focusing on whole-class discussion. Below, we will describe how the seven categories are grounded in the literature.

Communicating objectives related to historical thinking and reasoning. Historical thinking and reasoning requires an understanding of second-order concepts (e.g. cause, change or

evidence) and knowledge of how, for example, to explain or critically assess historical sources (e.g. Lee, 2005; Nokes et al., 2007; Stoel et al., 2015). It also requires the understanding that history is always interpretation (e.g. Chapman, 2011; Maggioni et al., 2006). Historical thinking and reasoning not only contribute to a deep understanding of historical phenomena; in order to develop historical reasoning, students must deeply understand historical facts, concepts and chronologies (Van Boxtel & Van Drie, 2013). When teachers teach historical thinking and reasoning, they can aim at developing this knowledge and understanding and informing students about their particular goals. The items that are part of this category reflect these goals. For example: 'In this lesson we will look at how to critically assess sources. We will work with a format of how you can evaluate sources.'

Demonstration of historical thinking and reasoning. Based upon the historical thinking concepts of Seixas and Morton (2013) and the components of historical reasoning of Van Boxtel and Van Drie (Van Drie & Van Boxtel, 2008; Van Boxtel & Van Drie, 2013; 2018), we identified several ways in which teachers can demonstrate thinking and reasoning themselves when they explain new content or give instructions. Teachers can ask historical questions, contextualise, take a historical perspective, explain historical phenomena, discern aspects of change and continuity, compare historical phenomena or periods or assign historical significance. The items in this category include these activities. Historical thinking and reasoning related to the use of historical sources is a separate category, since it is a vast and well-researched field in itself (see below). An example is when the teacher says: 'Why did so many people die during this period?' (asks historical questions) or 'Although many things changed, it was still the nobility who had the power' (discerns aspects of continuity and change).

The use of sources to support historical thinking and reasoning. Using historical sources is an important component of doing history. The items within this category do not refer to the use of sources as illustrations but describe activities, such as sourcing, contextualisation, close reading and comparison of information from different sources (see Monte-Sano, 2011a; Reisman, 2012; Wineburg, 1991). The teacher can evaluate the usefulness of a source in relation to a specific question and refer to the role of sources as evidence in an interpretation or argument (e.g. Levesque, 2008; Seixas & Morton, 2013; Wiley & Voss, 1999). An example is when the teacher asks: 'What does the presence of a skeleton in this painting tell

us?' (close reading of sources) or 'Does her letter shed light on the conditions of the emigrants?' (evaluates the usefulness/reliability of a source).

Presenting multiple perspectives and interpretations. In history, there are always multiple perspectives. The items in this category include different types of multi-perspectivity, for example, at the level of historical agents (e.g. how they perceived a particular event), different dimensions of society (e.g. economic or political), scale (e.g. local or global) and historical interpretations (e.g. Chapman, 2011; Lee & Shemilt, 2004; Seixas & Morton, 2013; Stradling, 2003; Van Hover et al., 2012). An example of presenting the perspectives of different historical actors is: 'This negative account comes from their neighbours and enemies who were not impressed by their endeavours.'

Explicit instructions on historical thinking and reasoning strategies. Explicit instruction is one of the strategies that is advocated when aiming at the development of generic and domain-specific strategies (see Bain, 2000; Nokes et al., 2007; Reisman, 2012; Stoel et al., 2017). The teacher can, for example, give explicit instruction on how to explain historical phenomena, evaluate historical sources or assign historical significance. The teacher can also make a remark about the nature of history or the construction of historical knowledge. An example is when the teacher provides instruction about how to explain ('Remember that we always need to search for multiple causes and think about different types of causes that may play a role. Can you think of an economic cause?') or how to assign historical significance ('Okay but we talked about the five criteria for establishing historical significance. What were they again?').

Engaging students in individual or group tasks that require historical thinking and reasoning. The 'doing history' approach emphasises that students should be actively engaged in historical thinking and reasoning (e.g. Barton & Levstik, 2003). This asks for learning tasks and activities in which students can apply historical knowledge and strategies (e.g. Nokes et al., 2007; Havekes et al., 2010; Reisman, 2012; Seixas & Morton, 2013; Van Boxtel & Van Drie, 2013). An example is an assignment that asks for historical thinking and reasoning activities, such as historical contextualisation: students have to draw on previous knowledge and use sources to figure out why it was particularly serious in the seventeenth century for a married woman to get pregnant by another man. Tasks may also engage students in the evaluation and analysis of historical sources and/or argumentation. For example, an

assignment that asks students to defend the stance that Columbus was not the discoverer of America using historical evidence.

Engaging students in a whole-class discussion that asks for historical thinking and reasoning. In the instrument, we make a distinction between individual or group tasks that engage students in historical thinking and reasoning and whole-class discussions that aim at prior knowledge activation, a deeper understanding of a particular topic or a debriefing of individual or group tasks that require historical thinking and reasoning (Van Drie & Van Boxtel, 2011; Havekes et al., 2017). An example is when several students working on modern China discussed the success of the one-child policy. Teacher: 'Would you conclude, from this information [pointing at a graph], that the one-child policy was successful?' Student 1: 'They did what they were trying to do but the cost of it...' Student 2: 'I think it was more a sort of contraception, rather than anything else.' Student 3: 'Isn't it more about the development of the country as a whole?' Teacher: 'As a whole?' Student 3: 'People are having more education.' (Whole-class discussion in which students are provoked to think/reason historically in order to deepen a particular topic.)

3.2 Results of expert consultation

The consultation of international experts resulted in some refinement of the examples of teacher behaviour as well as a stronger emphasis on how teachers demonstrate certain elements of HTR. The experts provided detailed comments and made many suggestions for improvement. In general, they considered all categories and items to be important, either 'somewhat important' but almost always 'very important'. Only once did an expert consider a category 'not at all important' whereas all the others ticked 'very important', and twice an expert considered a category 'not too important' while all the others considered it 'somewhat' or 'very important'. In some cases, experts asked for simpler coding (e.g. by splitting up items or making them more concrete). They also stressed the need for more elaborate examples to enhance the clarity of the items in question. This resulted in considerable clarification as repetitions between categories were abolished, and also increased the consistency in the vocabulary as certain concepts were highlighted and others cleared away. Because the category 'actively engaging students in historical thinking and reasoning' contained a broad variety of ways to actively engage students, we split it in two. Examples of items that were added are 'The teacher makes it clear that people in the past thought differently than we do

now' (category: demonstrates HTR) and 'The teacher provides explicit instructions on how to contextualise the events or actions of people in the past/take a historical perspective' (category: explicit instruction). Several accompanying examples were changed, and a number of new ones added according to the suggestions of the experts. Appendix 1 shows the items of the instrument that was piloted in Iceland and the Netherlands.

3.3 Results of the first pilot

Table 1 shows the percentages of agreement between the two coders, which ranged from 60% to 90%. The ICCs range between 0.23 and 0.72 (see Table 1). Good agreement (ICC = 0.72) was reached for the category 'engaging students in individual or group tasks that ask for HTR'. There was fair agreement (ICC = 0.59) for 'the use of historical sources'. In one case, the first coder observed three items in this category, whereas the second coder observed that the teacher used sources only to illustrate content. Fair agreement (ICC = 0.51) was also reached on 'demonstration of HTR'. In two lessons, the first coder observed more items that were part of this category than the second coder. There was poor agreement (ICC = 0.36) on providing multiple perspectives or interpretations. The coders, for example, disagreed on whether the behaviour 'presents and explores perspectives of different historical actors'. This might have to do with the difficulty of making a distinction between mentioning several historical actors and making clear the different perspectives of these actors. Poor agreement (ICC = 0.23) was also reached about the communication of objectives that focus on historical thinking and reasoning. It appeared difficult to make a distinction between a teacher communicating what the lesson is about and communicating objectives that focus on a deeper understanding of the topic. For the categories 'explicit instruction' and 'engaging students in a whole-class discussion that asks for HTR', no ICC could be calculated, as it was only scored once by one coder. For 9 out of 10 lessons, the two coders agreed that there was no explicit instruction on HTR strategies. In 8 lessons, the coders agreed that they did not observe a whole-class discussion in which students were engaged in HTR. When there was a difference between the two coders, in almost all cases, the first coder assigned a higher score than the second coder. The fact that the first author is an expert on teaching historical thinking and reasoning and has more teaching and observation experience might have accounted for this finding. She probably identified more easily than the student teacher certain behaviours that can be considered the teaching of HTR.

Table 1

Intra-Class Correlation Coefficients (ICC's) for the rating of ten lessons in upper secondary education in Iceland by two coders

Categories	% agreement	ICC
1. The teacher communicates learning objectives that focus on historical thinking and reasoning goals	70	.23
2. The teacher demonstrates historical thinking or reasoning	70	.51
3. The teacher uses historical sources to support historical thinking and reasoning	80	.59
4. The teacher makes clear that there are multiple perspectives and interpretations	60	.36
5. The teacher provides explicit instructions on historical thinking and reasoning strategies	90	*
6. The teacher engages students in historical thinking and reasoning by individual or group assignments	70	.72
7. The teacher engages students in historical thinking and reasoning by a whole class discussion	80	*

* ICC could not be calculated

Using the scores of the first coder, Table 2 shows the coding of the 10 lessons. Although we have to take into account that for some categories we did not reach sufficient inter-rater reliability, the table shows that the categories that operationalised the teaching of historical thinking and reasoning were hardly observed in this sample of 10 lessons.

We computed the internal consistency of the scale with seven categories (using the codes of the first coder) using Cronbach's coefficient alpha. The scale reached an internal consistency of 0.61.

Table 2

*Scores on a 4 point Likert scale for 10 lessons observed in Iceland on the seven categories of the observation instrument**

Lesson	Objectives	Demon-strating	Using sources	Multiple persp.	Explicit instruction	Assign-ments	Whole class discussion	Mean
1	2	3	2	3	1	1	1	1.86
2	1	3	1	2	1	2	1	1.57
3	1	4	1	2	1	1	1	1.57
4	2	2	1	2	1	1	1	1.43
5	1	3	2	1	2	1	1	1.57
6	2	3	3	1	1	1	1	2.14
7	1	3	1	1	1	1	1	1.29
8	1	2	1	1	1	1	1	1.14
9	3	2	3	2	1	4	4	2.71
10	2	3	1	2	1	1	3	1.86
Mean	1.60	2.80	1.60	1.70	1.10	1.40	1.50	

* We used the scores of the first coder.

In the lesson with only one HTR category observed (lesson 8), the teacher was discussing political science in a wide historical context. In the lesson with the highest mean score (lesson 9), the teacher demonstrated historical thinking and reasoning, used sources to support historical thinking and reasoning and made clear that there are multiple perspectives, as well as actively engaging students in a group assignment that asked for historical thinking and reasoning. The class had been working on World War I, and in this lesson all the threads were being tied together as they conducted a role play on the Treaty of Versailles. She had already provided several sources, both texts and photographs, to contextualise and enable the students to see different national perspectives (e.g. to differentiate between the views of the leaders of

the USA, Britain and France towards the defeated Germany—category 4). She encouraged a close reading of the sources instead of giving the students the answers they were looking for (e.g. when a student from the group representing the UK wondered why they should be compliant towards Denmark—category 3) and put historical questions to the students (e.g. why were some of the participants angry?—category 2). Apart from that, the teacher stayed outside the process, leaving the students to make their own corrections in response to questions that the teacher posed as she circulated. The students themselves had to, for example, correct incidents of presentism, such as when someone referred to what would happen later on in World War II.

3.4 Results of the second pilot

Because, in the first pilot, we experienced that in some lessons a large number of items were not observed and it appeared difficult to reach substantial agreement on some categories, we conducted a second pilot. The observations in the first pilot were used to add some examples and more elaborate directions for coding to the instructions document. The coders in this pilot were experienced teacher trainers with a range from some to extensive research experience. Table 3 shows the percentages of agreement and the ICCs for the seven categories of the instrument. In this pilot, all 33 subitems were observed. The percentage of agreement ranged from 10% for ‘making clear that there are multiple perspectives’ to 70% for several categories. Good agreement (ICC = 0.70) was reached for the total score (mean score of the seven categories). Table 3 shows that excellent agreement was reached for the categories ‘communicating learning objectives’ (ICC = 0.77), ‘demonstrating HTR’ (ICC = 0.83), ‘using historical sources’ (ICC = 0.80), ‘explicit instruction’ (ICC = 0.85) and ‘engaging students in individual or group tasks that ask for HTR’ (ICC = 0.86). Fair agreement (ICC = 0.55) was reached for ‘engaging students in a whole-class discussion that asks for HTR’. Poor agreement (ICC = 0.27) was reached for ‘making clear that there are multiple perspectives’.

We inspected the differences between the coders for the category about multiple perspectives. In five lessons, coder 2 observed more items belonging to the category than the other coder. In two lessons, for example, coder 2 observed ‘presents two or more perspectives: economic/political/sociocultural’, whereas the other coder did not observe this behaviour. In two other lessons, coder 2 observed ‘presents different historical interpretations, for example,

of causes and consequences, change and historical significance or shows that interpretations change through time', whereas the other coder did not.

In some cases, the two coders evaluated the whole-class discussion differently. In three lessons, one of the coders assigned a score of 2, whereas the other assigned a score of 3. In these cases, it appeared difficult to make a distinction between more or less comprehensive whole-class discussions in which students are engaged in HTR.

Table 3

Intra-Class Correlation Coefficients (ICC's) for the rating of ten lessons in upper secondary education in the Netherlands by two coders

Categories	% agreement	ICC
1. The teacher communicates learning objectives that focus on historical thinking and reasoning goals	50	.77
2. The teacher demonstrates historical thinking or reasoning	70	.83
3. The teacher uses historical sources to support historical thinking and reasoning	60	.80
4. The teacher makes clear that there are multiple perspectives and interpretations	10	.27
5. The teacher provides explicit instructions on historical thinking and reasoning strategies	70	.85
6. The teacher engages students in historical thinking and reasoning by individual or group assignments	70	.86
7. The teacher engages students in historical thinking and reasoning by a whole class discussion	40	.55

The coding in the second pilot also raised some questions about the co-occurrence of categories of the instrument. In some cases, two categories were observed during the same classroom activity. For example, during a whole-class discussion, the teacher compared past and present phenomena and thus also 'demonstrated historical thinking and reasoning'. When debriefing an assignment in which the students had to describe a process of change, the

teacher also provided explicit instruction on how to identify historical change. When a coder is only focusing on the category 'whole-class discussion', the items belonging to other categories can easily be missed. Therefore, we made this clear in the users' instructions. Furthermore, we decided to include extra instruction on coding of the subcategory 'explicit teaching'. Explicit instruction is not necessarily a comprehensive instruction that takes a substantial part of the lesson, but can also consist of only a few utterances. For example, when discussing an assignment about processes of change, the teacher can remark that it is important to note that changes are often long-term processes that take more than 100 years.

We computed the internal consistency of the scale with seven categories (using the codes of the first coder) using Cronbach's coefficient alpha. The scale reached an internal consistency of 0.82. Deleting the category about making clear that there are multiple perspectives (because of its low inter-rater agreement) resulted in an alpha of 0.79, which can be considered sufficient.

Using the codes of the first coder, for each of the 10 lessons in the Netherlands, we calculated the mean score for the seven categories (see Table 4).

Table 4 shows that in lessons 2, 4 and 9, the teaching of historical thinking and reasoning, as operationalised in the instrument, was hardly observed (a mean score of 2 or lower). Two of these lessons were characterised by a teacher-centred approach, while the third was more student-centred. In lesson 4, for example, the teacher taught about the relations between China and Europe during the Middle Ages. He mainly lectured, although he also asked his students many questions. However, there was no real whole-class discussion because the students were not encouraged to respond to each other or to elaborate their answers. His instruction was very rich in terms of demonstrating historical thinking and reasoning (category 2). He asked historical questions, identified aspects of change and continuity (e.g. China closing its doors to the outside world after a long period of international trade), discussed causes and consequences (e.g. why there was less trade) and explained the significance of historical developments and events (e.g. the fact that there is still quite a large Muslim community in China going back to the Middle Ages when Arab traders brought Islam to China).

In two lessons (lessons 7 and 8), a considerable amount of teaching of historical thinking and reasoning was observed (a mean score above 3). Lesson 7 was from an experienced history teacher who also worked as a teacher trainer. The teacher gave students three historical

documents from different countries (Declaration of Independence, 1776; Bill of Rights, 1688; Act of Abjuration, 1581), which the students had to compare, looking at what was meant by 'the people' and 'freedom'. Furthermore, they had to explain what these sources had to do with the Enlightenment and whether the Declaration of Independence could be considered a democratic revolution. In the second half of the lesson, the teacher guided a whole-class discussion to debrief the assignment. Several students actively participated in this discussion, and together they collaboratively reasoned about how the thinking about freedom and equality had changed through time. The fragment below shows part of this whole-class discussion.

Table 4

Scores on a 4 point Likert scale for 10 lessons observed in the Netherlands on the seven categories of the observation instrument

Lesson	Objectives	Demon-strating	Using sources	Multiple persp.	Explicit instruction	Assign-ments	Whole class discussion	Mean
1	1	1	3	1	3	3	2	1.86
2	1	4	2	1	1	2	2	1.86
3	3	4	4	3	2	3	2	3.14
4	1	4	3	2	1	1	1	1.86
5	1	4	3	2	2	3	3	2.43
6	2	3	3	2	2	2	3	2.14
7	4	4	4	3	4	4	4	3.86
8	4	3	4	2	3	4	3	3.42
9	4	2	1	2	1	2	2	2.00
10	1	2	2	2	2	4	3	2.14
Mean	2.20	3.10	2.90	2.00	2.10	2.80	2.50	

* We used the scores of the first coder.

Together with the students, the teacher identified consequences of the Enlightenment and aspects of change and continuity. The teacher closed the lesson by providing explicit

instruction about how to answer a question about whether something can be considered revolutionary or not. He emphasised the importance of using historical concepts (e.g. estates, absolutism, not in the example) and seeing an event as part of a long-term process:

Example 1

A teacher engages students in a whole class discussion that asks for historical thinking and reasoning.

Teacher	Vala says that, and I can already say that she describes very well a consequence of Enlightenment thinking. Because we look at equal opportunities for people, the qualities of people, and the second thing you told us, the last sentence, what were you saying again?
Vala	That people got more equal opportunities.
Teacher	And that, as a result of that, people got more equal opportunities, it was a consequence of Enlightenment thinking. Karl?
Karl	That is correct, because in the source, the source tells that subjects, people are more equal, so to say, there were less social inequalities, and then people were considered people instead of subjects.
Teacher	Okay.
Karl	During the Enlightenment, previously they didn't do that, previously they were more, they defended social classes and when the Enlightenment came, they became an estate, and everyone has to be equal.
Teacher	Okay, together we work this out, it goes very well.
Einar	You can say that in the past it was very strict, there really were estates, they stayed within these estates, and you can notice that little by little it went further, people descended from how do you call it, they were going to believe less in God and thought more about themselves and they saw more and more (incomprehensible)
Teacher	That's what we often discussed, didn't we? They started to believe in God differently, religion continues to be very important.

…

Teacher It is important that you understand exactly what has been said, that it goes step by step. That they didn't wake up on July 4th and think, let's make up something completely new. It took two hundred years, maybe more, these are the characteristic features that already started before and then, at this point, that's all coming together. Try to do it in a nuanced way.

4. Conclusion and discussion

Historical thinking and reasoning is an important goal in secondary history education. However, teaching historical thinking and reasoning is quite challenging for teachers. An observation instrument might be helpful to advance teachers' professional development in this respect, but thus far, no such instrument for the teaching of historical thinking and reasoning has been available. The purpose of this study was to discover which teaching behaviours are characteristic of a teaching approach that stimulates historical thinking and reasoning and to determine how it can be observed in the classroom. Based on the literature, this study and feedback from experts using a content validity rating form, we developed the Teach-HTR observation instrument consisting of seven categories and 33 items. After the first pilots, the instrument is promising regarding inter-rater reliability, which was evaluated using the ICCs and internal consistency, which was evaluated using Cronbach's alpha. Relatively low levels of inter-observer agreement have been found in several studies on observation instruments (e.g. Strong et al., 2011). Therefore, using observation instruments to assess the quality of teachers or schools is problematic. Since the instrument Teach-HTR is not intended for the assessment of the quality of teachers but for the professional development of experienced history teachers or for teacher trainees, moderate agreement is sufficient. Even so, at the levels of the seven individual items, the interrater reliability for the category 'making clear that there are multiple perspectives' appeared insufficient. This category might be more difficult to observe because it represents very different types of multi-perspectivity (e.g. perspectives of historical actors and perspectives of historians) and because the perspectives can be present without being verbalised explicitly, or a second perspective might be introduced much later. More extensive training of the raters, and better examples in the manual, may result in a sufficient level of inter-rater reliability. The training needs to be very

carefully conducted, as the authors of other observation instruments have concluded as well (e.g. Van Hover et al., 2012). More research is needed to improve this part of the instrument.

Despite these specific difficulties, Teach-HTR immediately revealed a considerable difference between lessons. It was very easy to spot the lessons in which historical thinking and reasoning were promoted and those in which this was hardly visible. We found differences in the way and the extent to which history teachers showed the behaviour that we defined as teaching HTR. In some lessons we found, for example, that teachers were demonstrating historical thinking and reasoning and used historical sources when lecturing but hardly actively engaged students in HTR. Although several studies found positive effects of explicit teaching of HTR strategies (e.g. Nokes et al., 2007; Reisman, 2012; Stoel et al., 2017), in the lessons that we observed, hardly any teacher demonstrated this behaviour. These findings show that the instrument has the potential to identify (student) teachers' strengths and room for development. The identification of concrete examples of teaching HTR can help teachers to further develop their ability to teach HTR and integrate it into more lessons, should they so wish.

The instrument in its current form can be a practical tool in the context of professional development or initial teacher training to discuss with students/teachers examples of teaching HTR and to evaluate particular lessons. Using the results of observations with Teach-HTR as a basis for discussion with experienced history teachers has already been attempted in several interviews by the first author. It turned out that teachers were not too keen on watching a whole lesson, but going back to specific instances of the teaching of HTR or missed opportunities during the lesson helped them to reflect on their behaviour. In future research, we want to investigate this potential. The observations might be used, for example, in post-observation conferences with students or experienced teachers or in the context of lesson study in which history teachers design and implement lessons that aim at the development of students' HTR abilities (e.g. Halvorsen & Kesler Lund, 2013). A recent study in the UK with secondary school teachers showed that only observing or being observed did not result in better English and maths scores (Worth et al., 2017). However, in this study, there were no requirements for post-observation discussions to take place. Opportunities for practice with possibilities for feedback and the availability of materials and resources are mentioned as important characteristics of effective professional development programmes (Van Veen et al., 2012).

We want to emphasise that the instrument is not meant to assess teachers' ability to teach historical thinking and reasoning. More research would be necessary to determine how many lessons need to be observed in order to draw valid conclusions about the teaching of an individual teacher or to investigate how the teaching of HTR correlates with teacher beliefs or student performance and interest in history. Our samples were small—10 history lessons in Iceland and another 10 in the Netherlands. However, it is easy to envisage that in future research Teach-HTR might also be usable to make comparisons between countries. National curricula might influence if and how teachers promote historical thinking and reasoning. Such requirements are hardly present in the Icelandic curriculum, which may explain why several items of the instrument were not observed at all in the 10 lessons that the two observers coded. Building upon the promising results of this study, we hope to further develop the Teach-HTR observation instrument as a useful tool for teachers who want to improve their teaching of historical thinking and reasoning.