Eur J Psychol Educ (2017) 32:159–181
DOI 10.1007/s10212-016-0295-8

Teaching historical contextualization: the construction of a reliable observation instrument

Tim Huijgen[1] **· Wim van de Grift**[1] **· Carla van Boxtel**[2] **· Paul Holthuis**[1]

Received: 21 October 2015 / Revised: 22 January 2016 / Accepted: 18 February 2016 /
Published online: 28 March 2016

Abstract Since the 1970s, many observation instruments have been constructed to map teachers' general pedagogic competencies. However, few of these instruments focus on teachers' subject-specific competencies. This study presents the development of the *Framework for Analyzing the Teaching of Historical Contextualization* (FAT-HC). This high-inference observation instrument focuses on history teachers' competency in promoting historical contextualization in classrooms. The results of the study demonstrate the instrument's content validity. Generalizability studies were conducted to further assess the instrument's dimensionality and reliability by decomposing the instrument's variance. A large proportion of the variance was explained by differences between observed teachers, and a small proportion of the variance was explained by lessons and observers, demonstrating the instrument's reliability. Furthermore, a decision study was conducted to determine the optimal number of observers and lessons needed for a reliable scoring design. The developed instrument could be used to gain greater insight into history teachers' subject-specific competencies and to focus teacher professionalization on teachers' specific needs.

Keywords Historical reasoning · Teacher education · History education · Historical contextualization · Observation instrument

Since the 1970s, increasing attention to the evaluation of teachers' generic competencies has resulted in the development of a variety of observation instruments that are widely used to assess elementary and secondary education, such as the *Stallings Observe System* (Stallings and Kaskowitz 1974), the *Framework for Teaching* (Danielson 1996), the *International*

✉ Tim Huijgen
t.d.huijgen@rug.nl

[1] Department of Teacher Education, Faculty of Behavioural and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

[2] Research Institute of Child Development and Education and Amsterdam School of Historical Studies, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS Amsterdam, The Netherlands

System for Teacher Observation and Feedback (Teddlie et al. 2006), the *International Comparative Analysis of Learning and Teaching* (Van de Grift 2007), and the *Classroom Assessment Scoring System* (Pianta et al. 2008). Other instruments used to examine teacher behavior include, for example, teachers' self-reports, (semi-structured) interviews, and student questionnaires (e.g., Kyriakides et al. 2002; Kyriakides 2008; Maulana et al. 2015; Muijs 2006). However, despite its labor-intensive nature, classroom observation is viewed as a more unbiased form of data collection to examine teacher behavior (Pianta and Hamre 2009; Wragg 1994).

The development and implementation of observation instruments can be very useful in more effectively shaping teacher education and professional development programs and in evaluating classroom-based interventions (e.g., Darling-Hammond 2012; Lavigne and Good 2015; O'Leary 2014; Yoder and Symons 2010). However, most of these instruments focus on teachers' generic competencies rather than teachers' subject-specific competencies. Therefore, scholars such as Grossman and McDonald (2008), Desimone (2009), and Schoenfeld (2013) emphasized the importance of adding subject-specific observation instruments to research on teaching and teacher education.

Although some recently developed observation instruments focus on more specific teacher competencies, such as classroom talk (Mercer 2010), project-based learning (Stearns et al. 2012), and the reform of learning and instruction (Sawada et al. 2002), only a few observation instruments focus on teachers' subject-specific strategies, such as English reading (Gertsen et al. 2005), content- and language-integrated learning (De Graaff et al. 2007), English language arts (Grossman et al. 2010), and mathematical instruction (Hill et al. 2012; Matsumura et al. 2008; Schoenfeld 2013).

To date, however, there are no validated and reliable observation instruments that evaluate secondary-school history teaching. This is unfortunate, especially because, as noted by Bain and Mirel (2006), Grant and Gradwell (2009), and Achinstein and Fogo (2015), current teacher education and professional development programs may not meet history teachers' needs so that they can achieve the aims set by history curricula. Observation instruments that evaluate history teachers' subject-specific strategies could identify history teachers' specific needs and, thus, further improve teacher education and professional development programs for history teachers.

Van Hover et al. (2012) attempted to construct a validated observation instrument to evaluate secondary-school history teaching. Their *Protocol for Assessing the Teaching of History* (PATH) is promising, but information about the measure's reliability is lacking. In contrast to PATH, the observation instrument that we developed focuses on a single but highly important history teacher competency, promoting students' ability to perform historical contextualization. Historical contextualization is considered an important component of historical thinking and reasoning and is incorporated into history curricula worldwide (Lévesque 2008; Seixas and Morton 2013; Van Drie and Van Boxtel 2008). In previous research, we examined how students performed on a historical contextualization task and found that secondary-school students of different ages experience difficulties in performing historical contextualization tasks (Huijgen et al. 2014).

Therefore, we must gain greater insight into how history teachers promote students' ability to perform historical contextualization in classrooms. The purpose of the present study is, therefore, to construct a reliable high-inference observation instrument and scoring design to assess history teachers' competency in promoting historical contextualization in classrooms. In this study, we first present the theoretical framework and our research questions. Then, we

present our methodology and results. Finally, we discuss our findings and present the practical implications of the results and directions for future research.

Theoretical framework

Teaching historical reasoning competencies

Scholars and other educational professionals widely agree that secondary-school history education should involve more than the simple learning of facts (e.g., Lévesque 2008; Van Drie and Van Boxtel 2008; Wineburg 2001). Therefore, historical reasoning competencies, such as determining causality, investigating sources, asking rich historical questions, and performing historical contextualization, have become increasingly important in Western history education over the last two decades (Erdmann and Hasberg 2011; Seixas and Morton 2013). Some scholars also stress the importance of historical reasoning competencies for promoting students' democratic citizenship (e.g., Barton 2012; Saye and Brush 2004). To achieve historical reasoning competencies, students in history classes must be involved in engaging learning tasks and activities (Levstik and Tyson 2008; Gerwin and Visone 2006; Grant and Gradwell 2010) and history lessons should extend beyond factual recall to achieve deep subject understanding (Bransford et al. 2000; Darling-Hammond et al. 2009).

However, both novice and experienced history teachers seem to struggle when they are asked to develop engaging learning tasks and teach students historical reasoning competencies (Monte-Sano 2011; Van Hover and Yeager 2004; VanSledright 2010; Virta 2002). Many history lessons might, therefore, have a strong focus on historical content knowledge (Saye and Social Studies Inquiry Research Collaborative (SSIRC) 2013; VanSledright 2011).

Observing history education

To explore the challenges and problems that history teachers face, qualitative research studies have been conducted (e.g., Bain and Mirel 2006; Fogo 2014; Monte-Sano and Cochran 2009; Virta 2007). However, few quantitative research studies using standardized instruments have been conducted to explore history teachers' competencies (Adler 2008; Ritter 2012). For example, only two studies used observation instruments to examine how teachers actually teach historical content knowledge and historical reasoning competencies. Thus, the use of standardized observation instruments in research on history education is an underexamined topic, as Van Hover et al. (2012) noted.

> While the field of history education elucidates a clear and ambitious vision of high-quality history instruction, a current challenge for history educators (including teacher educators, curriculum specialists, and school-based history and social science supervisors) becomes how to illuminate and capture this when observing classrooms to research history instruction or to provide useful discipline-specific feedback to preservice (and inservice) history teachers (p. 604).

Nokes (2010) used an observation instrument and focused on history teachers' literacy-related decisions about the types of texts they used and how students were taught to learn with these texts. Eight secondary-school history teachers were observed over a 3-week period using two frequency counting observation instruments, one instrument to record the type of texts and

one to record teachers' activities and instruction; however, detailed information about the instruments' validity and (inter-rater) reliability is lacking. The other study was conducted by Van Hover et al. (2012), the only researchers who attempted to construct a subject-specific observation instrument, called the PATH, with the goal of evaluating and improving history instruction. PATH has the same structure as Pianta and Hamre's (2009) *Classroom Assessment Scoring System-Secondary* (CLASS-S) and consists of six dimensions: (1) lesson components, (2) comprehension, (3) narrative, (4) interpretation, (5) sources, and (6) historical practices. Each dimension includes indicators and behavioral markers that are scored "high," "middle," and "low" by observers. The authors tested the inter-rater reliability for PATH and found positive indicators, but detailed information about the instrument's validity and reliability is lacking.

Historical contextualization: a conceptualization

Rather than constructing an observation instrument for all historical reasoning competencies, we focus on how history teachers promote historical contextualization in classrooms. This focus provides us with the opportunity to spend sufficient time on item development and to test whether it is possible to observe history teachers' subject-specific strategies using an observation instrument. We chose historical contextualization because it is considered a key competency of historical reasoning (Davies 2010; Lévesque 2008; Seixas and Morton 2013) and is, therefore, included in the formal history curricula of many countries, such as Australia, Belgium, Canada, Finland, Germany, the Netherlands, Spain, and the UK (Huijgen et al. 2014).

In history education, it is possible to contextualize historical sources and phenomena, including persons, events, and developments (Havekes et al. 2012). Historical contextualization is the ability to situate a historical phenomenon or person in a temporal, spatial, and social context to describe, explain, compare, or evaluate it (Van Boxtel and Van Drie 2012). Wineburg and Fournier (1994) defined historical contextualization as building a context of circumstances or facts that surround a particular historical phenomenon to render it more intelligible. Endacott and Brooks (2013) viewed historical contextualization as "a temporal sense of difference that includes deep understanding of the social, political, and cultural norms of the time period under investigation as well as knowledge of the events leading up to the historical situation and other relevant events that are happening concurrently" (p. 43). Historical events and historical agents' decisions must be placed in the specific socio-spatial and socio-temporal locations in which they emerged. For example, students must know that in ancient Roman times, Julius Caesar could not have had breakfast in Rome and dinner in the Gaul region of France on the same day because the transportation modes needed for such a trip was not available (Lévesque 2008).

Teachers' strategies for promoting historical contextualization

Research has been conducted to conceptualize the instructional practices that effective teachers employ to promote historical contextualization in classrooms (e.g., Doppen 2000; Rantala 2011; Van Boxtel and Van Drie 2012). To teach historical reasoning competencies such as historical contextualization, teachers must not only possess expert levels of subject content knowledge but also activate students to acquire knowledge and help them apply this knowledge to gain different historical reasoning

competencies (Haydn et al. 2015). Additionally, Hattie's meta-analysis (2008) indicated that effective teachers activate student learning. Other meta-analyses on effective teaching seem to confirm this finding (e.g., Kyriakides et al. 2013; Seidel and Shavelson 2007). Exposure to information alone is not sufficient for students to gain deep subject-specific understanding and historical reasoning competencies. Based on research that focused on historical contextualization, we identified four main teaching strategies for promoting historical contextualization in classrooms: (1) reconstructing the historical context, (2) fostering historical empathy, (3) performing historical contextualization to explain the past, and (4) raising awareness of present-oriented perspectives when examining the past.

First, the historical context of a phenomenon must be reconstructed to perform historical contextualization. Foster (1999) argued that students must possess historical context knowledge, including knowledge about chronology, before they can perform historical contextualization. Reisman and Wineburg (2008) also stressed the importance of background knowledge for the performance of historical contextualization. To reconstruct the historical context, students and teachers can use different frames of reference such as the chronological frame of reference, spatial frame of reference, or social frame of reference (e.g., De Keyser and Vandepitte 1998; Pontecorvo and Girardet 1993; Van Boxtel and Van Drie 2012). The chronological frame includes knowledge of time and period, significant events, and developments (Dawson 2009; Wilschut 2012). The spatial frame focuses on knowledge of (geographical) locations and scale (Havekes et al. 2012). The social frame includes not only knowledge of human behavior and the social conditions of life but also knowledge of socio-economic, socio-cultural, and socio-political developments (Van Boxtel and Van Drie 2004).

To reconstruct the historical context, teachers and students should explore the different frames of reference. For example, in previous research, we found that most students who used and combined different types of knowledge (e.g., chronological, spatial, economic, political, and cultural knowledge) obtained higher scores on a historical contextualization task than students who used a single type of knowledge (Huijgen et al. in press). Teachers could use different sources to build the different frames of knowledge, such as movies (Marcus 2005; Metzger 2012), written documents, objects, and images (Fasulo et al. 1998; Van Drie and Van Boxtel 2008).

Second, although some scholars claim that historical empathy is idealistic and can never be fully achieved because most historical agents are dead (e.g., Kitson et al. 2011; Riley 1998; Wineburg 2001), most scholars agree that historical empathy could promote historical contextualization (e.g., Cuningham 2012; Davis 2001; Endacott and Brooks 2013; Lee and Ashby 2001; Skolnick et al. 2004). Historical empathy focuses on empathizing with people in the past based on historical knowledge that explains their actions. Colby (2008) noted that the primary purpose of historical empathy is to enable students to transcend the boundaries of presentism by developing a rich understanding of the past from multiple viewpoints. In history lessons, teachers could focus on a historical agent to gain insight into the views and values of people who lived in the past (e.g., Foster 1999; Wooden 2008) or discuss historical agents' decisions with a group of students (Kohlmeier 2006). Teachers could also promote historical empathy by promoting the formation of affective connections with the historical agent based on students' own similar yet different life experiences (Endacott and Pelekanos 2015; Kitson et al. 2011) and focusing on understanding historical agents' prior knowledge and positions (Berti et al. 2009; Hartmann and Hasselhorn 2008; Huijgen et al. 2014).

Third, students should be able to explain the past based on their historical context knowledge (Lévesque 2008; Seixas and Morton 2013; Wineburg 2001). For example, students must explain why the Great Depression of 1929 spread to Europe or the differences between governance in ancient Greece and governance in the Middle Ages. To answer such historical questions, students must link the Great Depression and the different types of governance to their historical context (Seixas 2006). Furthermore, the successful performance of different historical reasoning competencies, such as identifying indirect and direct causes (Stoel et al. 2015), understanding change and continuity (Haydn et al. 2015), reasoning with historical sources (Reisman and Wineburg 2008), and asking historical questions (Logtenberg et al. 2011), requires an analysis of the broader historical context. Teachers should, therefore, create opportunities for students to practice these competencies with these types of questions. Hallden (1997) suggested that teachers should focus their instruction on the relationship between historical factual details (lower-level context) and large historical developments (larger context). Kosso (2009) also noted that "Individual events and actions are understood by being situated in the larger context. However, the larger context is understood by being built of individual events. It is a hermeneutic circle and perhaps the only way to understand other people" (p. 24). Presenting and evaluating historical phenomena from different perspectives is also considered an effective approach (e.g., Ciardiello 2012; Levstik 1997; McCully 2012; Stradling 2003). For example, to understand and explain the Cuban Missile Crisis of 1962, students should examine this phenomenon from not only a capitalist Western perspective but also a communist Soviet perspective.

Finally, teachers should raise awareness of students' present-oriented perspective and the consequences of this perspective when examining the past (Barton and Levstik 2004; Huijgen et al. 2014; Wineburg 2001). Students must know that the past differs from the present (Seixas and Peck 2004); however, social psychology research illustrates that young students especially find it very difficult to take another persons' perspective, particularly when that other person does not have the same knowledge that the students have (Bloom and German 2000; Wellman et al. 2001). This inability could cause problems in history education, as students must be aware that much of the information that they know was not available to people in the past. Students' present-oriented thinking or *presentism* is considered one of the main reasons why they fail to achieve historical contextualization and could cause misconceptions among students, leading them to reach incorrect conclusions about historical phenomena (Lee and Ashby 2001; Huijgen et al. 2014; VanSledright and Afflerbach 2000).

Although we can never be perfectly non-presentist (e.g., Pendry and Husbands 2000; Wineburg 2001), teachers should foster students' awareness of their own contemporary values and beliefs and the consequences of this perspective when explaining the past. To achieve this goal, teachers could present the past as tension for students (e.g., Savenije et al. 2014; Seixas and Morton 2013), present conflicting historical sources (Ashby 2004), not present the past as progress (Wilschut 2012), and promote intellectual conflict regarding historical phenomena that might be difficult for students to understand and explain (Foster 2001; Huijgen and Holthuis 2015). Furthermore, to prevent students from viewing the past from a present-oriented perspective, teachers should explicitly model or scaffold how historical contextualization can be performed successfully, for example, by providing learning strategies. Explicit teaching of domain-specific strategies, such as how to perform historical contextualization, could promote students' ability to explain historical events (Stoel et al. 2015). Reisman and Wineburg (2008) stressed the importance of explicitly providing students with an illustration

of contextualized thinking, for example, by providing videos of good examples of professional historians who scaffold their contextualization processes.

Research questions

A subject-specific observation instrument could provide insight into the instructions and methods that history teachers employ to promote students' ability to perform historical contextualization. Therefore, we aimed to construct a reliable subject-specific observation instrument and scoring design that measures how history teachers promote historical contextualization in classrooms. To address this central aim, we specify the following three research questions:

1. What is the observation instruments' dimensionality when used to observe how history teachers promote historical contextualization?
2. What are the reliability outcomes when the observation instrument is used to observe how history teachers promote historical contextualization?
3. How many lessons and observers are necessary to establish a reliable and optimal scoring design?

Method

Structure of the observation instrument

To design and construct our observation instrument, we used the guidelines described by Colton and Covert (2007), which focus on the development of valid and reliable instruments in social sciences. Our instrument could be characterized as a high-inference observation instrument. In contrast to low-inference instruments (such as time sampling and time logs), high-inference instruments provide a more qualitative verdict (Chávez 1984). However, these instruments are more susceptible to subjectivity; therefore, thorough inter-rater reliability procedures are necessary.

We modeled our instrument on Van de Grift's (2007, 2009) *International Comparative Analysis of Learning and Teaching* (ICALT) observation instrument. We chose this instrument's format because it also seeks to observe teachers' professional strategies and calculate scores based on these strategies. Similar to the ICALT instrument, our instrument utilizes a four-point Likert scale to score the items. In our instrument, scores 1 and 2 represent a negative verdict, while scores 3 and 4 represent a positive verdict. Score 1 should be used only if teachers do not use a particular strategy in their lessons.

Formulating and refining the items

Based on the four main strategies identified in our theoretical framework (reconstructing the historical context, fostering historical empathy, performing historical contextualization to explain the past, and raising awareness of a present-oriented perspective) and a review of literature on teaching historical contextualization, we formulated observable items to assess classroom teachers' behavior in regards to historical contextualization.

Furthermore, during two national teacher professionalization conferences, we asked 25 history teachers (after an introduction of the concept of historical contextualization) to each formulate 20 items that assess classroom teachers' behavior in regards to historical contextualization. Combining these items with the items that we formulated resulted in a total of 121 items.

Meta-analyses on effective teaching illustrate that promoting different types of interactions in classrooms (i.e., student-student interactions and teacher-student interactions) could promote student learning (Kyriakides et al. 2013; Seidel and Shavelson 2007). Therefore, we formulated three items ("the teacher asks evaluative questions," "the teacher uses classroom discussion," and "the teacher uses group work"), focusing on more generic teacher strategies and different (social) interactions in the classrooms. We included these three generic items because history education research shows that these types of interaction could promote historical reasoning competencies (e.g., Brooks 2008; Van Drie et al. 2006; Van Drie and Van Boxtel 2008; Stoel et al. 2015). Therefore, the total list included 124 items.

By excluding double items and items that might be very difficult to evaluate, we shortened the list to 82 items. For example, we first included individual items for all time indicators (e.g., year, period, and century), but we then incorporated these items into one item, "the teacher gives time indicators." Another example is that we excluded items focusing on the specific economic, political, and social circumstances (e.g., form of government, welfare, scientific knowledge, wars, and laws) of historical phenomena. Because these specific circumstances are difficult to observe in one history lesson, we included only items such as "appoints political/governance characteristics at the time of phenomena" and "appoints social-cultural characteristics at the time of phenomena." This method might result in a less nuanced image of a lesson, but we preferred to develop an instrument that allows us to observe all behavior indicators in a single history lesson.

Next, we organized two expert panel discussions to further shorten the list of 82 items and ensure the instrument's face and content validity. The first panel discussion was held with two history teacher educators and seven secondary-school history teachers. The second panel discussion was held with one history teacher educator and four secondary-school history teachers. All experts had more than 7 years of work experience. The experts were asked to (1) remove unnecessary items that did not measure history teachers' competency in terms of promoting historical contextualization, (2) remove possible multiple items that might cover the same teacher behavior, (3) reformulate unclear items, and (4) formulate new items that they thought were missing. In total, the experts excluded 24 items, reformulated 12 items, and created no new items, resulting in a list of 58 items.

Subsequently, we trained ten student history teachers on the use of the observation instrument, and they observed one videotaped history lesson using the instrument. We calculated Cronbach's alpha (jury alpha) for their observation scores to explore the instrument's internal consistency. This jury alpha was 0.58 (poor internal consistency). After deleting ten items that threatened internal consistency, the jury alpha increased to 0.81 (good internal consistency). Examples of the deleted items are "appoints relations between historical phenomena," "uses substantive concepts when explaining historical phenomena," and "uses general schemas to explain historical phenomena." We asked the experts in the first panel session to determine whether the ten deleted items could jeopardize the instrument's face and content validity; they found no threats.

The same experts were also asked to observe three videotaped history lessons taught by three different history teachers using the 48 items. After discussing each lesson, three items ("explains the importance of placing phenomena in a chronological framework," "explains the importance of placing phenomena in a spatial framework," and "explains the importance of viewing phenomena from different dimensions") led to strong disagreement among the experts; thus, we deleted these items. This resulted in a total list of 45 items in the first version of the *Framework for Analyzing the Teaching of Historical Contextualization* (FAT-HC).

Research design

Following Hill et al. (2012), we adopted generalizability theory to explore the instrument's dimensionality and to determine its reliability (Brennan 2001; Cronbach et al. 1972; Shavelson and Webb 1991). Compared to the classical test theory, generalizability theory is more informative and useful in educational systems because the classical test theory considers only one source of measurement error at a time. Additionally, it does not result in specific information on how many forms, items, occasions, or observers are required (Shavelson et al. 1989). A generalizability study (G-study) can accommodate any observational situation and is restricted by only the practical limitations of data collection and software (Lei et al. 2007). A G-study views a behavioral measurement (for example, an observed score) as a sample from a universe of admissible observations. Each aspect (called a facet) in the measurement procedure is considered a possible source of error. A G-study provides estimates of the variance contributed by persons, observers, occasions of measurement, and each of the possible interactions between these facets. Generalizability theory distinguishes a decision study (D-study) from a G-study. A D-study uses information from a G-study to construct a scoring design that minimizes error for a particular purpose (Shavelson and Webb 1991). In addition to a G-study, a D-study can identify the optimal data collection for a desired score reliability (Hill et al. 2012).

Sample and data collection

Non-probability sampling was used to select five teachers to observe and five observers (see Tables 1 and 2 for the teachers' and observers' characteristics). In the Netherlands, the average

Table 1 Teachers' characteristics

Teacher	Gender	Age	Educational qualification	Years' work experience	Nationality	Historical expertise	Students' performance[a]
A	Male	60	Bachelors	37	Dutch	Modern history	=
B	Male	34	Masters	8	Dutch	Modern history	>
C	Male	43	Masters	17	Dutch	Early modern history	>
D	Male	63	Masters	41	Dutch	Middle Ages	>
E	Male	41	Masters	14	Dutch	Early modern history	>

[a] Students' mean score on the formal history exam compared to the national mean score on the formal history exam

Table 2 Observers' characteristics

Observer	Gender	Age	Educational qualification	Years' work experience	Nationality
1	Female	29	Masters	7	Dutch
2	Female	32	Masters	7	Dutch
3	Male	32	Masters	8	Dutch
4	Male	33	Masters	7	Dutch
5	Male	29	Masters	8	Dutch

age of male teachers is 46 years and that of female teachers is 42 years. The gender distribution of the teachers was 48 % female and 52 % male. In total, there are 1785 history teachers with a master's degree and 3944 history teachers with a bachelor's degree working in the Netherlands (Dutch Ministry of Education 2011). The teachers in the sample worked at different schools, and these schools did not differ significantly from the total population in regards to student enrollment, location (rural or urban), or graduation rate (Statistics Netherlands 2014). The national students' mean score on the formal history exam for general secondary education and pre-university education was 6.35 on a ten-point scale.

We videotaped two different lessons for each teacher ($n = 5$), and all lessons were taught in the two highest tracks of secondary education in the Dutch educational system. We observed only the lessons for upper secondary-school students in the two highest tracks because the Dutch formal exam program considers the ability to perform historical contextualization to be an important aim for these students (Dutch Ministry of Education 2015). A total of 267 students, with a mean age of 16.2 (SD = 0.7) years old, were involved. The mean duration of analyzed lessons was 39 min (SD = 2.4). Each observer individually evaluated the ten videotaped lessons using the developed observation instrument, yielding a total of 50 observations.

Training observers to use the instrument

All observers received a 4-h training. In this training, we used three videotaped history lessons taught by three history teachers (one female teacher with more than 15 years of work experience, one male teacher with 4 years of work experience, and one male teacher with more than 25 years of work experience) from three different schools as training materials. One lesson was about the Ancient Roman period, one was about the Middle Ages, and one was about the Second World War. These three lessons were not used in our data analyses. The observers received an explanation of the 45 items and evaluated the videotaped lessons using a training version of the observation instrument that included more in-depth explanations of the items. After the observers observed each videotaped lesson, their results were discussed, and some items were clarified by the trainers to minimize inter-rater bias.

Data analysis

To explore the instrument's dimensionality, we conducted a G-study at the item level with seven facets in a crossed design. To estimate the reliability of our instrument and

produce a composite of scores with maximum generalizability, we conducted a new G-study and employed multivariate generalizability using a "$t \times l \times o$" design, where t represents the observed history teachers, l represents the number of observed lessons, and o represents the number of observers. To determine the optimal number of observers and lessons needed in a scoring design to achieve acceptable reliability, we conducted a D-study using the information from the earlier conducted G-study that estimated the reliability of our instrument.

Results

The instrument's dimensionality

Based on our theoretical framework, we consider our instrument to be one-dimensional because all items should measure teachers' ability to promote historical contextualization. The first data analysis indicated that five items ("the teacher asks evaluative questions," "the teacher uses classroom discussion," "the teacher uses group work," "the teacher compares phenomena with the present," and "the students compare phenomena with the present") displayed a low correlation (<0.30) with the other items. These five items also obtained a standard deviation above 1.00 and were excluded from further data analysis, resulting in a total list of 40 items in the final version of the FAT-HC observation instrument (see Appendix A).

To further explore the instrument's dimensionality, we conducted a G-study at the item level with seven facets in a crossed design using the collected data of the five observers who each evaluated two lessons taught by five teachers (50 observations in total). If our instrument is, in fact, one-dimensional, the *item* facet should explain the main part of the overall variance and the other facets (including the interaction effects) should explain a lesser part of the variance (e.g., Brennan 2001; Shavelson and Webb 1991). As shown in Table 3, the item facet was responsible for most of the variance (47.25 %), indicating that our instrument is one-dimensional in regards to observing how history teachers promote historical contextualization in classrooms.

Table 3 Variance decomposition for the item level

Variance components	Estimated variance	Percentage of variance
Item	0.43	47.25
Item * teacher	0.10	10.99
Item * observer	0.03	3.30
Item * lesson	0.00	0.00
Item * teacher * observer	0.02	2.20
Item * teacher * lesson	0.08	8.79
Item * lesson * observer	0.00	0.00
Residual	0.25	27.47
Total	0.91	100.00

*Interaction effect

The instrument's reliability

To determine the reliability of our instrument, a new G-study was conducted using the same data set (50 observations). The analysis was conducted on the final version of our observation instrument, which consisted of 40 items (see Appendix A). Table 4 displays the results of this G-study and presents the variance decomposition to assess the instrument's reliability. A reliable instrument should have a high proportion of the variance explained by differences between the observed teachers and a low proportion of the variance explained by lessons and observers.

The difference between the observed teachers accounted for 59.12 % of the variance, the difference between the observers accounted for 4.58 % of the variance, and the difference between the lessons accounted for 1.63 % of the variance. The residual was 34.67 %. The results show that the influence of the observers and lessons was very low, indicating that the observers and lessons can be considered to be inter-changeable and that the observers understood the observation items. Interaction effects between the different facets (observers * lessons, observers * teachers, and teacher * lessons) were also calculated and did not display any variance, indicating small differences between the observers' observations of the different teachers and lessons.

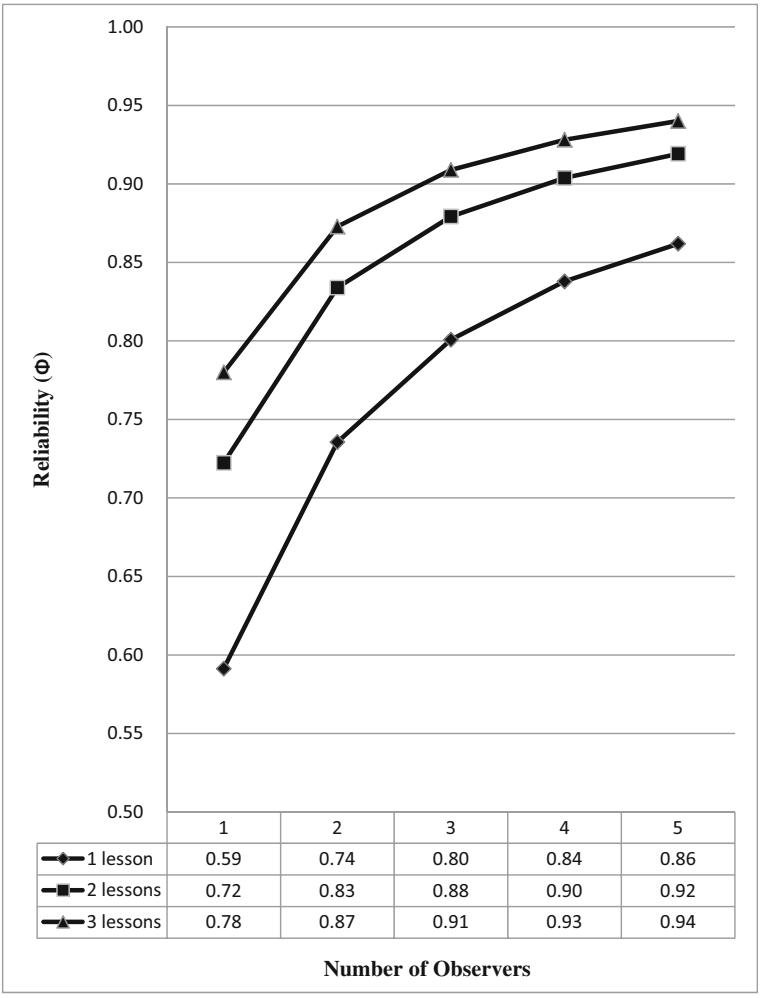
The optimal reliable scoring design

To identify the optimal number of observers and lessons needed for a reliable scoring design, we conducted a D-study based on the results of our G-study, which estimated the instrument's reliability. Because we are interested in the absolute level of an individual's performance independent of others' performance, we calculated the index of dependability coefficient (Φ) to identify the optimal number of observers (Shavelson and Webb 1991). The Φ should be ≥0.7 research purposes, ≥0.8 for formative evaluations, and ≥0.9 for summative evaluations (Brennan and Kane 1977).

The results of our D-study can be found in Fig. 1. A scoring design with one observer evaluating one lesson taught by a teacher yields a Φ of 0.59 (poor reliability), and this value increases to $\Phi = 0.72$ when one observer evaluates two lessons taught by the same teacher. Because we are interested in research purposes and formative evaluations, the optimal scoring design would use two observers who each evaluate two different lessons taught by the same teacher ($\Phi = 0.83$) or three observers who each evaluate the same lesson taught by a teacher ($\Phi = 0.80$).

Table 4 Variance decomposition for the observation instrument

Variance components	Estimated variance	Percentage of variance
Teachers	30.05	59.12
Observers	2.33	4.58
Lessons	0.83	1.63
Residual	17.62	34.67
Total	50.83	100.00



	1	2	3	4	5
1 lesson	0.59	0.74	0.80	0.84	0.86
2 lessons	0.72	0.83	0.88	0.90	0.92
3 lessons	0.78	0.87	0.91	0.93	0.94

Fig. 1 Results of the D-study

Conclusion and discussion

The aim of the present study was to develop a reliable observation instrument and scoring design to assess how history teachers promote historical contextualization in classrooms. This study resulted in the FAT-HC observation instrument. Using expert panels, we found positive indicators of the instrument's content validity. Furthermore, generalizability theory analysis provides indicators that the instrument is one-dimensional when used to evaluate how history teachers promote historical contextualization. Generalizability theory analysis also showed that a large proportion of the instrument's variance was explained by the differences between the observed teachers and a small proportion of the variance was explained by the differences in lessons and observers, which demonstrates the instrument's reliability (Brennan 2001; Hill et al. 2012; Shavelson and Webb 1991). Our D-study showed a reliable scoring design, with one observer evaluating two lessons as the most

effective method for research purposes. For formative teacher evaluations, a reliable scoring design in which two observers each evaluate two lessons or three observers each evaluate one lesson is most effective.

Van Hover et al. (2012, p. 604) noted that instruments that provide "useful discipline-specific feedback to preservice (and inservice) history teachers" are lacking. Additionally, Darling-Hammond et al. (2012) emphasized that most current teacher evaluation programs do little to help teachers improve their teaching. The FAT-HC instrument could provide insight into teachers' subject-specific needs, resulting in a valuable addition to existing generic observation instruments (Grossman and McDonald 2008). For example, if a teacher obtains low scores on the instrument, attention could be devoted to specific items of the instrument in teacher education or professional development programs. The pre-observation and post-observation interviews also could be structured based on the instrument's items, resulting in more concrete feedback for the observed teacher.

The instrument could also help researchers examine the instructions and methods that teachers employ to promote historical contextualization in classrooms. In the history education literature, there is a clear view of high-quality teaching and learning of history; however, research instruments that capture this view when observing history teachers while they work do not exist (Van Hover et al. 2012). Furthermore, our instrument could be used to gain more insight into the association between history teachers' instructions and methods and student achievement. Do teachers who activate their students to reconstruct a historical context better promote students' historical understanding than teachers who do not? The instrument could also be used to evaluate intervention studies, for example, to examine the effects of training teachers in the use of instructions incorporated into the observation instrument.

In addition to the function of the research instrument and feedback instrument, the instrument could be used as a framework for teachers who want to reshape and improve their instruction on historical contextualization. Slavin (1996) noted that teachers who explicitly model and scaffold their instructions contribute to their students' academic success. The instrument's strategies and items could provide direction for designing meaningful learning tasks and scaffolds for students. This is important, especially because, as noted by Grant and Gradwell (2010), many history teachers focus on recalling factual knowledge despite the fact that the teaching and learning of history includes far more activities, such as investigating sources and evaluating the past (Van Sledright 2008). Bain and Mirel (2006) and SSIRC (2013), therefore, argued that instruction models that help teachers learn how to promote students' ability to perform historical contextualization or other historical reasoning competencies are needed. In a post-observation interview, one of our observed teachers noted that he now uses the instrument as a checklist when designing his lessons. Prior to the study, he would forget the spatial context of historical phenomena. However, he now structurally includes the geographical context in his lessons when reconstructing the historical context of phenomena.

Despite the positive indicators of the instrument's reliability, some limitations must be acknowledged. We used a research design with only five observers and five teachers, who participated voluntary and, thus, might be more eager to learn (Desimone 2009; Desimone et al. 2006). More observers, teachers, and lessons (cf. Hill et al. 2012) are needed to provide greater insight into the instrument's dimensionality, reliability, and optimal scoring design. Including teachers and observers with more varied backgrounds (e.g., differences in gender, student performance, age, and

educational qualification) might also provide useful insights to further strengthen the instrument and scoring design.

Furthermore, when examining the instrument's reliability, nearly 35 % of the variance (residual) could not be explained by teacher, observer, or lesson variance. Future research and analyses must be conducted to decrease the residual variance and achieve greater reliability. The observers also noted that it is difficult to evaluate 40 items when observing one history lesson. Because the observation instrument must be practical and suitable for observing a single lesson, more research is needed to decrease the number of items while maintaining good reliability. A larger G-study including a D-study, which focuses on how many items are necessary to achieve reliability, could provide these insights (Brennan 2001). We also used videotaped lessons. Although videotaped lessons have many benefits and are widely used for constructing and validating observation instruments (e.g., Yoder and Symons 2010), they differ from "live" classroom observations. Future research should include live observations to assess possible differences in the instrument's reliability for live vs. videotaped sessions. Live video classroom observations (e.g., Liang 2015) could also be an interesting method to examine possible differences in reliability.

To further assess the instrument's construct validity, intervention studies with a quasi-experimental design and pre- and post-tests to further test the framework's efficiency for promoting historical contextualization are needed. The use of other methods to assess teacher factors, such as student questionnaires and teachers' self-reports on historical contextualization, could also provide important insights into the instrument's construct validity (e.g., Kyriakides 2008; Muijs 2006). Additionally, Rasch modeling could provide information on the instrument's reliability, which items history teachers find more difficult to perform and which items they consider easier to perform (e.g., Fischer and Molenaar 1995; Maulana et al. 2014; Van de Grift et al. 2014).

In conclusion, Ball and Forzani (2009) noted that current teacher education programs are often centered on teachers' beliefs and knowledge and argued that teacher education programs should mainly focus on the task and activities of teaching. They concluded that far more research is needed to gain insight into the tasks and activities of teaching across different subjects. We hope that our instrument can contribute to further insights into teachers' subject-specific activities for the teaching and learning of historical contextualization. Our instrument is not designed to assess history teachers; rather, it should function as a tool used to improve history instruction. Marriott (2001) noted that "Teachers seldom have a clear idea about their strengths and weaknesses. This is often because they have not been systematically observed and constructively debriefed" (p. 6). History teachers could observe each other using the instrument, discuss their lessons and findings, and collaboratively design new lessons with the instrument as framework, which might result in a giant step forward in the teaching and learning of history.

Acknowledgments The authors wish to thank the reviewers for their valuable comments and Rikkert van der Lans for his research assistance.

Compliance with ethical standards

Funding This work was funded by The Netherlands Organization for Scientific Research (NWO) under Grant Number 023.001.104.

Appendix A Framework for Analyzing the Teaching of Historical Contextualization (FAT-HC)
Explanatory notes: *1* weak, *2* more weak than strong, *3* more strong than weak, and *4* strong

	The teacher...	***1***	***2***	***3***	***4***
1	Activates relevant prior knowledge	0	0	0	0
2	Shows visual material	0	0	0	0
3	Uses historical sources	0	0	0	0
4	Gives time indicators regarding phenomena (e.g., year / century / period)	0	0	0	0
5	Gives the duration of phenomena	0	0	0	0
6	Shows phenomena on a timeline	0	0	0	0
7	Gives geographical/spatial indicators regarding phenomena	0	0	0	0
8	Shows phenomena on a geographical map	0	0	0	0
9	Appoints political/governance characteristics at the time of phenomena	0	0	0	0
10	Appoints economic characteristics at the time of phenomena	0	0	0	0
11	Appoints social-cultural characteristics at the time of phenomena	0	0	0	0
12	Appoints causes and consequences of phenomena	0	0	0	0
13	Appoints change and continuity regarding phenomena	0	0	0	0
	The students...	***1***	***2***	***3***	***4***
14	Give time indicators regarding phenomena	0	0	0	0
15	Give the duration of phenomena	0	0	0	0
16	Give geographical/spatial indicators regarding phenomena	0	0	0	0
17	Appoint political/governance characteristics at the time of phenomena	0	0	0	0
18	Appoint economic characteristics at the time of phenomena	0	0	0	0
19	Appoint social-cultural characteristics at the time of phenomena	0	0	0	0
20	Appoint causes and consequences of phenomena	0	0	0	0
21	Appoint change and continuity regarding phenomena	0	0	0	0
	The teacher...	***1***	***2***	***3***	***4***
22	Centralizes a historical actor	0	0	0	0
23	Moves self into the past to explain phenomena (if I ..)	0	0	0	0
24	Outlines a recognizable role for students to foster empathy (as a businessman / like a father)	0	0	0	0
	The students..	***1***	***2***	***3***	***4***
25	Make affective / emotional connections with historical actors	0	0	0	0
26	Consider the role of the historical actor to explain historical decisions	0	0	0	0
27	State what they would have decided regarding historical decisions	0	0	0	0

	The teacher...	***1***	***2***	***3***	***4***
28	Compares phenomena with other times	0	0	0	0
29	Compares phenomena with other places	0	0	0	0
30	Places phenomena in long-term developments	0	0	0	0
31	Outlines phenomena from different perspectives	0	0	0	0
	The students...	***1***	***2***	***3***	***4***
32	Compare phenomena with other times	0	0	0	0
33	Compare phenomena with other places	0	0	0	0
34	Place phenomena in long-term developments	0	0	0	0
35	Outline phenomena from different perspectives	0	0	0	0
	The teacher...	***1***	***2***	***3***	***4***
36	Does *not* use anachronisms	0	0	0	0
37	Does *not* present the past as progress	0	0	0	0
38	Creates historical tension (the past as different)	0	0	0	0
39	Presents conflicting historical sources	0	0	0	0
40	Presents learning strategies for historical contextualization	0	0	0	0

References

Achinstein, B., & Fogo, B. (2015). Mentoring novices' teaching of historical reasoning: opportunities for pedagogical content knowledge development through mentor-facilitated practice. *Teaching and Teacher Education, 45*(2), 45–58.

Adler, S. (2008). The education of social studies teachers. In L. S. Levstik & C. A. Tyson (Eds.), *Handbook of research in social studies education* (pp. 329–351). New York, NY: Routledge.

Ashby, R. (2004). Developing a concept of historical evidence: students' ideas about testing singular factual claims. *International Journal of Historical Learning, Teaching and Research, 4*(2), 44–55.

Bain, R. B., & Mirel, J. (2006). Setting up camp at the great instructional divide. Educating beginning history teachers. *Journal of Teacher Education, 57*(3), 212–219.

Ball, D. L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education, 60*(5), 497–511.

Barton, K. (2012). Agency, choice and historical action: how history teaching can help students think about democratic decision making. *Citizenship Teaching & Learning, 7*(2), 131–142.

Barton, K., & Levstik, L. (2004). *Teaching history for the common good*. Mahwah, NJ: Lawrence Erlbaum Associates.

Berti, A. E., Baldin, I., & Toneatti, L. (2009). Empathy in history. Understanding a past institution (ordeal) in children and young adults when description and rationale are provided. *Contemporary Educational Psychology, 34*(4), 278–288.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition, 77*(1), 25–31.
Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: brain, mind, experience, and school*. Washington, DC: National Academy Press.
Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.
Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*(3), 277–289.
Brooks, S. (2008). Displaying historical empathy: what impact can a writing assignment have. *Social Studies Research and Practice, 3*(2), 130–146.
Chávez, R. C. (1984). The use of high-inference measures to study classroom climates: a review. *Review of Educational Research, 54*(2), 237–261.
Ciardiello, A. V. (2012). Is Angel Island the Ellis Island of the West? Teaching multiple perspective-taking in American immigration history. *The Social Studies, 103*(4), 171–176.
Colby, S. (2008). Energizing the history classroom: historical narrative inquiry and historical empathy. *Social Studies Research and Practice, 3*(3), 60–79.
Colton, D., & Covert, R. W. (2007). *Designing and constructing instruments for social research and evaluation*. San Francisco, CA: John Wiley & Sons.
Cronbach, L., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York, NY: Wiley & Sons.
Cuningham, D. L. (2012). Understanding pedagogical reasoning in history teaching through the case of cultivating historical empathy. *Theory & Research in Social Education, 35*(4), 592–630.
Danielson, C. (1996). *Enhancing professional practice: a framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.
Darling-Hammond, L. (2012). *Creating a comprehensive system for evaluating and supporting effective teaching*. Stanford, CA: Stanford Center for Opportunity Policy in Education.
Darling-Hammond, L., Wei, R., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession*. Washington, DC: National Staff Development Council.
Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8–15.
Davies, I. (Ed.). (2010). *Debates in history teaching*. London: Routledge.
Davis, O. L. (2001). In pursuit of historical empathy. In O. L. Davis, E. A. Yeager, & S. J. Foster (Eds.), *Historical empathy and perspective taking in the social studies* (pp. 1–12). New York, NY: Rowman and Littlefield.
Dawson, I. (2009). What time does that tune start? From thinking about "sense of period" to modelling history at key stage 3. *Teaching History, 135*, 50–57.
De Graaff, R., Jan Koopman, G., Anikina, Y., & Westhoff, G. (2007). An observation tool for effective L2 pedagogy in Content and Language Integrated Learning (CLIL). *International Journal of Bilingual Education and Bilingualism, 10*(5), 603–624.
De Keyser, R., & Vandepitte, P. (1998). *Historical formation. Design of vision*. Brussel, Belgium: Flemish Board for Catholic Secondary Education.
Desimone, L. M. (2009). Improving impact studies of teachers' professional development: toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199.
Desimone, L. M., Smith, T. M., & Ueno, K. (2006). Are teachers who need sustained, content-focused professional development getting it? An administrator's dilemma. *Educational Administration Quarterly, 42*(2), 179–215.
Doppen, F. H. (2000). Teaching and learning multiple perspectives: the atomic bomb. *The Social Studies, 91*(4), 159–169.
Dutch Ministry of Education. (2011). Data file secondary school teachers. Retrieved from the website: www.duo.nl/.
Dutch Ministry of Education. (2015). History exam program. Retrieved from the website: www.examenblad.nl/.
Endacott, J. L., & Brooks, S. (2013). An updated theoretical and practical model for promoting historical empathy. *Social Studies Research & Practice*, *8*(1), 41-58.
Endacott, J. L., & Pelekanos, C. (2015). Slaves, women, and war! Engaging middle school students in historical empathy for enduring understanding. *The Social Studies, 106*(1), 1–7.
Erdmann, E., & Hasberg, W. (Eds.). (2011). *Facing–mapping–bridging diversity* (Foundation of a European discourse on history education. Part 1 & 2). Schwalbach, Germany: Wochenschau Verlag.
Fasulo, A., Girardet, H., & Pontecorvo, C. (1998). Seeing the past: learning history through group discussion and iconographic sources. In J. F. Voss & M. Carretero (Eds.), *Learning and reasoning in history. International review of history education* (Vol. 2, pp. 132–153). London, United Kingdom: Woburn Press.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: foundations, recent developments, and applications*. New York, NY: Springer Science & Business Media.

Fogo, B. (2014). Core practices for teaching history: the results of a Delphi panel survey. *Theory & Research in Social Education, 42*(2), 151–196.

Foster, S. (1999). Using historical empathy to excite students about the study of history: can you empathize with Neville chamberlain? *The Social Studies, 90*(1), 18–24.

Foster, S. (2001). Historical empathy in theory and practice: some final thoughts. In O. L. Davis, E. A. Yeager, & S. J. Foster (Eds.), *Historical empathy and perspective taking in the social studies* (p. 167). Lanham, MD: Rowman and Littlefield.

Gertsen, R., Baker, S. K., Haager, D., & Graves, A. W. (2005). Exploring the role of teacher quality in predicting the reading out comes for first-grade English learners: an observational study. *Remedial and Special Education, 26*(4), 197–206.

Gerwin, D., & Visone, F. (2006). The freedom to teach: contrasting history teaching in elective and state-tested courses. *Theory and Research in Social Education, 34*(2), 259–282.

Grant, S., & Gradwell, J. (2009). The road to ambitious teaching: creating big idea units in history classes. *Journal of Inquiry and Action in Education, 2*(1), 1–26.

Grant, S., & Gradwell, J. (2010). *Teaching with big ideas: cases of ambitious teaching*. Lanham: R&L Education.

Grossman, P., & McDonald, M. (2008). Back to the future: directions for research in teaching and teacher education. *American Educational Research Journal, 45*(1), 184–205.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: the relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores (No. w16015)*. Cambridge, MA: National Bureau of Economic Research.

Hallden, O. (1997). Conceptual change and the learning of history. *International Journal of Educational Research, 27*(3), 201–210.

Hartmann, U., & Hasselhorn, M. (2008). Historical perspective taking: a standardized measure for an aspect of students' historical thinking. *Learning and Individual Differences, 18*(2), 264–270.

Hattie, J. (2008). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

Havekes, H., Coppen, P., Luttenberg, J., & Van Boxtel, C. (2012). Knowing and doing history: a conceptual framework and pedagogy for teaching historical contextualisation. *International Journal of Historical Learning, Teaching and Research, 11*(1), 72–93.

Haydn, T., Stephen, A., Arthur, J., & Hunt, M. (2015). *Learning to teach history in the secondary school: a companion to school experience*. New York, NY: Routledge.

Hill, H. C., Charalambous, C. Y., & Kraft, M. (2012). When rater reliability is not enough: observational systems and a case for the G-study. *Educational Researcher, 41*(2), 56–64.

Huijgen, T., & Holthuis, P. (2015). "Why am I accused of being a heretic?" A pedagogical framework for stimulating historical contextualisation. *Teaching History, 158*, 56–61.

Huijgen, T., Van Boxtel, C., Van de Grift, W., & Holthuis, P. (2014). Testing elementary and secondary school students' ability to perform historical perspective taking: the constructing of valid and reliable measure instruments. *European Journal of Psychology of Education, 29*(4), 653–672.

Huijgen, T., Van Boxtel, C., Van de Grift, Holthuis, P. (in press). Toward historical perspective taking: students' reasoning when contextualizing the actions of people in the past. *Theory and Research in Social Education*.

Kitson, A., Husbands, C., & Steward, S. (2011). *Teaching and learning history 11–18: understanding the past*. Maidenhead: Open University Press.

Kohlmeier, J. (2006). "Couldn't she just leave?": the relationship between consistently using class discussions and the development of historical empathy in a 9th grade world history course. *Theory and Research in Social Education, 34*(1), 34–57.

Kosso, P. (2009). Philosophy of historiography. In A. Tucker (Ed.), *A companion to the philosophy of history and historiography* (pp. 9–25). Malden, MA: Wiley-Blackwell.

Kyriakides, L. (2008). Testing the validity of the comprehensive model of educational effectiveness: a step towards the development of a dynamic model of effectiveness. *School Effectiveness and School Improvement, 19*(4), 429–446.

Kyriakides, L., Campbell, R. J., & Christofidou, E. (2002). Generating criteria for measuring teacher effectiveness through a self-evaluation approach: a complementary way of measuring teacher effectiveness. *School Effectiveness and School Improvement, 13*(3), 291–325.

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: a meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education, 36*(1), 143–152.

Lavigne, A., & Good, T. (2015). *Improving teaching through observation and feedback: beyond state and federal mandates*. New York, NY: Routledge.

Lee, P., & Ashby, R. (2001). Empathy, perspective taking, and rational understanding. In O. L. Davis, E. A. Yeager, & S. J. Foster (Eds.), *Historical empathy and perspective taking in the social studies* (pp. 1–12). New York, NY: Rowman and Littlefield.

Lei, P. W., Smith, M., & Suen, H. K. (2007). The use of generalizability theory to estimate data reliability in single-subject observational research. *Psychology in the Schools, 44*(5), 433–439.

Lévesque, S. (2008). *Thinking historically. Educating students for the twenty-first century.* Toronto, ON: Toronto University Press.

Levstik, L. S. (1997). "Any history is someone's history": listening to multiple voices from the past. *Social Education, 61*(1), 48–51.

Levstik, L. S., & Tyson, C. A. (Eds.). (2008). *Handbook of research on social studies education*. New York, NY: Routledge.

Liang, J. (2015). Live video classroom observation: an effective approach to reducing reactivity in collecting observational information for teacher professional development. *Journal of Education for Teaching, 41*(3), 235–253.

Logtenberg, A., Van Boxtel, C., & Van Hout-Wolters, B. (2011). Stimulating situational interest and student questioning through three types of historical introductory texts. *European Journal of Psychology of Education, 26*(2), 179–198.

Marcus, A. S. (2005). "It is as it was": feature film in the history classroom. *The Social Studies, 96*(2), 61–67.

Marriott, G. (2001). *Observing teachers at work*. Portsmouth, NH: Heinemann.

Matsumura, L., Garnier, H., Slater, S., & Boston, M. (2008). Toward measuring instructional interactions "at-scale". *Educational Assessment, 13*(4), 267–300.

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2014). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: a Rasch modelling approach. *School Effectiveness and School Improvement, 26*(2), 169–194.

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015). Pupils' perceptions of teaching behaviour: evaluation of an instrument and importance for academic motivation in Indonesian secondary education. *International Journal of Educational Research, 69*(1), 98–112.

McCully, A. (2012). History teaching, conflict and the legacy of the past. *Education, Citizenship and Social Justice, 7*(2), 145–159.

Mercer, N. (2010). The analysis of classroom talk: methods and methodologies. *British Journal of Educational Psychology, 80*(1), 1–14.

Metzger, S. (2012). The borders of historical empathy: students encounter the holocaust through film. *Journal of Social Studies Research, 36*(4), 387–410.

Monte-Sano, C. (2011). Learning to open up history for students: preservice teachers' emerging pedagogical content knowledge. *Journal of Teacher Education, 62*(3), 260–272.

Monte-Sano, C., & Cochran, M. (2009). Attention to learners, subject, or teaching: what takes precedence as preservice candidates learn to teach historical thinking and reading? *Theory & Research in Social Education, 37*(1), 101–135.

Muijs, D. (2006). Measuring teacher effectiveness: some methodological reflections. *Educational Research and Evaluation, 12*(1), 53–74.

Nokes, J. D. (2010). Observing literacy practices in history classrooms. *Theory & Research in Social Education, 38*(4), 515–544.

O'Leary, M. (2014). *Classroom observation: a guide to the effective observation of teaching and learning*. London, United Kingdom: Routledge.

Pendry, A., & Husbands, C. (2000). Research and practice in history teacher education. *Cambridge Journal of Education, 30*(3), 321–334.

Pianta, R., & Hamre, B. (2009). Conceptualization, measurement, and improvement of classroom processes: standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119.

Pianta, R., La Paro, K., & Hamre, B. (2008). *Classroom assessment scoring system (CLASS)*. Baltimore, MD: Paul H. Brookes.

Pontecorvo, C., & Girardet, H. (1993). Arguing and reasoning in understanding historical topics. *Cognition and Instruction, 11*(3–4), 365–395.

Rantala, J. (2011). Assessing historical empathy through simulation. How do Finnish teacher students achieve contextual historical empathy? *Nordidactica-Journal of Humanities and Social Science Education, 1*, 58–76.

Reisman, A., & Wineburg, S. (2008). Teaching the skill of contextualizing in history. *The Social Studies, 99*(5), 202–207.
Riley, K. L. (1998). Historical empathy and the Holocaust. *International Journal of Social Education, 13*(1), 32–42.
Ritter, J. K. (2012). Modeling powerful social studies: bridging theory and practice with preservice elementary teachers. *The Social Studies, 103*(3), 117–124.
Savenije, G., Van Boxtel, C., & Grever, M. (2014). Sensitive 'heritage' of slavery in a multicultural classroom: pupils' ideas regarding significance. *British Journal of Educational Studies, 62*(2), 127–148.
Sawada, D., Piburn, M., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: the reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 245–253.
Saye, J., & Brush, T. (2004). Promoting civic competence through problem-based history learning experiments. In G. E. Hamot, J. J. Patrick, & R. S. Leming (Eds.), *Civic learning in teacher education* (Vol. 3, pp. 123–145). Bloomington, IN: The Social Studies Development Center.
Saye, J., & Social Studies Inquiry Research Collaborative (SSIRC). (2013). Authentic pedagogy: its presence in social studies classrooms and relationships to student performance on state-mandated tests. *Theory and Research in Social Education, 41*(1), 89–132.
Schoenfeld, A. (2013). Classroom observations in theory and practice. *ZDM, 45*(4), 607–621.
Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: the role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499.
Seixas, P. (2006). *Benchmarks of historical thinking: a framework for assessment in Canada*. Vancouver, BC: University of British Columbia. Retrieved from the website: www.penseehistorique.ca/sites/default/files/files/docs/Framework_EN.pdf.
Seixas, P., & Morton, T. (2013). *The big six historical thinking concepts*. Toronto, ON: Nelson Education.
Seixas, P., & Peck, C. (2004). Teaching historical thinking. In A. Sears, & I. Wright (Eds.), *Challenges and prospects for Canadian social studies*, (pp. 109-117). Vancouver, BC: Pacific Educational Press
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: a primer*. Newbury Park, CA: Sage.
Shavelson, R., Webb, N., & Rowley, G. (1989). Generalizability theory. *American Psychologist, 44*(6), 922–932.
Skolnick, J., Dulberg, N., & Maestre, T. (2004). *Through other eyes: developing empathy and multicultural perspectives in the social studies*. Toronto, ON: Pippin.
Slavin, R. (1996). *Education for all*. Lisse, the Netherlands: Swets & Zeitlinger Publishers.
Stallings, J., & Kaskowitz, D. (1974). *Follow through classroom observation evaluation, 1972–1973*. Menlo Park, CA: Stanford Research Institute.
Statistics Netherlands. (2014). Data file secondary schools. Retrieved from the website: www.cbs.nl/.
Stearns, L., Morgan, J., Capraro, M., & Capraro, R. (2012). A teacher observation instrument for PBL classroom instruction. *Journal of STEM Education: Innovations and Research, 13*(3), 7–16.
Stoel, G., Van Drie, J., & Van Boxtel, C. (2015). Teaching towards historical expertise. Developing a pedagogy for fostering causal reasoning in history. *Journal of Curriculum Studies, 47*(1), 49–76.
Stradling, R. (2003). *Multiperspectivity in history teaching: a guide for teachers*. Strasbourg, France: Council of Europe.
Teddlie, C., Creemers, B., Kyriakides, L., Muijs, D., & Yu, F. (2006). The international system for teacher observation and feedback: an evolution of an international study of teacher effectiveness constructs. *Educational Research and Evaluation, 12*(6), 561–582.
Van Boxtel, C., & Van Drie, J. (2004). Historical reasoning: a comparison of how experts and novices contextualise historical sources. *International Journal of Historical Learning, Teaching and Research, 4*(2), 89–97.
Van Boxtel, C., & Van Drie, J. (2012). "That's in the time of the Romans!" knowledge and strategies students use to contextualize historical images and documents. *Cognition and Instruction, 30*(2), 113–145.
Van de Grift, W. (2007). Quality of teaching in four European countries: a review of the literature and an application of an assessment instrument. *Educational Research, 49*(2), 127–152.
Van de Grift, W. (2009). Reliability and validity in measuring the value added of schools. *School Effectiveness and School Improvement, 20*(2), 269–285.
Van de Grift, W., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation, 43*(6), 150–159.

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.
Van Drie, J., Van Boxtel, C., & Van der Linden, J. L. (2006). Historical reasoning in a computer-supported collaborative learning environment. In A. M. O'Donnell, C. E. Hmelo, & G. Erkens (Eds.), *Collaborative learning, reasoning and technology* (pp. 265–296). Mahwah, NJ: Lawrence Erlbaum Associates.
Van Hover, S., & Yeager, E. (2004). Challenges facing beginning history teachers: an exploratory study. *International Journal of Social Education, 19*(1), 8–21.
Van Hover, S., Hicks, D., & Cotton, S. (2012). Can you make "historiography" sound more friendly?: towards the construction of a reliable and validated history teaching observation instrument. *History Teacher, 45*(4), 603–12.
VanSledright, B. (2008). Narratives of nation-state, historical knowledge, and school history education. *Review of Research in Education, 32*(1), 109–146.
VanSledright, B. (2010). What does it mean to think historically… and how do you teach it? In W. C. Parker (Ed.), *Social studies today: research and practice* (pp. 112–120). New York, NY: Routledge.
VanSledright, B. (2011). *The challenge of rethinking history education: on practices, theories, and policy*. New York, NY: Routledge.
VanSledright, B., & Afflerbach, P. (2000). Reconstructing Andrew Jackson: prospective elementary teachers' readings of revisionist history texts. *Theory and Research in Social Education, 28*(3), 411–444.
Virta, A. (2002). Becoming a history teacher: observations on the beliefs and growth of student teachers. *Teaching and Teacher Education, 18*(6), 687–698.
Virta, A. (2007). Historical literacy: thinking, reading and understanding history. *Journal of Research in Teacher Education, 14*(4), 11–25.
Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development, 72*(3), 655–684.
Wilschut, A. (2012). *Images of time: the role of an historical consciousness of time in learning history*. Charlotte, NC: Information Age Publishing.
Wineburg, S. (2001). *Historical thinking and other unnatural acts. Charting the future of teaching the past*. Philadelphia, PA: Temple University Press.
Wineburg, S., & Fournier, J. (1994). Contextualized thinking in history. In J. F. Voss & M. Carretero (Eds.), *Cognitive and instructional processes in history and the social sciences* (pp. 285–308). Mahwah, NJ: Lawrence Erlbaum Associates.
Wooden, J. A. (2008). "I had always heard Lincoln was a good person, but…": a study of sixth Graders' reading of Lincoln's views on black–white relations. *The Social Studies, 99*(1), 23–32.
Wragg, T. (1994). *An introduction to classroom observation*. New York: Routledge.
Yoder, P., & Symons, F. (2010). *Observational measurement of behavior*. New York, NY: Springer.

Tim Huijgen, MA. Department of Teacher Education, Faculty of Behavioural and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands. E-mail: t.d.huijgen@rug.nl, website: www.rug.nl/staff/t.d.huijgen.

Current themes of research:

Teaching and learning of history, educational measurement, teacher and teaching quality, and educational design.

Most relevant publications:

Huijgen, T., & Holthuis, P. (2014). Towards bad history? A call for the use of counterfactual reasoning in history education. *Historical Encounters*: *A Journal of Historical Consciousness, Historical Cultures, and History Education*, *1*(1), 103–110.
Huijgen, T., & Holthuis, P. (2015). 'Why am I accused of being a heretic?' A pedagogical framework for stimulating historical contextualisation.*Teaching History*, *158*, 50–55.
Huijgen, T., Van Boxtel, C., Van de Grift, W., & Holthuis, P. (2014).Testing elementary and secondary school students' ability to perform historical perspective taking: The constructing of valid and reliable measure instruments. *The European Journal of Psychology of Education*, *29*(4), 653-672.

Prof. dr. Wim van de Grift. Department of Teacher Education, Faculty of Behavioural and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands. Email: w.j.c.m.van.de.grift@rug.nl, website: www.rug.nl/staff/w.j.c.m.van.de.grift.

Current themes of research:

Professional development of teachers, teacher and teaching quality, and school effectiveness.

Most relevant publications:

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement*, *26*(2), 169–194.

Van de Grift, W. (2007). Quality of teaching in four European countries: a review of the literature and an application of an assessment instrument.*Educational Research*, *49*(2), 127–152.

Van de Grift, W. (2009). Reliability and validity in measuring the value added of schools. *School Effectiveness and School Improvement*, *20*(2), 269-285.

Prof. dr. Carla van Boxtel. Research Institute of Child Development and Education and Amsterdam School of Historical Studies, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS Amsterdam, The Netherlands. E-mail: C.A.M.vanBoxtel@uva.nl, website: www.uva.nl/over-deuva/organisatie/medewerkers/content/b/o/c.a.m.vanboxtel/c.a.m.vanboxtel.

Current themes of research:

Teaching of history and heritage, historical thinking, and reasoning abilities.

Most relevant publications:

Van Boxtel, C., Grever, M., & Klein, S. (2015). Heritage as a resource for enhancing and assessing historical thinking. Reflections from the Netherlands. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 40-50). New York, NY: Routledge.

Van Boxtel, C., & Van Drie, J. (2012). "That's in the time of the Romans!" Knowledge and strategiesstudents use to contextualize historical images and documents.*Cognition and Instruction*, *30*(2), 113–145.

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.

Dr. Paul Holthuis. Department of Teacher Education, Faculty of Behavioural and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands. E-mail: p.holthuis@rug.nl, website: www.rug.nl/staff/p.holthuis.

Current themes of research:

Subject-specific instructions, history education, and heritage education.

Most relevant publications:

Huijgen, T., & Holthuis, P. (2014). Towards bad history? A call for the use of counterfactual reasoning in history education. *Historical Encounters: A Journal of Historical Consciousness, Historical Cultures, and History Education*, *1*(1), 103–110.

Huijgen, T., & Holthuis, P. (2015). 'Why am I accused of being a heretic?' A pedagogical framework for stimulating historical contextualisation.*Teaching History*, *158* 50–55.

Huijgen, T., Van Boxtel, C., Van de Grift, W., & Holthuis, P. (2014). Testing elementary and secondary school students' ability to perform historical perspective taking: The constructing of valid and reliable measure instruments. *The European Journal of Psychology of Education*, *29*(4), 653–672.