Testing elementary and secondary school students' ability to perform historical perspective taking: the constructing of valid and reliable measure instruments

Huijgen, T.; van Boxtel, C.; van de Grift, W.; Holthuis, P.

Link to publication

Testing elementary and secondary school students' ability to perform historical perspective taking: the constructing of valid and reliable measure instruments

Tim Huijgen · Carla van Boxtel · Wim van de Grift · Paul Holthuis

Abstract Historical reasoning competencies play an important role in history education. However, valid and reliable large-scale measurement instruments to assess these competencies are scarce. This study considers two instruments for measuring students' ability to perform historical perspective taking (HPT) as a historical reasoning competency. The instruments have been tested for validity and reliability among 1,270 Dutch upper elementary and secondary school students, ranging in age from 10 to 17 years. One instrument offers effective validity and reliability and can map HPT performance among a large and heterogeneous student population. The results show that even upper elementary school students are capable of performing HPT. However, as students age, their ability to perform HPT increases. Differences regarding the ability to perform HPT were also found between educational levels. Pre-university students performed HPT more successfully compared to students at lower educational levels. The results of this study can be used to gain insight into the construct of HPT and into how historical reasoning competencies such as HPT can be measured. Furthermore, the results provide insight into how differences between students, such as age and educational levels, influence the performance of HPT.

Keywords History education · Thinking skills · Historical reasoning · Assessment · Educational measurement

Introduction

During one observation of classroom practice, we heard a history teacher asking his students the following question: 'Can you explain why people in Germany voted for Hitler in the

T. Huijgen (✉) · W. van de Grift · P. Holthuis
Department of Teacher Education, Faculty of Behavioural and Social Sciences, University of Groningen, Landleven 1, Postbus 800, 9700 AV Groningen, The Netherlands
e-mail: t.d.huijgen@rug.nl

C. van Boxtel
Research Institute of Child Development and Education, University of Amsterdam, Nieuwe Prinsengracht 130, 1018 VZ Amsterdam, The Netherlands

1930s?' Most students answered that they could not understand why anyone would vote for such a terrible and evil leader, who was responsible for the deaths of millions. Just one student in this class described the historical context of Germany in the 1930s, coming to the conclusion that some people may well have voted for Hitler in response to the poor economic circumstances, German anger over the Treaty of Versailles, and widespread calls for a strong leader. This last one student was the only one to display historical perspective taking (HPT).

Historical reasoning competencies including HPT have become increasingly important for learning history (Barton and Levstik 2004; Foster and Yeager 2001; Haydn et al. 1997; Haydn and Counsell 2002; Lévesque 2008; Maggioni et al. 2004; Osborne 2006; O'Reilly 1991; Perfetti et al. 1995; Spoehr and Spoehr 1994; Seixas and Morton 2013; Van Drie and Van Boxtel 2008; Wineburg 2001). Historical reasoning competencies therefore have been incorporated in the history curricula of many countries such as the USA, Canada, the UK, Australia, New Zealand, Belgium, Germany, Finland and the Netherlands.

Despite the growing importance of historical reasoning competencies, valid and reliable large-scale measurement instruments for assessing these competencies are scarce. Rothstein (2004) noted that history teachers often assess only the factual background of history and not students' ability to perform historical reasoning. The reason for this, according to Rothstein, is the difficulty of constructing valid and reliable standardised tests. This difficulty is emphasised by Reich (2009), who was one of the few to attempt to measure historical reasoning competencies using multiple-choice items. However, he concluded that multiple-choice items merely tested history content, literacy, and test-wiseness but not important discipline-based thinking, such as HPT. Peck and Seixas (2008) noted that the focus of classroom assessment relies on factual recall and that, as a result, there is a lack of systematic assessment of students' progression in historical reasoning competencies. Students, teachers and educational professionals might therefore have an uncertain grasp on what progress in history education means, as Haydn (2011) noted. Recently, Fordham (2013) and VanSledright (2013) also argued for new assessment formats, if educational professionals wish to make sense of how students learn history and how they improve in it. Increasing numbers of research studies, projects, conferences and books concentrate on the assessment of history education to gain insight into its benefits and problems (e.g., Breakstone et al. 2013; Davies 2011; Harris and Foreman-Peck 2004; Martin et al. 2011; Seixas and Colyer 2012; SERVE 2006).

Our study should be placed in this context, and we took up the key challenge of constructing a reliable and valid measure instrument that could assess historical reasoning competencies within a large and heterogeneous student population and which was also time- and cost-effective. We focused on HPT because this student ability is crucial to learning history. Failing to perform HPT leads to important misunderstanding about the past (Barton and Levstik 2004; Davis 2001; Foster and Yeager 2001; Husbands 1996; Lee and Ashby 2001; Leinhardt et al. 1994; Lévesque 2008; Seixas and Morton 2013; Van Boxtel and Van Drie 2012; Wineburg 2001; Wineburg and Fournier 1994). Scholars also argue that HPT can contribute to citizenship competencies because recognising other people's views is necessary in a multicultural democracy (e.g., Barton and Levstik 2004; Den Heyer 2003).

Hartmann and Hasselhorn (2008) recently designed a measurement instrument that offers positive indicators for assessing students' ability of performing HPT. However, they tested their instrument only among a homogenous group of 170 10th-grade German students (16 years old) and focused on only one historical topic. Our study focuses on testing the instrument format among students in a larger and more heterogeneous student population and with two different historical topics to map possible differences between students. In this study, we first present the theoretical framework, starting with the conceptualisation of HPT and how it relates to historical reasoning. Subsequently, we look at what is already known about

students' ability to perform HPT and focus on the opportunities and difficulties that exist for measuring HPT. Then, our research questions, method, results, conclusions, and discussion will be presented.

Historical perspective taking: a conceptualisation

Without the ability to perform HPT, it is impossible to achieve historical reasoning and thinking (Lévesque 2008; Seixas and Morton 2013; Van Drie and Van Boxtel 2008). Seixas and Peck (2004) conceptualise HPT as an understanding of the social, cultural, intellectual, and emotional setting that shaped people's lives and actions, and they emphasise the importance of being aware of the difference between the past and present. Hartmann and Hasselhorn (2008) follow the definition of Lee and Ashby (2001) and define HPT as the application of the knowledge that historical agents had particular perspectives on their world that affected their actions. Foster and Yeager (2001) and Van Boxtel and Van Drie (2012) talk about the application of the knowledge and understandings of the historical context and chronology.

Based on a review of the literature, we distilled three elements necessary for performing HPT successfully. First, the ability to perform *historical contextualisation* was identified (e.g., Britt and Aglinskas 2002; Doppen 2000; Havekes et al. 2012; Leinhardt and McCarthy Young 1996; Nokes et al. 2007; Rouet et al. 1997; Van Boxtel and Van Drie 2012; Wineburg 1998). Historical contextualisation refers to building a context of circumstances or facts that surround the particular historical phenomenon to describe, compare, explain, or evaluate it (Van Drie and Van Boxtel 2008; Wineburg 1991). In history, it is possible to contextualise historical sources or historical phenomena, including persons, events, developments, or structures. In HPT, the focus is the contextualisation of actions of people and groups in the past. Students can therefore use chronological, spatial, and sociocultural frames of reference (De Keyser and Vandepitte 1998).

Second, students need to exhibit *historical empathy* (e.g., Davis 2001; Endacott 2010; Lee and Ashby 2001; Skolnick et al. 2004). Without the ability to imagine oneself in a situation that he or she is not likely to experience, the past remains an unopened book. However, historical empathy is not sympathy, as Eisenberg (2000) notes. Sympathy is compassion, sorrow or concern for another person. Historical empathy focuses on identifying with people in the past based on historical knowledge to explain their actions in the past.

Third, students have to avoid *presentism*, the bias by which people assume that the same goals, intentions, attitudes, and beliefs existed in the past as they exist today (e.g., Barton and Levstik 2004; Barton 1996; Lee and Ashby 2001; Stahl et al. 1996; Seixas and Morton 2013; Shemilt 1983; VanSledright and Afflerbach 2000; Wineburg 2001). The failure to perform HPT—and, therefore, the failure to explain, evaluate, or describe the past—often stems from this type of reasoning (Lee and Ashby 2001; Wineburg 2001). Its danger is explicitly mentioned in the *American National Standards for History*, which demands that students 'avoid present-mindedness, judging the past solely in terms of the norms and values of today' (National Center for History in the Schools 1996).

History education research has debated the extent to which HPT is an affective or cognitive achievement (e.g., Barton and Levstik 2004; Davis 2001; Endacott 2010; Foster and Yeager 1998). Some researchers claim that it is predominately a cognitive function (e.g., Foster 1999; Lee and Ashby 2001; Stern 1998), and others claim that it is more an affective process (Riley 1998; Skolnick et al. 2004). Although affective processes, such as connecting with known and familiar emotions of people in the past, may be at work during HPT, we consider it to be predominately a cognitive process in which students, based on historical evidence, perform historical contextualisation and historical empathy and avoid presentism.

Addressing the different needs of students

Unfortunately, we know relatively little about which students suffer from presentism and which students can perform HPT successfully. In accordance with Piaget's theory of the stages of cognitive development, researchers, such as Hallam (1970), have concluded that historical thinking is not possible for people younger than 16 years of age. These students cannot be expected to cope with abstract concepts or investigation, analysis, and interpretation—all of which are elements required to perform HPT successfully. However, Brophy and VanSledright (1997) argue that fifth graders (ages 10–11 years) can overcome their tendencies toward presentism and other biases to identify and empathise with people from the past. A general consensus among scholars concurs that children are capable of historical reasoning and HPT much earlier than Hallam suggested (e.g., Barton 1997; Foster and Yeager 1999; Levstik and Smith 1996; VanSledright 2002).

Specific information about which students perform HPT successfully is still lacking, however. This is a great concern with regard to the tendency for classrooms and schools to become increasingly diverse (Forsten et al. 2002; McCoy and Ketterlin-Geller 2004; Subban 2006; Tomlinson 2002; Tomlinson and Kalbfleisch 1998). Teachers should therefore know their students' competency levels, such as for HPT, to adapt their teaching and to reshape history curricula to fit it to student's needs (Jonassen and Grabowski 1993). However, one of the most important conclusions in the annual report of the Dutch Inspectorate of Education (2012) was that most teachers do have the basic skills to offer good teaching but are not able to provide teaching tailored to the different needs of students. The use of reliable and valid measurement instruments can help teachers and other educational professionals gain insight into student performance and can assist them in achieving the important ability of addressing the different needs of students.

Measuring the ability to perform historical perspective taking

Measuring historical reasoning competencies is a very difficult challenge (e.g., Haydn 2011; Peck and Seixas 2008; Reich 2009; VanSledright 2013). HPT can be measured through semistructured interviews (e.g., Berti et al. 2009; Lee and Ashby 2001; Shemilt 1987) and think-aloud assignments (e.g., Van Boxtel and Van Drie 2004; Wineburg 2001; Wooden 2008), but these methods are time- and cost-ineffective. Hartmann and Hasselhorn (2008) have recently developed an instrument using a hypothetical scenario with an item-rating format. Their study offers positive indicators for measuring HPT among a large and heterogeneous student population.

The scenario refers to the rise of the Nazi Party in Germany in the 1930s. The central historical agent is a young man who is deciding which political party to vote for in the next election. In relation to the historical story, the authors formulated nine items, corresponding to three stages of HPT: the present-oriented perspective, the role of the historical agent, and the historical contextualisation (Hartmann and Hasselhorn 2008). The three present-oriented perspective items display contemporary views on the past, whereas the three items pertaining to the role of the historical agent refer to his personal situation: What is his family like? Is he a member of the elite? This category is marked by the authors as an intermediate category between the present-oriented perspective and the historical contextualisation items. These latter items display historical contextualised thinking. The student's assignment is to place himself or herself in the historical context of this agent and decide if Hannes is willing to vote for the Nazi Party.

Hartmann and Hasselhorn (2008) found positive initial results for their instrument's reliability and validity. Their instrument is also a time- and cost-effective measurement instrument that can easily be implemented by, for example, teachers and test administrators. However, no study has tested the instrument in a large, heterogeneous population of students. Hartmann and Hasselhorn also raise the question about the instrument's reliability and validity should it incorporate a different historical topic. In this study, we took up these challenges. We tested the instrument in a different country among both upper elementary and secondary school students and developed a second version of the instrument to test the reliability and validity effects when a different historical topic was used.

Research questions

Despite the importance of historical reasoning competencies, almost no reliable and valid instruments exist to measure HPT among upper elementary and secondary school students. This results in little knowledge about the differences between students in terms of this capability. Therefore, we specify three research questions:

1. Does the instrument developed by Hartmann and Hasselhorn (2008) have positive reliability and validity outcomes when it is used to measure the ability to perform HPT among a large, heterogeneous student population in a different country?
2. What are the reliability and validity outcomes when the instrument format developed by Hartmann and Hasselhorn (2008) focuses on a different historical topic?
3. Which differences arise among students of different ages and educational levels regarding their ability to perform HPT?

Method

Constructing and adjusting the instruments

The first step was translating the hypothetical scenario and the accompanying items of the Nazi Party instrument developed by Hartmann and Hasselhorn (2008) into Dutch without affecting the instruments' interpretative framework. Hartmann and Hasselhorn excluded one instruments' item (ROA1) from their analysis because factor analysis showed that it violated the two-dimensional structure of their conceptualisation of HPT. We included this item in our instrument because our study has been conducted in a larger and more heterogeneous student population and therefore might fit in our conceptualisation of HPT.

As a second step, to investigate the effect of topic choice on a student's ability to perform HPT, we developed three other hypothetical scenarios and items about different historical topics, with the same item-rating format Hartmann and Hasselhorn (2008) used. The first scenario was about medieval witchery, the second scenario was about the Nazi occupation of the Netherlands from 1940 to 1945, and the last scenario focused on 19th century slavery. Constructing the scenarios and items was a difficult challenge because every historical topic has its own historical context with different related historical phenomena. HPT was embedded in different ways into the scenarios and with different student tasks. In the medieval witchery scenario, students had to explain the burnings of witches; in the Nazi occupation scenario, students had to decide what Dutch policemen would have done when asked to sign a document of collaboration with the Nazis. In the slavery scenario, we triggered HPT in the context of a

question to evaluate information from a historical source. All three newly developed scenarios and items intentionally were designed to give rise to students' emotions and their present values and beliefs just as Hartmann and Hasselhorn's instrument did, because we wanted to examine whether students could set aside their first emotional reaction, create a historical context and explain people's actions in the past.

To decide which additional instruments were the most suitable for use in our research and whether such instruments would be practically used by teachers in the classroom, we organised an expert panel composed of four history teacher educators from two universities (two with more than 4 years' work experience; two with more than 14 years' work experience), six secondary school history teachers (all six with more than 22 years' work experience), and two elementary school teachers (both with more than 16 years' work experience). The meeting took place in the context of a 1-day teacher-training program at the University of Groningen, and all teachers and teacher educators participated voluntarily.

All secondary school teachers and teacher educators were optimistic about the use of these instruments in classroom practice not only for assessing the ability to perform HPT but also as a practice and training instrument for their students. The secondary school teachers noted that history textbooks do not provide these types of assessment formats but focus more on assessing factual knowledge. The teachers also noted that using these instruments also supports other historical thinking and reasoning competencies, such as a critical evaluation of historical sources or providing solid argumentation. Furthermore, the secondary school teachers were optimistic about the use of the instruments as starting point for a whole-classroom discussion about, for example, the rise of Hitler in Nazi Germany.

The elementary school teachers were more restrained because they did not explicitly see the relevance of the instruments regarding the government's goals for elementary history education. However, they were positive about the 'empathy' aspect of the instruments and expected that such assignments would help students developing a better understanding of decisions made by a historical actor. The experts concluded that the topics of the Nazi occupation of the Netherlands and medieval witchcraft needed too detailed historical content knowledge, which would result in comprehension difficulty for upper elementary and young secondary school students. Therefore, we excluded these two scenarios and selected the slavery-related instrument as the second instrument.

The third and final step was shaping the two final instruments (see Appendix 1 for the Nazi Party instrument and Appendix 2 for the slavery instrument) in a manner that would make them suitable for both upper elementary and secondary school students. Therefore, we first conducted a qualitative pilot study among upper elementary ($n=6$) and secondary school students ($n=9$) to test the comprehension difficulty of the two instruments' hypothetical scenarios. Specifically, while students performed the assignment and thought aloud, their answers were transcribed and analysed to examine comprehension difficulty. We also asked the students to highlight difficult words in the scenarios and the accompanying items. The analysis of the pilot study showed that some abstract concepts in the hypothetical scenarios and question items were too difficult for upper elementary children. For example, the word *master* as a designation for a plantation owner in the slavery scenario caused confusion. In the hypothetical scenario of the Nazi Party, some upper elementary and secondary students also experienced difficulties with abstract concepts such as *conservative*.

Second, we asked elementary school teachers ($n=4$) and secondary school history teachers ($n=6$) in an expert panel to review both hypothetical scenarios and items for their levels of comprehension difficulty. All teachers involved in the expert panel had more than 15 years' work experience. The experts noted concerns about a few substantive concepts in the hypothetical scenarios that were found to be too difficult, especially for children in upper

elementary schools, such as *conservative*, *policy of appeasement*, and the name of the German political party *DVNP*.

The results of the qualitative pilot study and the expert panel meeting showed that both instruments needed minor revisions. We replaced difficult concepts with more specific terms or else removed them without affecting the interpretive framework of the hypothetical scenarios. In a second session with different upper elementary ($n=4$) and secondary ($n=5$) school students, we noticed that there was no more comprehension difficulty.

Sample and procedure

The study was conducted on 1,383 students in elementary ($n=178$) and secondary ($n=1,205$) schools—specifically, four elementary and 18 secondary schools in the northern part of the Netherlands. Missing data led us to exclude 113 cases, leaving 1,270 cases for further analysis. In the Dutch educational system, students begin their elementary education around the age of four and continue in elementary education for 8 years. In the last 2 years of their elementary education, students are advised about their further (secondary) education, including lower secondary professional education (4 years), senior general secondary education (5 years), or pre-university education (6 years). We included students undertaking elementary education, senior general secondary education, and pre-university education, as described in Table 1. Lower secondary professional education was not included in the research sample because of the different history curriculum of this type of education in which the ability to perform HPT played a far less substantial role compared to senior general secondary education and pre-university education.

The mean student age was 14.2 years (SD=2.2). In terms of gender, the distribution in the research sample was 45 % boys and 55 % girls; in the Netherlands, overall, the distribution between male and female students is 48 % and 52 %, respectively (Statistics Netherlands 2012). The participating schools generally matched the total population in terms of the number of students and graduation rates (Statistics Netherlands 2012).

The data collection took place during March and April 2012. Participating schools and teachers received hard copies of the instruments. Students were instructed at the beginning of a lesson to complete the instruments individually, in silence and without asking the teacher or other students for help. No time limit was given, but they all completed each instrument within 15 min. To assess students' prior knowledge about a topic, we included four multiple-choice items for each instrument. The multiple-choice items focused on historical content knowledge. For example, we asked for the year in which Hitler came to power in Germany and in which year the great worldwide economic depression was. Related to the slavery instrument, we

Table 1 Participants by age, educational level and gender ($n=1,270$)

Age	10	11	12	13	14	15	16	17	Total
Elementary male	20	43	14	*	*	*	*	*	77
Elementary female	34	50	17	*	*	*	*	*	101
Senior general secondary male	*	1	34	16	39	60	58	59	267
Senior general secondary female	*	2	45	13	42	63	61	62	288
Pre-university male	*	1	28	9	39	61	37	59	234
Pre-university female	*	2	37	17	53	76	44	74	303
Total	54	99	175	55	173	260	200	254	1,270

* No students of this age occur in the educational level

asked them to define the triangular trade and in which part of America slavery was most prominent in the 19th century. Furthermore, we asked for the students' ages, history grades, genders, and scores on a Dutch standardised final test (*Citotoets*) that is administered to upper elementary students. This optional test, commissioned by the Dutch Minister of Education and developed by the Dutch National Institute for Educational Measurement, aims to measure pupils' attainment of certain standards in elementary education. The test contains 290 multiple-choice items in the fields of language (100 items), mathematics (60 items), learning skills (40 items), and world orientation (90 items). World orientation is a combination of history and geography multiple-choice items and forms a substantial part of the test. The history items focus on content knowledge and historical reasoning competencies. For example, students have to date historical pictures and choose periods in which there was war in The Netherlands (Dutch National Institute for Educational Measurement 2013).

Data analysis

To answer the first two research questions, we began by examining the psychometric quality of the Nazi Party instrument and the slavery instrument. To be able to do this, we needed a coding system. In contrast with Hartmann and Hasselhorn (2008), who worked with latent class analysis, we used student mean scores on both instruments. Hartmann and Hasselhorn conducted their research on a small and homogeneous population. In our study, working with student mean scores showed the best results regarding the large and heterogeneous research population.

The present-oriented perspective items of both instruments used the following coding system from left to right for the answer columns (see Appendix 1 and Appendix 2 for the four columns). Selecting the first column yielded four points, the second column three points, the third column two points, and the last column one point. The role of the historical agent and historical contextualisation items had the opposite coding system from left to right. Selecting the first column yielded one point, the second column two points, the third column three points, and the last column four points. A mean category score was calculated by summing the category items' scores and dividing this score by three (because each category has three items). A total mean score of HPT was calculated by adding up the different mean category scores and dividing this score by three (because the instrument has three categories).

To test the content validity of both instruments, we selected ten history teachers as an expert panel. The ten teachers were randomly extracted from the teacher network of the Department of Teacher Education of the University of Groningen, which consisted of 52 history teachers. All teachers participated voluntarily and had more than 10 years' work experience. We also randomly selected ten historians from a pool of 44 historians as a second expert panel. The pool was created by making a list of historians who held a position at a university or at a university of applied sciences. Because they are professional historians accustomed to taking historical perspectives, they ought to score consistently high on the role of the historical agent and historical contextualisation items and low on the present-oriented perspective items. All historians in the pool held university degrees in the field of history and participated voluntarily. The instruments' content validity was tested on both expert panels. Furthermore, we performed a principal component analysis (PCA) and a reliability analysis using the Cronbach's alpha coefficient to explore the data structure and internal consistency of both instruments. Finally, we examined the predictive validity and calculated correlations between the scores of both instruments. To answer the third research question, we used the different mean category scores, plotted this by age and calculated correlations between the students' HPT scores and different student characteristics (viz., age and educational level).

Results

The psychometric quality of the Nazi Party and slavery instruments

The first two research questions focus on the reliability and validity of the instrument format developed by Hartmann and Hasselhorn (2008) when used in a different country, among a far larger and more heterogeneous student population and with a different historical topic. To answer both research questions, we looked at the instruments' content validity, dimensionality (i.e., whether the three categories of each instrument form one or multiple factors), internal consistency, and predictive validity.

Content validity of both instruments

We asked ten expert secondary school history teachers to sort the nine items of each instrument into the three categories (viz., the present-oriented perspective, the role of the historical agent, and the historical contextualisation) to confirm the categories' and items' face validity. A brief description of each category was provided, and they were instructed to place the items in the appropriate category. For both instruments, we calculated the agreement among the ten experts using the jury alpha and Fleiss's kappa, which we preferred to Cohen's kappa so that we could calculate the agreement among more than two raters. Fleiss's kappa values above 0.61 indicate substantial agreement; values greater than 0.81 are almost perfect agreement (Landis and Koch 1977). For the Nazi Party instrument, the jury alpha was 0.96, and Fleiss's Kappa was 0.64. The jury alpha for the slavery instrument was 0.98, and the Fleiss' kappa was 0.71.

Beyond face validity, we wanted to test the instruments for accuracy, so we invited ten professional historians to complete the measures. We calculated mean item scores for all three categories using a four-point scale. The expert scores on the historical contextualisation items were 3.88 (Nazi Party) and 3.77 (slavery); those for the role of the historical agent items were 3.56 (Nazi Party) and 3.23 (slavery). The scores on the present-oriented perspective items (using a reverse-coding scheme, in contrast to the role of the historical agent items and historical contextualisation items) were 3.93 (Nazi Party) and 3.89 (slavery). As we expected, the experts scored the role of the historical agent and historical contextualisation items high and did not reason from a present-oriented perspective.

In accordance with these findings and to refine our content validity results, we derived two hypotheses, in which we predicted higher HPT scores among (1) older students and (2) students with more topic knowledge. The mean student score (on a four-point scale) for the Nazi Party prior-topic knowledge test was 2.77 compared to 2.10 for the slavery prior-topic knowledge test. We calculated the correlation of students' total HPT scores with their ages and their prior topic knowledge scores. The results appear in Table 2.

Dimensionality and internal consistency of both instruments

The principal component analysis (PCA) served to examine the structure of our data collected using our instruments. In line with Hartmann and Hasselhorn (2008), we expected to find two

Table 2 Correlations of student HPT scores with age and prior knowledge (*n*=1,270)

[*]Correlations are significant at the 0.01 level

Instrument	Age	Prior knowledge
Nazi Party	0.35[*]	0.27[*]
Slavery	0.21[*]	0.24[*]

dimensions: one representing the two poles of a present-oriented perspective vs. a historical contextualisation and the other representing the role of the historical agent. The results of the PCA for the Nazi Party instrument in Table 3 reveal two factors extracted with eigenvalues greater than 1. They accounted for 42 % of the variance (factor 1: 28 %, factor 2: 14 %). The factor loadings after Varimax rotation with Kaiser normalisation also indicate that the present-oriented perspective items and historical contextualisation items constituted one factor. The three items pertaining to the role of the historical agent constituted the second factor. In contrast with Hartmann and Hasselhorn, our item ROA1 did not violate the simple structure.

The PCA results for the slavery instrument data (see Table 4), however, highlight three factors extracted with eigenvalues greater than 1. They accounted for 52 % of the variance (factor 1 21 %, factor 2 18 %, and factor 3 13 %). The factor loadings after Varimax rotation with Kaiser normalisation indicate that the present-oriented perspective items constituted one factor, the historical contextualisation items represented another factor, and the items pertaining to the role of the historical agent constituted a third factor.

Furthermore, we performed a reliability analysis using Cronbach's alpha coefficient to determine the internal consistency of both instruments (see Table 5). The slavery instrument showed a very low internal consistency score ($\alpha=0.25$), compared with the Nazi Party instrument ($\alpha=0.62$). Further analysis of the data showed that the historical agent items for both instruments were primarily responsible for this low internal consistency. Excluding these items from the analysis resulted in higher internal consistency scores for both instruments (slavery: $\alpha=0.49$; Nazi Party: $\alpha=0.69$).

Predictive validity

To assess the predictive validity of the instruments, we tested two hypotheses: namely, that the highest HPT scores would come from students with (1) high scores on the Dutch standardised final test for upper elementary students (*Citotoets*) and (2) high grades in history. Because historical reasoning and historical content knowledge form a substantial part of the final test, high scores on this test should be successful predictors for HPT performance. In line with Hartmann and Hasselhorn (2008), we also used history grades as a predictor for HPT performance. In Table 6, we present these correlation coefficients; the missing data are due to the non-obligatory nature of the *Citotoets*, such that not every Dutch elementary school (approximately 15 %) has implemented this test (Dutch National Institute for Educational Measurement 2013). The missing data regarding students' history grades exist because elementary school students do not have separate grades for history. We found a small but

Table 3 Principal component analysis results (rotated), Nazi Party instrument

Items	Factor 1	Factor 2
POP1	0.59	0.06
POP2	0.71	−0.03
POP3	0.72	0.01
ROA1	0.22	0.71
ROA2	−0.09	0.70
ROA3	−0.21	0.47
CONT1	−0.53	0.24
CONT2	−0.52	0.10
CONT3	−0.66	0.04

POP present-oriented perspective, *ROA* role of the historical agent, *CONT* historical contextualization

Table 4 Principal component analysis results (rotated), slavery instrument

Items	Factor 1	Factor 2	Factor 3
POP1	0.71	−0.11	0.06
POP2	0.77	0.02	0.09
POP3	0.73	0.01	0.05
ROA1	−0.04	−0.16	0.58
ROA2	0.08	0.04	0.77
ROA3	0.20	0.12	0.67
CONT1	0.09	0.69	0.02
CONT2	−0.07	0.78	−0.03
CONT3	−0.02	0.67	−0.01

POP present-oriented perspective, *ROA* role of the historical agent, *CONT* historical contextualisation

significant correlation between students' HPT scores and their *Citotoets* scores for the Nazi Party instrument but not for the slavery instrument. In addition, in contrast with Hartmann and Hasselhorn (2008), we did not find a significant correlation between students' history grades and their HPT scores.

Because we assume that both instruments test the same abilities of students, we calculated correlations between the category scores of the two tests across all students. The correlation coefficient between the present-oriented perspective category scores was 0.24, which was significant at the 0.01 level. We did not find a significant correlation between the two instruments category scores for the role of the historical agent. Between the historical contextualisation category scores, there was a significant correlation of 0.23 at the 0.01 level. The correlation coefficient between the total HPT scores of all students was 0.23, which was significant at the 0.01 level.

Differences among the students when executing HPT

The third research question focuses on possible differences between students regarding their ability to perform HPT. The data obtained from the slavery instrument offered too low of an internal consistency to support the reliability of the data; therefore, we decided to work only with the Nazi Party instrument's data. Using these data, we investigated student mean scores for the three different categories (viz., the present-oriented perspective, role of the historical agent, and historical contextualisation), plotted by age and educational level.

Figure 1 presents the three mean category scores for students between the ages of 10 and 17 years. Both the declining trend for the present-oriented perspective category and the ascending trend for the historical contextualisation category are notable. Starting at approximately 11 years of age, students began scoring higher in the historical contextualisation

Table 5 Internal consistency of two instruments (*n*=1,270)

Instrument	(α) POP	(α) ROA	(α) CONT	(α) POP, ROA, and CONT	(α) POP and CONT
Slavery	0.60	0.42	0.54	0.25	0.49
Nazi Party	0.62	0.30	0.51	0.62	0.69

POP present-oriented perspective, *ROA* role of the historical agent, *CONT* historical contextualisation

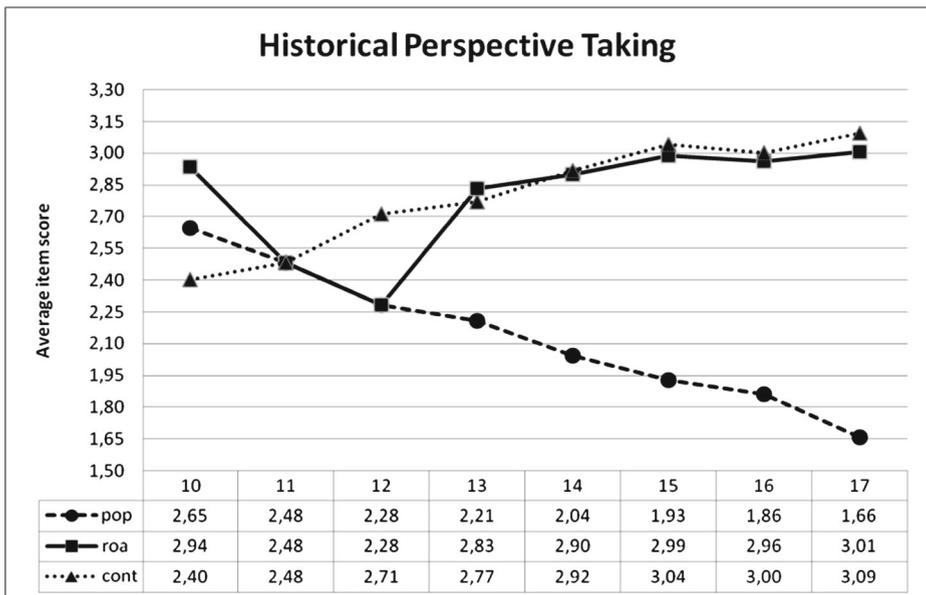**Table 6** Correlations between HPT scores and student characteristics

Instrument	*Citotoets*	Students' history grades
Nazi Party	0.17[*]	−0.02
Slavery	0.06	−0.01
N	659	885

[*]Correlations are significant at the 0.01 level

category than in the present-oriented perspective category. With regard to the role of the historical agent, a decline occurred between the ages of 10 and 12 years, then after the age of 12, the line began to ascend, similar to the historical contextualisation scores.

We calculated correlations for further analysis. Between 13 and 17 years (secondary education), the students showed a small but significant correlation of 0.11 (at the 0.01 level) between their scores in the category measuring the role of the historical agent and in the historical contextualisation category. We did not find such a significant correlation (at the 0.01 or 0.05 level) when students were between 10 and 12 years of age (elementary education). Both general secondary and pre-university education showed the same trend (as plotted in Fig. 1) between the ages of 12 and 17.

When compared with students in other educational levels, the pre-university students scored the highest on HPT. A one-way analysis of variance-based post hoc multiple comparison with assumed Scheffé equal variance was used to test for any significant differences across the different educational levels. The difference between general secondary education

Historical Perspective Taking

	10	11	12	13	14	15	16	17
pop	2,65	2,48	2,28	2,21	2,04	1,93	1,86	1,66
roa	2,94	2,48	2,28	2,83	2,90	2,99	2,96	3,01
cont	2,40	2,48	2,71	2,77	2,92	3,04	3,00	3,09

Notes: POP = present-oriented perspective, ROA = role of the historical agent, CONT = historical contextualization.

Fig. 1 Historical perspective taking, plotted by age, Nazi Party instrument (*n*=1,270)

(total score of 2.44, SD=0.51) and pre-university education (total score of 3.15, SD=0.50) was significant at the 0.05 level. The comparison of elementary education with both general secondary education (total score of 2.90, SD=0.54) and pre-university education showed significant differences at the 0.01 level.

Discussion and conclusions

Our study focused on the reliability and validity of the instrument of Hartmann and Hasselhorn (2008) when tested among a large and heterogeneous student population in a different country and when applied to a different historical topic. Furthermore, we explored possible differences between students on HPT performance. In this section, we discuss our findings, outline limitations of our study, and present suggestions for further research.

Regarding the first research question, we found—in line with Hartmann and Hasselhorn (2008)—positive indicators for the reliability and validity for the Nazi Party instrument. We also concluded that HPT is a two-dimensional construct consisting of (1) a dimension characterised by present-oriented perspective and historical contextualisation poles and (2) items pertaining to the role of the historical agent. A PCA performed on our data from the Nazi Party instrument confirmed this. The reliability analyses indicated very acceptable (nearly characterisable as good) internal consistency for the Nazi Party instrument when we excluded the items tapping the role of the historical agent. We do not know the implications of the role of the historical agent items and its relation with the historical reasoning competency of HPT. Thinking-aloud methods could provide more insight if the role of historical agent items can contribute to students' ability to perform HPT.

To examine the second research question, we used a different historical scenario about 19th century slavery with the same item-rating format. When examining the psychometric qualities of this slavery instrument, we did find positive evidence for content validity but not for predictive validity or internal consistency. In line with Hartmann and Hasselhorn's conclusions, our findings using the data obtained from the Nazi Party instrument showed HPT emerging as a two-dimensional construct. However, with the slavery instrument, our PCA identified three dimensions that were separately associated with each perspective (viz., present-oriented perspective, role of the historical agent, and historical contextualisation).

Regarding the third research question, using the data obtained from the Nazi Party instrument, we found that upper elementary school students, starting at the age of 10 years, successfully performed some historical contextualisation efforts. This is in line with research conducted by Barton (1997), Field (2001), and Brophy and VanSledright (1997). However, they also displayed more presentism, which resulted in higher scores on the present-oriented perspective items. Older students achieved higher scores for historical contextualisation than younger students, and pre-university students held the highest HPT scores compared with students in general secondary and elementary education groups.

There may be several reasons for the differences in reliability and validity observed between the slavery instrument and the Nazi Party instrument. Because we embedded testing students' ability to perform HPT into determining the usefulness of a source for making statements about the past, the observed differences might have stemmed from the specific instructions provided for the slavery instrument. For this instrument, students had to approach the story about how the enslaved people were treated as historical source. In addition to performing HPT, students also had to execute other historical thinking and reasoning

competencies related to the use of historical evidence (e.g., assessing the reliability of the source) when completing the slavery instrument successfully. This dimension is missing from the Nazi Party instrument.

The differences also might be explained by the students' having less prior knowledge of slavery. Students scored lower on the slavery prior knowledge questions compared with the topic knowledge questions related to the Nazi Party. Van Boxtel and Van Drie (2012) concluded that knowledge of key historical concepts and dates plays an important role in a student's ability to contextualise a historical source. Thinking-aloud methods could be a valuable addition for gaining insight about whether students use knowledge (and what knowledge they do use) when responding to the slavery items and whether they notice differences in how the items are constructed.

Although we found a significant correlation between students' scores between the Nazi Party instrument and the slavery instrument, the results show that the slavery instrument did not meet our reliability and validity criteria. The secondary and elementary school teachers who were consulted were encouraging about the use of these instruments in classroom practice as both an assessment format and as a training exercise to stimulate HPT. Still, we do not exactly know if it is possible to test the ability to perform HPT in a reliable and valid way using items reflecting a present-oriented perspective, the role of the historical agent and historical contextualisation in the context of different historical topics. Our results illustrate the difficulties that are encountered when trying to construct a new instrument with this item-rating format using the same types of items used in the Nazi Party instrument.

We must take into account the limitations of our study. Our instruments focus on a student's ability to consider the historical actors' personal situations (i.e., the role of the historical agent) and the broader historical context (i.e., the present-oriented perspective and historical contextualisation). This is a narrow view of HPT because scholars also refer to students' awareness of the differences between past and present (e.g., Seixas and Peck 2004), the sense of a period (Dawson 2009), and the application of different frames of references, specific (prior) knowledge, and understanding of the historical context and chronology (e.g., De Keyser and Vandepitte 1998; Van Boxtel and Van Drie 2012). For example, if students have little prior knowledge about a topic, they might refer more to specific characteristics of the historical actor to perform HPT (Berti et al. 2009). These are difficult abilities to measure using only the instruments described in this study.

A more comprehensive measurement procedure might be necessary if we want to include the measurement of students' underlying knowledge and understanding. Constructing items that take into account, different frames of knowledge might provide insight into which different frames of references are used by students when performing HPT (e.g., De Keyser and Vandepitte 1998). The addition of thinking-aloud methods could also facilitate improved insight into whether students apply specific knowledge of topics and whether they combine this with knowledge about the specific characteristics of the historical agent. Combining the instruments with related historical empathy tasks and historical content tests might also provide insight into the roles played by distinguished elements (viz., historical contextualisation, historical empathy, and avoiding presentism) when students perform HPT.

Another limitation is that both instruments focused purposefully on topics that give strong rise to students' emotions, such as anger and compassion, and these emotions may hinder efforts to better understand the past (Von Borries 1994). It would be interesting to see how students perform HPT with respect to historical topics that do not explicitly give rise to emotions such as the invention of the steam engine. Furthermore, the items and the scenarios

do not represent the whole historical context of the historical phenomena. The instruments had to be suitable for elementary school students; therefore, the items might consist of more simple functional explanations about the past (e.g., Bermúdez and Jaramillo 2001). Constructing more items for each category or using different instruments focusing on the same historical topic might tackle this problem.

Further research should focus on the question of whether it is possible to construct a reliable and valid measurement of the ability to perform HPT, without the dependency of a specific historical topic and without being embedded in different tasks, such as historical empathy tasks in which students are asked to take the perspective of a fictional or genuine historical person or to examine the trustworthiness and usefulness of a historical account. More research is also needed to investigate how students perform when the central historical agent of the instrument is, for example, a child or a politician. Students might identity themselves more with other children or heroic figures than politicians, and this might affect their ability to perform HPT (Brophy and VanSledright 1997).

Additionally, the differences between taking the historical perspective of a group vs. taking the perspective of an individual should be further elaborated, following an interesting question raised by Berti et al. (2009). Furthermore, we only used one type of source: a textual story and its accompanying items. Textual sources play a very important role in history education, but so do visual sources. Further examination needs to be made of the differences that exist in HPT performance when the source is non-textual. Finally, but not less important, further research should focus on the role of the teacher. What types of instruction do teachers use to stimulate HPT in elementary and secondary education? Can the role of the historical agent be used as a scaffold for stimulating HPT? Such research could provide more insight into how to stimulate the important ability to perform HPT.

Appendixes

Appendix 1. The Nazi Party instrument

Dusseldorf, Germany in 1930. Hannes (20 years old) is the son of a man who owns a small factory that makes handmade shoes. One day Hannes meets with his friend Gerd. They talk about the situation in Germany and the upcoming elections. Hannes says: "My father's company might close down. Since the war ended, everything is getting worse and worse. After the economic crisis of 1923, we began to feel some hope again. But now, it is worse than ever. I don't know how this is going to end. Right now, I still have a job in my father's business. But when he closes down, I have no idea where to get a job. We have always been wealthy people—and look at us now!" Gerd replies. "You are right. What has happened to our country? Look at what is going on today. No one has work. Hannes replies: "My father always says that we were better off during the time of the German Empire. What can we do if our country is suffering from a crisis and the winners of the war are hurting us wherever they can? Our politicians are not decisive and do us no good. It's time that Germany is ruled by someone who knows what he is doing and who really takes the lead. During the last election, I supported the German Democratic Party, but I do not know if they have got the right people to save our country."

Below you will find some statements. Read through all the statements first. Then try to take Hannes' perspective and mark for every statement how well it fits his situation. Could Hannes vote for an anti-democratic party like the NSDAP?

Item		Doesn't fit his situation at all	Doesn't fit his situation too much	Fits his situation somewhat	Fits his situation very well
POP 1	He will definitely not vote for the NSDAP. No one can approve what this party has done to the world.	☐	☐	☐	☐
POP 2	He will see that only in a democracy can people take part in decision-making. That is why he will decide wisely and that does not mean NSDAP.	☐	☐	☐	☐
POP 3	He will not vote for the NSDAP. Their ideas are easy to see through. It is clear that this party wants a war.	☐	☐	☐	☐
ROA 1	As a member of a wealthy family, he would like to return to the German Empire, where his family was better off. He can vote on an anti-democratic party.	☐	☐	☐	☐
ROA 2	As son of a businessman, he could vote for a party that strives to keep things as they are, but not necessarily for the NSDAP.	☐	☐	☐	☐
ROA 3	Because his father's business is almost bankrupt, he could vote for a party that protects small business owners.	☐	☐	☐	☐
CONT 1	Hannes has little experience with democracy. He probably does not know the risk that the NSDAP entails and thus will probably vote for the NSDAP.	☐	☐	☐	☐
CONT 2	To him, Hitler probably represents a strong leader. He probably would not think too much about the threats connected with the NSDAP.	☐	☐	☐	☐
CONT 3	In his situation, he only sees the disadvantages of democracy. Therefore, he might fall for the ideas of the NSDAP.	☐	☐	☐	☐

Items are placed in order of the category of historical perspective taking. In the original instrument, items were placed in random order

POP present-oriented perspective, *ROA* role of the historical agent, *CONT* historical contextualisation

Appendix 2. The slavery instrument

Harry Knox, a journalist working for the respected American newspaper *Austin Press*, interviewed in 1891 the 70-year-old Ben Simpson. The enslaved Simpson worked for over 20 years at an American plantation in Texas and told Knox the following story about his life:

"The plantation-owner was in charge of a large plantation. When he pulled me and the others of the boat, he chained us around our necks. The chains were fixed to the horses. With the chains we—my mother, my sister Emma, I and the other slaves had to walk all the way to his plantation in Texas. Somewhere along the way it started to snow, but the plantation-owner did not care about our bare feet. We had to sleep in the snow on the ground. The plantation-owner had a long whip, made of leather. And if one of us fell behind, then he would hit him with it. We had no tents. When the night came, he fixed our chains to a tree. The ground was our bed. A little raw meat and corn were the only things we ate. Often, I ate weeds and I was very hungry. He let us never eat during the day, and forced us to walk the whole day without any breaks. He branded me, my mother and my sister. At the border of Texas my mother couldn't go any further. Her feet were broken and bleeding, her legs were swollen.

The master took his gun and shot her. He didn't bury her, he left her lying where he had shot her."

Which statements are useful in order to describe the relationship between slaves and plantation owners in 19th century America?

Item		Not useful	Not too useful	Somewhat useful	Very useful
POP 1	The plantation owner committed a severe crime. He should be arrested and brought to trial.	☐	☐	☐	☐
POP 2	Slaves are also people. That is why the plantation owner should treat them right.	☐	☐	☐	☐
POP 3	Food is a primary need. This is why the plantation owner should provide the slaves with enough food.	☐	☐	☐	☐
ROA 1	Simpson was an eye-witness. He told his story to the *Austin Daily*, because he thought it was a reliable newspaper.	☐	☐	☐	☐
ROA 2	Simpson was the son of a woman who was killed by a plantation owner. That's why he might not be objective in his story.	☐	☐	☐	☐
ROA 3	Because Simpson was an enslaved man who suffered terribly, he probably only remembered the bad things.	☐	☐	☐	☐
CONT 1	The plantation owners were harsh, because they wanted the slaves to work harder in order to increase profit.	☐	☐	☐	☐
CONT 2	The plantation owners were strict, because they feared rebellion among the slaves.	☐	☐	☐	☐
CONT 3	Plantation owners saw slaves as economic products and not as people. They thought: if a product is broken, I will buy a new one.	☐	☐	☐	☐

Items are placed in order of the category of historical perspective taking. In the original instrument, items were placed in random order

POP present-oriented perspective, *ROA* role of the historical agent, *CONT* historical contextualisation.

References

Barton, K. (1996). Narrative simplifications in elementary children's historical understanding. In J. Brophy (Ed.), *Advances in research on teaching* (pp. 51–83). Greenwich: JAI Press.

Barton, K. (1997). "I just kinda know": elementary students' ideas about historical evidence. *Theory and Research in Social Education, 25*, 407–430.

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

Bermúdez, A., & Jaramillo, R. (2001). Development of historical explanation in children, adolescents and adults., vol. 3. In A. Dickinson, P. Gordon, & P. Lee (Eds.), *Raising standards in history education* (pp. 146–167). London: Woburn Press.

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.

Breakstone, J., Smith, M., & Wineburg, S. (2013). Beyond the bubble in history/social studies assessments. *Phi Delta Kappan, 94*(5), 53–57.

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

Brophy, J., & VanSledright, B. (1997). *Teaching and learning history in elementary schools*. New York: Teachers College Press.

Davies, I. (2011). *Debates in history teaching*. London: Routhledge.

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: 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 Keyser, R., & Vandepitte, P. (1998). *Historical formation. Design of vision*. Brussel: Flemish Board for Catholic Secondary Education.

Den Heyer, K. (2003). Between every "now" and "then": a role for the study of historical agency in history and citizenship education. *Theory & Research in Social Education, 31*(4), 411–434.

Doppen, F. (2000). Teaching and learning multiple perspectives: the atomic bomb. *The Social Studies, 91*(4), 159–169.

Dutch Inspectorate of Education. (2012). The State of Education in the Netherlands. *Annual Report*. http://www.onderwijsinspectie.nl/english. Accessed 28 May 2013.

Dutch National Institute for Educational Measurement. (2013). CITO file elementary education final test 2013. http://www.cito.nl. Accessed 25 July 2013.

Eisenberg, N. (2000). Empathy. In K. E. Alan (Ed.), *Encyclopedia of psychology, vol. 3* (pp. 179–182). London: Oxford University Press.

Endacott, J. (2010). Reconsidering affective engagement in historical empathy. *Theory and Research in Social Education, 38*(1), 6–47.

Field, S. (2001). Perspectives and elementary social studies: Practice and promise. In O. L. Davis, E. A. Yeager, & S. J. Foster (Eds.), *Historical empathy and perspective taking in the social studies* (pp. 115–138). New York: Rowman and Littlefield.

Fordham, M. (2013). O brave new world, without those levels in't: where now for key stage 3 assessment in history? *Teaching history. The Historical Association Curriculum Supplement, 153*(1), 16–23.

Forsten, C., Grant, J., & Hollas, B. (2002). *Differentiated instruction. Different strategies for different learners*. Peterborough: Crystal Springs Books.

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., & Yeager, E. A. (1998). The role of empathy in the development of historical understanding. *International Journal of Social Education, 13*(1), 1–7.

Foster, S., & Yeager, E. A. (1999). "You've got to put together the pieces": english 12-year-olds encounter and learn from historical evidence. *Journal of Curriculum and Supervision, 14*(4), 286–317.

Foster, S., & Yeager, E. A. (2001). The role of empathy in the development of historical understanding. In O. L. Davis, E. A. Yeager, & S. Foster (Eds.), *Historical empathy and perspective taking in the social studies* (pp. 13–21). Oxford: Rowman and Littlefield.

Hallam, R. (1970). Piaget and thinking in history. In M. Ballard (Ed.), *New movements in the study and teaching of history* (pp. 162–178). Bloomington: Indiana University Press.

Harris, R., & Foreman-Peck, L. (2004). 'Stepping into other people's shoes': teaching and assessing empathy in the secondary history curriculum. *International Journal of Historical Learning, Teaching and Research, 4*(2), 1–14.

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

Havekes, H., Coppen, P. A., 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. (2011). Secondary history. Current themes. In I. Davies (Ed.), *Debates in history teaching* (pp. 30–45). New York: Routledge.

Haydn, T., & Counsell, C. (2002). *History, ICT and learning*. London: Routledge-Falmer.

Haydn, T., Arthur, J., Hunt, M., & Stephen, A. (1997). *Learning to teach history in the secondary school*. London: Routledge.

Husbands, C. (1996). *What is history teaching? Language, ideas and meaning in learning about the past*. Buckingham: Open University Press.

Jonassen, D., & Grabowski, B. (1993). *Handbook of individual differences, learning and instruction*. Hillsdale: Lawrence Erlbaum Associates.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lee, P., & Ashby, R. (2001). Empathy, perspective taking, and rational understanding. In O. L. Davis, E. A. Yeager, & S. Foster (Eds.), *Historical empathy and perspective taking in the social studies* (pp. 21–50). Oxford: Rowman and Littlefield.

Leinhardt, G., & McCarthy Young, K. (1996). Two texts, three readers: distance and expertise in reading history. *Cognition and Instruction, 14*(4), 441–486.

Leinhardt, G., Beck, I. L., & Stainton, C. (1994). *Teaching and learning in history*. Hillsdale: Erlbaum.

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

Levstik, L., & Smith, D. (1996). "I've never done this before": Building a community of historical inquiry in a third-grade classroom. In J. Brophy (Ed.), *Advances in research on teaching* (pp. 85–114). Greenwich, CT: JAI Press.

Maggioni, L., Alexander, P., & VanSledright, B. (2004). At a crossroads? The development of epistemological beliefs and historical thinking. *European Journal of School Psychology, 2*(2), 169–197.

Martin, D., Maldonado, S.I., Schneider, J., & Smith, M. (2011). A report on the state of history education: State policies and national programs. http://teachinghistory.org/system/files/teachinghistory_special_report_2011.pdf. Accessed 7 March 2014.

McCoy, J., & Ketterlin-Geller, L. (2004). Rethinking instructional delivery for diverse student populations. *Intervention in School and Clinic, 40*(2), 88–95.

National Center for History in the Schools. (1996). *National standards for history* (basicth ed.). CA: Los Angeles.

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

O'Reilly, K. (1991). Informal reasoning in high school history. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 363–379). Hillsdale, NJ: Erlbaum.

Osborne, K. (2006). 'To the past': Why we need to teach and study history. In R. Sandwell (Ed.), *To the past: history education, public memory and citizenship in Canada* (pp. 103–131). Toronto: University of Toronto Press.

Peck, C., & Seixas, P. (2008). Benchmarks of historical thinking: first steps. *Canadian Journal of Education, 31*(4), 1015–1038.

Perfetti, C., Britt, M., & Georgi, M. (1995). *Text-based learning and reasoning: studies in history*. Hillsdale: Erlbaum.

Reich, G. (2009). Testing historical knowledge: standards, multiple-choice questions and student reasoning. *Theory & Research in Social Education, 37*(3), 325–360.

Riley, K. (1998). Historical empathy and the Holocaust: theory into practice. *International Journal of Social Education, 13*(1), 32–42.

Rothstein, R. (2004). We are not ready to assess history performance. *Journal of American History, 90*(4), 1381–1391.

Rouet, J., Favart, M., Britt, M., & Perfetti, C. (1997). Studying and using multiple documents in history: effects of discipline expertise. *Cognition and Instruction, 15*(1), 85–106.

Seixas, P., & Colyer, J. (2012). *Assessing historical thinking. A report on the national meeting of the historical thinking project*. http://historicalthinking.ca/sites/default/files/2012Report_EN.pdf. Accessed 3 August 2013.

Seixas, P., & Morton, T. (2013). *The big six historical thinking concepts*. Toronto: 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: Pacific Educational Press.

SERVE. (2006). *How to assess student performance in history: Going beyond multiple-choice tests*. http://www.serve.org/uploads/publications/AssessHistory.pdf. Accessed 7 June 2013).

Shemilt, D. (1983). The devil's locomotive. *History and theory, 22*, 1–18.

Shemilt, D. (1987). Adolescent ideas about evidence and methodology in history. In C. Portal (Ed.), *The history curriculum for teachers* (pp. 39–61). London: The Falmer Press.

Skolnick, J., Dulberg, N., & Maestre, T. (2004). *Through other eyes: developing empathy and multicultural perspectives in the social studies*. Toronto: Pippin.

Spoehr, K., & Spoehr, L. (1994). Learning to think historically. *Educational Psychologist, 29*(2), 71–77.

Stahl, S., Hynd, C., Britton, B., McNish, M., & Bosquet, D. (1996). What happens when students read multiple source documents in history? *Reading Research Quarterly, 31*(4), 430–456.

Statistics Netherlands. (2012). http://www.cbs.nl/english. Accessed 10 Augustus 2012.

Stern, B. (1998). Addressing the concept of historical empathy: Wit frauen: German women recall the Third Reich. *International Journal of Social Education, 13*(1), 43–48.

Subban, P. (2006). Differentiated instruction: a research basis. *International Education Journal, 7*(7), 935–947.

Tomlinson, C. (2002). Different learners different lessons. *Instructor, 112*(2), 21–25.

Tomlinson, C., & Kalbfleisch, M. (1998). Teach me, teach my brain: a call for differentiated classrooms. *Educational Leadership, 56*(3), 52–55.

Van Boxtel, C., & Van Drie, J. (2004). Historical reasoning: a comparison of how experts and novices contextualize 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 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.

VanSledright, B. (2002). Fifth graders investigating history in the classroom: results from a researcher-practitioner design experiment. *The Elementary School Journal, 103*(2), 131–160.

VanSledright, B. (2013). *Assessing historical thinking and understanding. Innovative designs for new standards.* New York: Taylor & Francis.

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.

Von Borries, B. (1994). (Re-)Constructing history and moral judgment: On relationships between interpretations of the past and perceptions of the present. In M. Carretero & J. F. Voss (Eds.), *Cognitive and instructional processes in history and the social sciences* (pp. 285–308). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wineburg, S. (1991). Historical problem solving: a study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology, 83*(1), 73–87.

Wineburg, S. (1998). Reading Abraham Lincoln: an expert/expert study in the interpretation of historical texts. *Cognitive Science, 22*(3), 319–346.

Wineburg, S. (2001). *Historical thinking and other unnatural acts. Charting the future of teaching the past.* Philadelphia: Temple University Press.

Wineburg, S., & Fournier, J. (1994). Contextualized thinking in history. In M. Carretero & J. F. Voss (Eds.), *Cognitive and instructional processes in history and the social sciences* (pp. 285–308). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wooden, J. (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), 23–31.