Erschienen in: American Educational Research Journal ; 54 (2017), 3. - S. 444-484
https://dx.doi.org/10.3102/0002831217694833

American Educational Research Journal
June 2017, Vol. 54, No. 3, pp. 444–484
DOI: 10.3102/0002831217694833

Learning Historical Thinking With Oral History Interviews: A Cluster Randomized Controlled Intervention Study of Oral History Interviews in History Lessons

Christiane Bertram
Wolfgang Wagner
Ulrich Trautwein
University of Tübingen

*The present study examined the effectiveness of the oral history approach with respect to students' historical competence. A total of 35 ninth-grade classes (*N *= 900) in Germany were randomly assigned to one of four conditions—live, video, text, or a (nontreated) control group—in a pretest, posttest, and follow-up design. Comparing the three intervention groups with the control group, the intervention groups scored better on four of the five achievement tests. Comparing the live group with the video and text groups, students in the live condition were more convinced of their learning progress at both measurement points. However, they scored lower than the video/text group on two achievement measures and higher on one at the posttest.*

Christiane Bertram is an assistant professor at the University of Konstanz, 78457 Konstanz, Germany; e-mail: *christiane.bertram@uni-konstanz.de*. She graduated from the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen. Her main research interests include the assessment and improvement of historical thinking skills in the field of teaching history.

Wolfgang Wagner, PhD, is research assistant at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen. His main research interests include the assessment of characteristics of learning environments and their effects on the development of academic achievement as well as methodological issues in the field of multilevel structural equation models.

Ulrich Trautwein is professor at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen. His research interests include the development of student motivation, personality, academic effort, and achievement.

KEYWORDS: historical thinking competencies, intervention study, randomized controlled field study, teaching history

When Karl Dönitz, former Grand Admiral and Adolf Hitler's successor as of April 1945, delivered an oral report at a school on the internal affairs of the Nazi Regime in January 1963, students and teachers were fascinated by his personal authority and exciting stories. He told them confidently that he and the other men at the front had only done their duty as soldiers in war, and he convincingly concealed his own responsibility as a high-level Nazi leader. The auditorium was enthusiastic, and the local newspaper characterized his oral report about the past as "the best-done history lesson" (see Lutteroth, 2011).

Whereas it is likely that Grand Admiral Dönitz purposely provided a distorted historical account, other examples show that witnesses of the past are often not able to provide an adequate account of the past even when they want to (J. Assmann & Czaplicka, 1988/1995; Plato & Burley, 2009; Portelli, 1981/1991). In fact, there is evidence from several disciplines, including cognitive and clinical psychology (Welzer & Markowitsch, 2005), neuropsychology (Bridge & Paller, 2012), psychology of law (Loftus, 1999; Loftus & Palmer, 1974), and history education (Wineburg, 2001), that "remembering" is always a reconstructive process that is influenced by several factors rather than a "copy" of the past. Accordingly, oral witness reports might not constitute what students (and perhaps some teachers) think they do.

Despite these obvious risks, oral history as an instructional format in history lessons has become very popular in Germany, the United States, and many other countries. A Google search (on *oral history in school* yielding 10,400,000 results in 0.49 seconds on July, 22, 2016) highlights the growing number of "active" oral history projects in schools across the United States and around the world. It is believed that oral histories in schools provide exceptionally good learning experiences (e.g., Lanman, 1987; Lanman & Wendling, 2006b; Loewen, 2004; Ritchie, 2011; Whitman, 2004, 2011). In Germany, where the present study took place, the school authorities recommend working with the oral reports of contemporaries, especially in the ninth grade when the events of the late 20th century are commonly taught (Standing Conference, 2009).

But do oral history lessons actually have the effects that are intended by their use? Unfortunately, there have been very few attempts to put this question to a hard empirical test. The current study is based on a fairly large sample of 35 history classes (N = 900 students) that were randomly assigned to four conditions (interviewing an eyewitness of the past, working with a video interview, working with the transcript of an interview, no intervention). Moreover, the present study took a multidimensional approach by using several psychometrically sound instruments to measure the effects of working

with oral history interviews on students' performance on historical competence tests at three measurement points.

Competencies in Historical Thinking

For several decades, the major goals of history education have been shifting from factual knowledge to competencies in historical thinking. Scholars in Europe and in the Anglophone world define historical thinking as a familiarity and facility with disciplinary ways of interpreting and reasoning with historical texts, an appreciation of the slippery nature of historical knowledge, and the application of conceptual, narrative, and factual knowledge (Monte-Sano & Reisman, 2016). Students have to understand the distinction between the *past* (all events and incidents that have gone on in the world before) and *history* (the practice of interpreting the past on the basis of the residues we have). History makes the effort to interpret the past in a manner that is particularistic, selective, and laced with perspective. If students equate the past and history, they miss the whole complex interpretative process (in other words, *historical thinking*) that must be undertaken to make sense of the past (Rüsen, 2005; VanSledright, 2014). Research indicates that epistemic cognition in history (see the overview in VanSledright & Maggioni, 2016) as well as the processes of historical thinking (see the overview in Monte-Sano & Reisman, 2016) do not come easily or intuitively to students. Therefore, it is a main goal of history lessons in Western democracies to provide students with the competence to think historically (e.g., Erdmann & Hasberg, 2011; Köster, Thünemann, & Zülsdorf-Kersting, 2014; Stearns, Seixas, & Wineburg, 2000; VanSledright, 2004).

The shift from history lessons that primarily teach "facts" to a more complex approach can also be witnessed in the Teaching American History program (TAH; see www.teachingamericanhistory.org), which concluded in 2011. In the context of this program, teachers were trained to foster historical thinking in history lessons (e.g., by working with primary sources and reconciling different points of view). In a similar way, researchers argue that focusing on critical thinking incites increased curiosity and stimulates more attention in history lessons than merely studying textbooks and learning facts (Clark, 2008; VanSledright & Reddy, 2014). Teachers and students are required to go beyond the facts presented in textbooks (Barber & Peniston-Bird, 2009; Loewen, 1996), use primary sources in history lessons (De La Paz, Malkus, Monte-Sano, & Montanaro, 2011; Westhoff, 2009), work like historians (e.g., asking questions and using multiple sources; Mandell, 2008), and consider the strategies of sourcing, corroboration, and contextualization (see Wineburg, 1991, 1998, 2001). Similarly, according to the American Common Core State Standards Initiative (2010), students have to compare multiple historical accounts, consider source information, and use evidence in discussions. Furthermore, as a prerequisite for truly

addressing the nature and purposes of historical thinking in historical education research, questions must be asked not only about how students read and write about historical texts but also about how students become aware of how the past is used by themselves and those around them (Monte-Sano & Reisman, 2016).

When history is not just about "facts," there is a need for assessment instruments that reflect this broader understanding of historical competence and that can also be used in empirical studies with large sample sizes. Unfortunately, there is a lack of standardized measurement instruments that are focused on historical thinking skills (Ercikan & Seixas, 2011; Monte-Sano & Reisman, 2016). It is important to note that most of the available standardized historical competence tests (e.g., National Assessment of Educational Progress [NAEP]) have been criticized for a relative imbalance between the items that address facts versus historical thinking (Reich, 2009; VanSledright, 2014; Wineburg, 2004). Conversely, more "open" assessment formats seem to be valid for measuring historical thinking practices (Breakstone, 2013; also see Frost, de Pont, & Brailsford, 2012; Monte-Sano & De La Paz, 2012). One such example is the History Assessments of Thinking (HATs), which were developed in the context of the project "Beyond the Bubble" at the Stanford University and serve as formative assessments that measure students' ability to deal with primary sources (Breakstone, Smith, & Wineburg, 2013; Breakstone, Wineburg, & Smith, 2015; Smith & Breakstone, 2015). However, open-ended tasks are not feasible in studies with large samples.

Despite the lack of well-accepted standardized instruments that are able to cover historical competence in its full breadth, there have been successful attempts to measure some of its central dimensions. Of specific relevance for the present study, there has been an increasing interest in assessing insights into the epistemological principles of history—particularistic, selective, and laced with perspective—as a necessary underpinning of historical thinking (e.g., Havekes, Arno-Coppen, Luttenberg, & van Boxtel, 2012; Shemilt, 2000; Stoel, van Drie, & van Boxtel, 2016; Van Drie & van Boxtel, 2008; VanSledright, 2014). Historical thinking requires people to be aware of the nature of history, generate historical arguments on the basis of the available evidence, and evaluate the strength of such arguments (P. Lee & Shemilt, 2003; VanSledright, 2002; Wineburg, 2001). To measure epistemological principles of history, the Beliefs in History Questionnaire (BHQ; Maggioni, 2010) was developed in a series of studies (Maggioni, Alexander, & VanSledright, 2004; Maggioni, VanSledright, & Alexander, 2009). Three categories of epistemic beliefs were found via factor analyses—Objectivism, Subjectivism, and Criterialism. These categories formed reliable scales with Cronbach's alphas between .72 and .78 (Maggioni, 2010). According to VanSledright and Reddy (2014), these categories of epistemic beliefs are compatible with Kuhn and Weinstock's (2002) and King and Kitchener's

(2002) models of epistemic development as well as with the developmental trajectory of the concepts of the past, evidence, and historical accounts described in work by P. Lee and his colleagues (P. Lee, 2004; P. Lee & Ashby, 2000; P. Lee & Shemilt, 2003).

In addition, there have been attempts to measure whether students are able to deal with multiple documents by taking into account the differences between the sources (e.g., Rouet, Britt, Mason, & Perfetti, 1996). For this purpose, the students have to understand the difference between the past and history. Whereas the *past* encompasses all events that have occurred previously, *history* is the name we give to our efforts to interpret the past (VanSledright, 2014). Linked with this profound difference is the distinction between *(primary) sources*, which remain from the past, and *(historical) accounts*, which involve a telling of the past from a retrospective perspective. Prior attempts to measure basic historical concepts with standardized tests—in the context of the research on historical consciousness (Angvik & Borries, 1997; Trautwein et al., in press)—have not yet resulted in widely used standardized assessment instruments. But just recently, attempts have been made to capture the core concepts of historical thinking with standardized instruments, an approach that seems to be promising (Körber & Meyer-Hamme, 2015).

Oral History in School: Goals and Empirical Research

In U.S. schools, working with oral histories has become increasingly common (Lanman, 1987; Lanman & Wendling, 2006a). Similarly, working with eyewitnesses of the past is highly encouraged in the education standards of all 16 federal states in Germany (Henke-Bockschatz, 2014). The research literature concerning "oral history in school" mostly focuses on oral history projects conducted in history classes in the United States (e.g., Lanman & Wendling, 2006b; Peniston-Bird, 2009; Ritchie, 2011; Whitman, 2004) as well as in Germany (e.g., Henke-Bockschatz, 2014; Schreiber & Árkossy, 2009). In oral history projects, students are supposed to adopt the ways in which historians work by researching the historical context; conducting, transcribing, analyzing, and interpreting the interview; and presenting the results. In the oral history approach, working like a historian is believed to engage and inspire students and spark their interest in history (e.g., Boyd, Fernheimer, & Dixon, 2015; Lanman & Wendling, 2006a; C. R. Lee & Nasstrom, 1998; Peniston-Bird, 2009; Ritchie, 2011; Whitman, 2004, 2011).

Despite the popularity of the use of the oral history in school, only a limited number of empirical studies have explored the cognitive and motivational benefits and the potential pitfalls of this approach (Whitman, 2011). The study most similar to the present study (intervention group vs. control group at the classroom level, assessment of cognitive and motivational outcomes) was conducted by Lanman (1987). He observed differences between

the effects of the oral history approach (five classes) and traditional methods of instruction (five classes) in a sample of 210 students in U.S. history classes at a high school in Baltimore. The results of this study suggested that the increased affective attitude acquired through the oral history assisted cognitive performance, and the students noted an improvement in "their participation and involvement in history" (Lanman, 1987, p. 129) prompted by the oral history approach. Unfortunately, the measurement instruments and the teaching unit (e.g., Which kind of "active oral history approach" was used exactly?) were not described in detail. Apparently, the cognitive test addressed factual knowledge about immigration and Black history but not a broader understanding of historical competence as described previously.

To explore the effects of working with oral histories in history lessons, it is useful to distinguish between an "active oral history" (Lanman & Wendling, 2006b, p. xix)—talking directly with an eyewitness of the past—and a "passive oral history" (Lanman & Wendling, 2006b, p. xix)—integrating "ready-made" sources of oral history such as audio recordings, videotapes, or written interviews into history lessons. In our study, we examined the effectiveness of three ways of presenting oral history interviews that were prepared and evaluated in the context of a teaching unit: conducting a live interview with an eyewitness of the past compared with working with recorded oral histories such as videos or transcriptions.

Experiencing History in the Present: Opportunities and Challenges

Since the postwar renaissance of memory as a source for "people's history," much debate has ensued about the reliability of oral histories and the methodological challenges that arise when they are used to reconstruct the past. At the focus of such criticism was the assertion that memory is distorted by physical deterioration and nostalgia in old age, the personal biases of both the interviewer and interviewee, and the influence of collective and retrospective versions of the past (Thomson, 2015). However, for oral historians, the question is less what really happened but how the event was experienced and how it is remembered (A. Assmann, 2006; Plato & Burley, 2009; Portelli, 1981/1991). If one takes into consideration the idea that eyewitnesses' accounts switch between individual, social, and cultural memory, then oral history interviews are an important source for getting to know how historical events and periods are remembered in a specific social group (e.g., J. Assmann, 2010; Thomson, 2015). Analyzing oral history interviews in history lessons might therefore be suitable for fostering students' understanding of history as a reconstruction of the past. More than to simply promote students' historical thinking skills, the goal of the recently increasing use of the oral history approach in schools (e.g., Boyd et al., 2015; Goald & Gradowski, 2014) and in social life (e.g., Shopes, 2015) is to

motivate students by getting them to work like a historian (e.g., Lanman & Wendling, 2006a; Whitman, 2011), create an emotional involvement (Plato & Burley, 2009), or impress the students with the eyewitnesses' "aura of authenticity" (Sabrow, 2012, p. 27). Oral history use also tends to be connected with digital technologies and the new social phenomenon of storytelling (Freund, 2015).

From an educational perspective, we argue that three characteristics of live oral history interviews are of specific interest to researchers and have to be taken into account when gauging the advantages and liabilities of the oral history approach.

First, the characteristics of the live oral history interview—vividness and typically apparent honesty and sincerity—might create an "aura of authenticity" (Sabrow, 2012, p. 27), as illuminated by the "history lesson" with Hitler's successor, Karl Dönitz, described previously. Even more so, survivors or opponents of dictatorships from the past century (who are among the most sought after eyewitnesses in history education) are encircled by a specific aura, which stands in contrast to the peaceful situation in current Western democracies (Bernard-Donals & Glejzer, 2001). Thus, the live reports given by eyewitnesses of the past might make a very large impression on the recipients. One of the rare empirical studies exploring students' receptions of interviews with Holocaust survivors emphasized that students did not reflect critically on the witnesses' reports but took them as literal truth (Obens & Geißler-Jagodzinski, 2008). Another study about students' receptions of survivor interviews ($N = 779$) pointed out that most of the students were emotionally impressed with the eyewitness account, but they did not recognize the reconstructive character of the accounts (Galda, 2013). An international study that was based on a sample of 32,000 young people in Europe revealed that students place greater belief in oral testimonies than teachers' statements or history textbooks (Angvik & Borries, 1997). Therefore, it would be reasonable to speculate that students who encounter an eyewitness of the past might be more likely to take the report for granted than they would when working with a canned oral history interview (e.g., the video or the transcription).

Second, live oral history interviews get at the heart of what constitutes history. The account of a person who has witnessed the past might be helpful for understanding the distinction between the past and history as well as between primary sources and historical accounts. Students consistently confuse the past with history, often conflating the two, and tend to assume that primary sources and historical accounts map directly onto the past (Rouet et al., 1996; Stearns et al., 2000; VanSledright, 2014). However, the past is over, and our access to the past is mediated by traces, residue, and relics (i.e., primary sources), and it is based on such residue and relics that history (or more exactly histories; VanSledright, 2014, p. 27) is (are) told. Someone who witnessed the past could be considered a relic. But at the same time,

the eyewitness tells his or her story about the past from a current point of view, thus interpreting the past in a personal way (i.e., by providing a historical account). An encounter with a live eyewitness—combined with a discussion in the class about the ambiguity between a primary source and a historical account—might create an "aha effect," thus helping the students understand this important distinction.

Third, whereas many students lack an interest in history, by providing a link between the past and the present, oral history interviews might increase students' subjective relevance and interest. The direct communication offered by a face-to-face interview is considered to be one of the most personal and telling ways in which the past and present intersect (e.g., Lanman & Wendling, 2006b; Stricklin & Sharpless, 1988; Whitman, 2004), as shown by three empirical studies that explored students' reception of oral history accounts in school: two studies on students' reception of a Holocaust survivor's talk before a larger audience in the assembly hall (Galda, 2013; Obens & Geißler-Jagodzinski, 2008) and one study using oral history interviews on the topic of immigration and Black history in the United States (Lanman, 1987). On the basis of these studies, we can speculate that the students enjoyed interviewing witnesses of the past and that they embraced the oral history approach as valid and important for their own learning. Therefore, we can speculate that a real encounter with an eyewitness of the past enhances students' interest, at least temporarily.

Aims of the Current Study

Despite the great amount of public interest in and the increased awareness of history (see e.g., the aforementioned projects, Teaching American History and Beyond the Bubble), there is still a lack of sound experimental studies in classroom contexts with discipline-based conceptions of history and the goal of developing a better understanding of the conditions and environments that foster the learning of history (Maggioni, 2010; Monte-Sano & Reisman, 2016). The current study contributes to filling this gap by exploring the effects of working with oral history interviews in history lessons with the use of psychometrically sound instruments.

We scrutinized the effects of the oral history approach in three conditions (live, video, text) that were all embedded in an instructional learning environment and gauged the effect of the intervention with a comparison group that received their usual history lessons. According to the standards of randomized controlled field trials (e.g., Christensen, 2012; Schulz, Altman, & Moher, 2010), the control group could receive either an alternate intervention or no intervention (Shadish, Cook, & Campbell, 2001). We chose to use a nontreated control group, the results of which represented the baseline for the measurement of the dependent variable. We explored the following five hypotheses.

First, we examined the effectiveness of the—well-prepared and evaluated—oral history approach in active (live) and passive ways (video, text). We hypothesized that students in the three intervention groups would improve in their (a) insight into epistemological principles of history (by understanding that history is a reconstruction of the past, historical accounts have to be deconstructed, and oral history interviews offer specific opportunities and limitations), (b) understanding of basic history concepts (the distinction between primary sources and historical accounts), and (c) topic-related knowledge. Accordingly, we expected to find higher scores for the intervention groups compared with the control group on the tests measuring their historical competencies and their factual knowledge.

Second, we were interested in differential effects of the live group compared with the video and text groups with regard to the students' insight into epistemological principles of history (understanding reconstruction, understanding deconstruction, understanding oral history). Given the "aura of authenticity" (Sabrow, 2012, p. 27) of oral witnesses described previously, it seems possible that students in the live condition would not realize as well as the two other treatment groups that oral history accounts are reconstructions of the past that have to be evaluated carefully and critically just like any other historical account. We therefore expected that the live witness might have a less beneficial effect on students' insight into the epistemological principles than the video and text group interventions.

Third, we were interested in differential effects on students' comprehension of central historical concepts. The encounter with a live witness in the classroom clearly leads to the question of whether the account belongs to the present or the past: Should it be considered a primary source or a historical account? By contrast, the students in the video and text groups might have more problems grasping the distinction because they were working with a previously recorded or transcribed oral history.

Fourth, we also explored whether the three intervention conditions would be associated with differential effects with regard to students' topic-related factual knowledge, but we had no a priori expectations about the pattern of results.

Fifth, we examined differential effects on students' self-reported evaluation of the impact of oral history interviews on learning. In line with results from prior empirical studies (e.g., Lanman, 1987) and the didactic literature (e.g., Stricklin & Sharpless, 1988; Whitman, 2004), we expected that the students in the live group would report greater interest and provide a more favorable evaluation of the effectiveness of the oral history approach than students in the video and text groups.

Method

Our randomized controlled intervention study investigated the effects of using oral history interviews under three conditions (live, video, and text transcript) plus a (nontreated) control group in a pretest, posttest, follow-up test design.

Database and Sample

The current results were based on data from 35 ninth-grade school classes from 10 academic-track schools in Germany (N = 900 ninth-grade students, 46.7% female, M_{age} = 14.54, SD = 0.62). We acquired schools in the area next to our university. To qualify for participation, the schools had to have at least three history classes (preferably four) to allow for a within-school randomization. We contacted a total of 12 schools with a total of 51 classes. Of these, 10 schools agreed to participate, 5 of which had three classes and 5 of which had four, for a total of 35 classes. After the schools agreed to participate, in the schools with three classes, we randomly assigned the classes to the three intervention groups. In the schools with four classes, we randomly assigned the four classes to one of the three intervention conditions or the control group. In other words, a total of 10 classes took part in each of the intervention groups (live: 27.8% of the students; video: 28.4%; text: 29.2%), and there were five classes (14.6% of the students) in the control group.[1]

Teaching Unit

Overview

In our study, we focused on oral history interviews in active (live) and passive (video, text) ways that were embedded in a traditional instructor-directed learning environment. Thus, the setting of the three conditions was controlled—as much as possible in a real-life intervention—to avoid confounded effects. Therefore, the intervention unit was taught by the same teacher—the first author of the current article—and addressed the topic "Peaceful Revolution in the German Democratic Republic (GDR)." The intervention unit contained seven lessons (one lesson = 45 minutes) that were identical across the three interventions with the exception of two lessons (the fifth and sixth lessons) when the students worked with an oral history interview in three different ways: In 10 classrooms, students interviewed an eyewitness of the past; in another 10 classrooms, they analyzed a video interview; and in the last 10 classrooms, they worked with the transcript of the video interview. The pretest (65 minutes) took place immediately before the first lesson, and the posttest (90 minutes) was administered during the first regular history lesson after the intervention was completed.

Across the three conditions, the main goal of the teaching unit was to improve students' insight into epistemological principles regarding history (history as a reconstruction of the past, the need to deconstruct historical accounts, and the specific opportunities and limitations of oral history) and their understanding of the difference between primary sources and historical accounts. A detailed description of the teaching unit can be found in Appendix A in the online version of the journal. The three measurement points took place immediately before the intervention (pretest), immediately after the intervention (posttest), and 2 to 3 months later (follow-up test). The control group was taught various topics by the students' own teacher who used various teaching methods.[2] Using a nontreated control group (Shadish et al., 2001) allowed us to explore the total effect of the intervention and exclude the possibility that any learning progress made by the students in the intervention groups could be attributed to an effect of repeating the test (three measurements) or an effect of the usual cognitive development of ninth graders.

Study Context

Teacher and Rater

To control for the teacher's influence and improve the internal validity of the study, the same teacher taught the 30 intervention classes. Having the researcher teach the intervention classes meant that the teacher/researcher may have been tempted to influence the students' responses in the live classes. Another problem could be the teacher/researcher's exhaustion due to repeating the same unit in 30 classes. Therefore, every history lesson in the unit was observed by two raters who assessed the interaction between the class and the teacher. Using studies about teaching quality (Klieme, Pauli, & Reusser, 2009; Pianta & Hamre, 2009), we designed an observation protocol that was based on rating items that covered central aspects of classroom climate and instructional quality (six scales with a total of 33 items that assessed discipline, the supportiveness of the climate, use of time, students' cooperation, teacher's enthusiasm, and teacher's clarity). These ratings were used to determine whether there were any significant differences in instructional quality between the different intervention conditions.

Interviewees

A crucial factor in choosing the eyewitnesses was that they needed to have had similar experiences in the GDR to ensure that they would tell comparable stories about the past. Moreover, we needed the same eyewitnesses in the live, video, and text conditions. Therefore, we used the website www.jugendopposition.de, which offers videos and transcripts of interviews, lasting approximately 30 minutes, with members of the opposition

movement in the GDR. We also contacted four eyewitnesses and invited them to be interviewees in our live condition classes. Thus, all of the intervention conditions (live, video, text) involved these four interviewees who had been in their early 20s when the Wall fell. As students, they had struggled with the political system, contacted the opposition movement, and engaged in political activities and demonstrations against the regime near the end of the GDR. In the video and text classes, we used the videos and transcripts of the same four witnesses. The video and text interviews followed the same structure: The eyewitnesses told about their political socialization and their engagement and experiences during the Peaceful Revolution. Therefore, they mentioned the same points for the videos as they did in the live interviews. In the live interviews as well as in the videos and transcripts, the four eyewitnesses reported on the political system of the GDR from a critical point of view—based on similar experiences of being controlled and suppressed by the socialistic system. The interviewees were randomly allocated and balanced across the intervention conditions in order to minimize the impact of the specific personalities and experiences of the eyewitnesses on the intervention conditions.

Instruments

Outcomes

We used a total of eight outcome measures concerning factual knowledge (one test), historical competence (four tests), and students' evaluations of their learning success and interest (three scales) in our study. Students' knowledge of facts was captured at all three measurement points. We used fill-in-the-blank questions for which the students had to write the correct word in the blank. We assessed historical competence by administering three subtests that measured students' understanding of epistemological principles at all three measurement points. The subtests that captured their understanding of epistemological principles were inspired by the Beliefs in History Questionnaire (Maggioni, 2010; Maggioni et al., 2004, 2009) but were not exact copies of this instrument.[3] Furthermore, we assessed historical competence regarding students' understanding of basic history concepts (distinction between primary sources and historical accounts). Due to the time constraints—these items required students to carefully read three full historical documents—this test was administered only at t_2. Students in the three intervention groups evaluated their learning success and interest on the posttest and follow-up test. They evaluated how much the teaching unit with oral history interviews improved their learning about the subject of the GDR (content learning scale), enhanced their methodological learning (thinking history scale), and increased their interest (interest in oral history scale). Table 1 provides an overview of these outcome variables (achievement tests as well as students' self-reported ratings of the unit) and presents

Table 1

Description of the Outcome Variables, Reliability-Related Indices (Cronbach's α, Weighted Likelihood Estimators-Person Separation Reliability [WLE-PSR]), and Omnibus Tests for Pretest Group Mean Differences (Wald Test *p* Values)

Test	Explanation	Example	Items	Reliability t_1	Reliability t_2	Reliability t_3	*p* Values (Wald Test, t_1)
Test factual knowledge	Fill-in-the-blank questions divided into two parts at pretest and posttest; all items (14) are given on the follow-up test	After opening the border between _____ and Austria, a mass exodus from the GDR started.	7/7/14	Part A: α = .46 Part B: α = .62	Part A: α = .72 Part B: α = .73	Complete: α = .83	.696
Test understanding reconstruction	Understanding the difficulties of reconstructing the past out of primary sources and historical accounts	Historians have to take into account the likelihood that sources are incomplete.	24	WLE PSR = .87	See t_1	See t_1	.050
Test understanding deconstruction	Understanding that historical accounts have to be read critically	If you read something in a textbook, you can be sure that this was exactly what happened.	6	WLE PSR = .59	See t_1	See t_1	.029
Test understanding oral history	Understanding the specifics of oral testimonies, for example, the individual's perspective	Every eyewitness of history has his/her own view of the past.	10	WLE PSR = .73	See t_1	See t_1	.139
Test concepts of sources and accounts	Ability to distinguish between the concepts of primary sources and historical accounts on the basis of three historical documents	The **textbook excerpt** is a historical account because it is told from today's perspective, and the whole situation is presented. The **appeal** is a primary source because Prof. Kurt Masur spoke the text in this way so that the text is a relic of the GDR. The **eyewitness account** is a historical account because the text was written some years after the event. The **eyewitness account** is a primary source because the eyewitness' memory is a relic of the past.	23	Not tested	WLE PSR = .74	Not tested	Not tested at t_1

(continued)

Table 1 **(continued)**

Test	Explanation	Example	Items	Reliability t_1	Reliability t_2	Reliability t_3	*p* Values (Wald Test, t_1)
Content learning scale (student report)	Evaluation of the opportunity to learn topic-related knowledge through the oral history interview	In the oral history interview, I learned a lot about life in a dictatorial regime.	5	Not tested	$\alpha = .80$	$\alpha = .76$	Not tested at t_1
Thinking history scale (student report)	Evaluation of their methodical learning about the essentials of history itself	In the evaluation of the oral history interview, I learned about how historians work.	5	Not tested	$\alpha = .69$	$\alpha = .65$	Not tested at t_1
Interest in oral history scale (student report)	Evaluation of the motivational power of the oral history approach for themselves	Working with oral history interviews was exciting and interesting.	5	Not tested	$\alpha = .86$	$\alpha = .80$	Not tested at t_1

Note. The reported Cronbach's α and WLE-PSR coefficients are based on the total study sample, including students from three additional classes ("raw" data without multiple imputation: $N = 962$). Scales were based on data from $680 \leq N \leq 896$, excluding the factual knowledge test, which was administered at t_1 and t_2 as Parts A and B ($430 \leq N \leq 463$). The WLE scores for concepts of sources and accounts (assessed only at t_2) was based on data from 891 students. The remaining WLE scores were estimated on the basis of a data set with $2{,}643 \leq N \leq 2{,}649$ "virtual persons" for Measurement Points × Persons.

a brief explanation, an example item, the number of items, and the internal consistency at the three measurement points for each of the instruments. A more comprehensive account of all instruments is provided in Appendix B in the online version of the journal. Furthermore, all items from the three subtests for measuring students' understanding of epistemological principles are presented in Appendix C in the online version of the journal. The competence test capturing students' understanding of basic history concepts (i.e., the distinction between primary sources and historical accounts) is based on three historical documents (in the German language). This test and the fill-in-the-blank test for measuring students' factual knowledge can be obtained from the first author on request.

Covariates

Several variables including aptitudes, social background, and motivation were included as covariates in the regression analyses to control for pretreatment differences and improve the power of the statistical tests (Maxwell & Delaney, 2004). Table 2 presents the internal consistencies of the multi-item instruments that were used as covariates, and Table 3 provides an overview of all covariates that were included in the analyses as well as the results of a Wald test that was computed on the data at the first measurement point as a randomization check.

Methods of Data Analysis

Handling of Missing Data

There were relatively few missing pieces of data due to omitted answers. Across all variables (at the scale or score level) used in the analyses, on average, 9% (range, 7%-12%) of the responses were missing due to nonresponse or invalid responses.[4] Missing data were multiply imputed (10 imputations) with the software package Imputation and Variance Estimation Software (IVEWare; Raghunathan, Solenberger, & Hoewyk, 2002) from a data set with all dependent variables, predictors, and covariates across all measurement points.[5]

Scaling of the Competence Test

As explained in Note 1, we used the data from the total sample ($N = 962$) for scaling. For the competence tests, scores were estimated in separate unidimensional one-parameter (partial credit) item response theory models with the software ConQuest (Wu, Adams, Wilson, & Haldane, 2007) as weighted likelihood estimate (WLEs; Warm, 1989). Scores for the competence tests that were assessed at multiple measurement points (understanding reconstruction, understanding deconstruction, and understanding oral history) were estimated on the basis of a data set with "virtual persons"

Table 2

Internal Consistencies of the Scales Used for Adjustments (as Covariates in the Regression Analyses)

Test	Explanation	Example	Items	Cronbach's α	*M*	*SD*
Self-concept in history	Students' personal beliefs about their academic skills in history	"I am good at history."	4	.88	2.57	0.73
Interest in history	Personal interest in history	"I think history is not important for my future life (reversed)."	7	.86	2.43	0.65
Effort in history lessons	Willingness to work hard in history lessons	"I do my best in history lessons."	5	.72	2.80	0.59
Satisfaction with history lessons	Being happy about history lessons in school	"Time in history lessons just flies by."	3	.80	2.55	0.76
Role of history in the family	Importance of history in the family (books/films, talks, interest in history lessons)	"My parents are interested in what I am learning in history lessons."	4	.76	2.70	0.68
Interest in the German history from the 20th century	General interest in East and West Germany and the reunification of Germany	"I would like to learn more about the life of the people in the GDR."	5	.83	2.70	0.67
Talking about German historical topics	Talking with friends and family about specific historical topics of the 20th century regarding Germany	"In my family/with friends I talk about how Germany was divided."	7	.83	1.62	0.59
Trust in historical accounts	Trust in several historical accounts (*no trust, not much trust, some trust, great trust*)	"TV-documentary"	12	.75	2.98	0.36
Knowing topic-related terms (student report)	Knowing the meaning of topic-related terms (self-evaluation) (*I do not know, I once heard, I know roughly, I could explain well*)	"Class struggle"	7	.83	2.02	0.64

(continued)

Table 2 **(continued)**

Test	Explanation	Example	Items	Cronbach's α	*M*	*SD*
Working with eyewitness accounts	How to interview and to evaluate an eyewitness account	"If it is possible, the interviewer should take notes during the interview."	5	.71	0.39	0.33
Test motivation t_1	Effort in filling out the test at t_1	"I worked carefully on the test."	4	.73	2.87	0.58
Test motivation t_2	Effort in filling out the test at t_2	I worked carefully on the test."	4	.82	2.55	0.67
Test motivation t_3	Effort in filling out the test at t_3	I worked carefully on the test."	4	.80	2.59	0.69

Note. Scales were based on data from $680 \leq N \leq 896$. Unless otherwise indicated, the scale items were based on 4-point Likert-type response scales ranging from 1 (*does not apply at all*) to 4 (*fully applies*). Cronbach's alpha was based on the total study sample, including students from three additional classes ("raw" data without multiple imputation; $N = 962$). Unless otherwise stated, all scores referred to the measurement at t_1.

Table 3

Covariates (Observed Variables and Scales, see Table 2) Used in the Regression Analyses (Besides Pretest Scores) and Omnibus Tests for Group Mean Differences at t_1[a] (Wald Test *p* Values)

	Live Group		Video Group		Text Group		Control Group		Wald Test, t_1
Covariates	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*p*
Gender (dichotomous)	0.55	0.50	0.37	0.48	0.43	0.50	0.56	0.50	.012
Age	–0.01	1.03	–0.01	0.96	–0.06	0.97	0.16	1.06	.836
First language German (dichotomous)	0.89	0.31	0.96	0.19	0.93	0.26	0.89	0.31	.037
Birth place of parents in Germany (dichotomous)	0.88	0.33	0.93	0.25	0.89	0.31	0.85	0.36	.109
Birth place of parents in GDR (dichotomous)	0.11	0.31	0.08	0.28	0.06	0.24	0.05	0.22	.096
Talking with eyewitnesses about GDR	0.04	1.03	–0.05	0.97	–0.02	1.02	0.07	0.95	.634
Books at home	–0.13	1.06	0.00	1.01	0.05	0.98	0.14	0.87	.171
Role of history in the family	0.00	0.97	0.02	1.02	–0.07	0.98	0.10	1.06	.770
Cognitive ability	–0.02	0.95	0.06	0.96	0.03	1.02	–0.13	1.11	.773
Capacity for reading	–0.08	0.94	–0.12	0.92	0.19	1.09	0.00	1.02	.396
Self-concept in history	0.01	0.98	–0.05	0.96	–0.04	1.00	0.15	1.11	.377
Interest in history	0.04	0.99	–0.03	0.97	–0.08	0.98	0.14	1.10	.581
Effort in history lessons	0.10	0.95	–0.05	1.00	–0.08	1.00	0.06	1.08	.359
Satisfaction with history lessons	–0.04	0.90	0.04	0.96	–0.13	1.04	0.26	1.13	.191
Interest in German history from the 20th century	0.14	0.99	–0.06	0.99	–0.10	0.94	0.06	1.12	.163
Talking about German historical topics	0.00	0.97	0.00	0.98	–0.01	1.02	0.03	1.06	.995
Trust in historical accounts	–0.03	0.98	0.09	0.84	–0.05	1.11	–0.01	1.08	.570
Grade in history	0.00	0.95	0.06	1.01	–0.03	1.01	–0.05	1.08	.824
Grade in German	0.00	0.97	0.16	1.05	0.00	0.98	–0.31	0.92	.001
Grade in mathematics	–0.02	0.99	0.01	1.02	–0.03	0.99	0.09	1.00	.852
History as favorite subject (dichotomous)	0.09	0.28	0.07	0.25	0.05	0.22	0.14	0.34	.004

(continued)

Table 3 **(continued)**

Covariates	Live Group		Video Group		Text Group		Control Group		Wald Test, t_1
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*p*
Previous knowledge (one item, student report): 1 = *not at all* to 4 = *very much*	–0.02	0.92	0.07	1.04	–0.10	0.96	0.10	1.14	.050
Knowledge of topic-related terms, for example, class struggle (student report)	–0.07	0.91	0.00	0.93	–0.08	1.05	0.28	1.14	.048
Knowledge about the leading party SED (one item, test)	–0.01	0.98	–0.01	0.96	0.03	1.05	0.00	1.02	.995
Working with eyewitness accounts (five items, test)	–0.06	1.00	0.17	1.01	–0.04	1.03	–0.14	0.89	.068
Test motivation t_1	0.04	0.88	0.11	0.92	–0.08	1.06	–0.13	1.20	.636
Test motivation t_2	0.11	0.96	0.03	0.94	–0.13	1.04	0.02	1.09	.683
Test motivation t_3	0.04	0.97	–0.02	0.99	–0.06	1.03	0.10	1.02	.796

Note. Continuous variables were *z* standardized on the subsample used in this study ($N = 900$) with multiply imputed missing values. The group means for dichotomous (dummy-coded) variables refer to unstandardized values and may be interpreted as proportions.
[a]All variables refer to the first measurement point with the exception of test motivation, which also refers to the second and third measurement points.

for Measurement Points × Persons (i.e., each student was represented three times in this data set with the respective response vector for each measurement point). The weighted mean square statistics (WMNSQ) indicated acceptable item fit for all indicators (understanding reconstruction: $0.82 \leq$ WMSNQ ≤ 1.34; understanding deconstruction: $0.89 \leq$ WMSNQ ≤ 1.24; understanding oral history: $0.88 \leq$ WMSNQ ≤ 1.29; concepts of sources and accounts: $0.86 \leq$ WMNSQ ≤ 1.12), and the variances of the latent dimensions were reasonably large, ranging from 0.86 to 0.88, except for concepts of sources and accounts, which had a somewhat smaller variance (0.34). Table 1 shows the WLE-Personal Separation Reliability (PSR) of the four competence tests.

Standardization of Dependent Variables

All variables that were measured at each of the three measurement points (understanding reconstruction, understanding deconstruction, understanding oral history'') were *z* standardized with the mean and standard deviation of the respective (multiply imputed) scores of the first measurement point by subtracting the mean from each score and dividing by the standard deviation of the first measurement point for the total sample. This allowed for an interpretation of mean differences between the timepoints or groups relative to the respective standard deviation of the pretest. Dependent variables that were measured at t_2 and t_3 (only t_2: concepts of sources and accounts; t_2 and t_3: content learning [student report], thinking history [student report], interest in the oral history [student report]) were *z* standardized at each measurement point.

Inferential Statistical Analyses/Nested Structure

Statistical inferences were based on procedures that took into account the sampling design (with school classes as clusters) and the multiple imputation. All analyses were computed in SAS (Version 13.1; SAS Institute Inc., 2013) with PROC SURVEYREG and PROC MIANALYZE. In a first step, regressions using a generalized least squares approach with Taylor linearization were estimated with PROC SURVEYREG. PROC MIANALYZE was then used to combine the results from the different imputed data sets using adjusted degrees of freedom as proposed by Barnard and Rubin (1999). The combination of these two steps ensured that reliable statistical significance testing would be applied in the present study. Students were nested within classes in the present study, a condition that had the potential to bias the standard errors of the resulting estimates if this clustering had not been taken into account. In order to estimate the correct standard errors, we used the Taylor linearization, which is implemented in the SAS package and provides correct standard errors when data from a cluster sampling design are used.

Results

Descriptive Statistics, Randomization Check, Treatment Check

Means and Standard Deviations

Table 4 shows the means and standard deviations of the four groups at the three measurement points.

Randomization Check

For each of the 30 variables (5 pretest variables, 25 additional covariates) assessed at the first measurement point, we tested the hypothesis that the means were equal across the four groups by computing Wald tests (TEST option in PROC SURVEYREG) (see Tables 1 and 3, last column). The results of the Wald tests are omnibus tests whose interpretation is comparable to the results of a unifactorial ANOVA with the factor group membership (live, video, text, or control group), but the analysis also takes into account the sampling design (cluster sampling). The Wald test p values here reflect the probability of the given or more extreme group mean differences for a specific variable under the null hypothesis that all group means are equal (which is a reasonable assumption only in the case of randomization). With increasing numbers of variables that are tested for group differences, the chance of detecting small p values that can be interpreted as statistically significant group differences automatically increases (covariate imbalance), and the number of these significant group differences also depends on the correlation of the group differences with different variables (e.g., higher average achievement levels of students may go along with higher average self-concept levels). In the present study, statistically significant group mean differences were found for two pretest variables (Table 1; understanding reconstruction, understanding deconstruction) and six of the student background variables used as covariates in the analyses (Table 3; gender, German language, grade in German, history as one's favorite subject, previous knowledge, knowledge of topic-related terms). There was, however no clear pattern of group mean differences. The video group, for instance, descriptively showed the highest mean for grade in German (which indicates a lower performance compared with the other groups) but also the highest mean for the two pretest variables (understanding reconstruction, understanding deconstruction; indicating high performance). Because all of the variables from the pretest were included in the multiple regression analysis as covariates, the differences on the pretest should not be relevant for the interpretation of our results.[6]

Treatment Check

The intervention in the three groups was conceptualized to be identical, with the exception of the two lessons in which three different approaches to

Table 4
Group Means and Standard Deviations for the Outcomes at All Measurement Points

		Pretest		Posttest		Follow-Up Test	
Test	Group	*M*	*SD*	*M*	*SD*	*M*	*SD*
Test factual knowledge	Live	–0.08	0.99	1.63	1.60	1.02	1.62
	Video	–0.07	0.92	1.86	1.57	1.12	1.50
	Text	0.09	1.04	1.99	1.63	1.31	1.68
	Control	0.09	1.08	0.48	1.35	0.18	1.40
Test understanding reconstruction	Live	–0.10	0.88	–0.08	1.20	0.04	1.49
	Video	0.17	1.02	0.18	1.43	0.11	1.45
	Text	0.00	1.00	0.03	1.33	0.01	1.50
	Control	–0.14	1.12	–0.31	1.37	–0.24	1.44
Test understanding deconstruction	Live	–0.10	0.99	–0.13	1.08	–0.02	1.13
	Video	0.20	1.01	0.18	1.17	0.18	1.09
	Text	–0.06	1.00	0.11	1.17	0.05	1.19
	Control	–0.08	0.96	0.02	1.09	–0.01	1.15
Test understanding oral history	Live	–0.16	1.00	0.00	1.18	0.14	1.32
	Video	0.02	0.93	0.31	1.24	0.38	1.32
	Text	0.11	1.03	0.31	1.34	0.36	1.31
	Control	0.04	1.03	–0.08	1.08	–0.11	1.18
Test concepts of sources and accounts	Live	Not tested at t_1		0.11	1.03	Not tested at t_3	
	Video			0.04	1.02		
	Text			–0.05	1.00		
	Control			–0.18	0.88		
Content learning scale (student report)	Live	Not tested at t_1		0.53	0.82	0.32	0.97
	Video			–0.16	0.95	–0.03	0.99
	Text			–0.35	1.00	–0.27	0.95
Thinking history scale (student report)	Live	Not tested at t_1		0.30	0.89	0.19	0.98
	Video			–0.04	0.98	0.01	1.05
	Text			–0.25	1.04	–0.19	0.95
Interest in oral history scale (student report)	Live	Not tested at t_1		0.72	0.78	0.51	0.97
	Video			–0.20	0.88	–0.10	0.91
	Text			–0.49	0.91	–0.38	0.91

Note. $N = 900$. All variables that were measured at each of the three measurement points were standardized by using the mean and standard deviation of the respective (multiply imputed) scores from the first measurement point. Dependent variables that were measured at t_2 and t_3 only were z standardized at each measurement point.

oral history accounts were used. To probe for unintended differences in teaching across the three intervention conditions, the instructional quality in all intervention groups was assessed by trained observers (in every unit lesson) with a total of five scales covering central aspects of classroom

climate and teaching quality (Klieme et al., 2009; Pianta & Hamre, 2009). Interrater agreement between independent observers was assessed by computing the average deviation index (Burke & Dunlap, 2002; Burke, Finkelstein, & Dusig, 1999) for which low values indicate high agreement. Raters' average deviation ranged from 0.13 to 0.30 on a 4-point scale and was thus clearly below the cutoff value of 0.67 (Lüdtke, Trautwein, Kunter, & Baumert, 2006). We did not find any statistically significant differences in the average ratings, indicating that instructional quality was similar across all intervention groups.

Correlation Matrix

Table 5 shows the correlations for the achievement outcomes. They were based on linear regression models with robust standard error estimation (PROC SURVEYREG and PROC MIANALYZE for cluster sampling and multiple imputations) using z standardized variables. To do this, the data from the posttest (N = 900) were used because all achievement-related variables were included at this timepoint—including the time-consuming task of measuring students' conceptual understanding of primary sources and historical accounts. The correlation matrix showed that nearly all outcomes were statistically significantly correlated with each other. At the same time, the correlations were mostly modest in size (all rs $< .66$), indicating that the five outcomes represented clearly separable constructs.

Overall Effectiveness of the Intervention Compared With the Control Condition

Our first hypothesis referred to the difference between the three intervention groups and the control group. We expected higher scores on both the posttest and the follow-up test in the intervention groups. To test this hypothesis, we dummy coded group membership (0 = control group, 1 = intervention group) and ran a set of five regression analyses for the five outcomes. Adjusted mean differences between the three oral history intervention groups and the control group were based on a multiple regression analysis with the dummy-coded variable for group membership and covariates (students' achievement on the pretest and students' sociocultural background, grades, interest in history and the topic, and test motivation). Due to the standardization of the dependent variables (see the section "Standardization of Dependent Variables"), the estimates reported in column "b" in Table 6 are to be interpreted as adjusted effect sizes of the intervention concerning the achievement outcomes. For instance, b = 0.31 for understanding reconstruction means that the difference between the intervention and control groups—adjusted for (preintervention) group differences—was 0.31 standard deviations (in the metric of the respective pretest scores). The reported p values refer to two-sided tests.

Table 5
Correlations Between the Five Achievement Tests at t_2

	Test Factual Knowledge	Test Understanding Reconstruction	Test Understanding Deconstruction	Test Understanding Oral History
Test understanding reconstruction	.39			
Test understanding deconstruction	.22	.47		
Test understanding oral history	.28	.65	.45	
Test concepts of sources and accounts	.28	.38	.24	.30

Note. Correlations were estimated as regression weights based on *z* standardized variables at t_2 with robust standard errors (cluster sampling, multiple imputation). All coefficients are statistically significant ($p < .001$). $N = 900$.

As predicted in the first hypothesis, the results demonstrated that the students in the intervention groups showed statistically significantly higher historical competence and more factual knowledge than the control group at both measurement points (see Table 6), the only exception being the understanding deconstruction subtest. The largest effects at both measurement points were found for factual knowledge (up to 1.46 *SDs*). This large effect was due to the fact that the control group classes attended their usual history lessons, which were taught without addressing the topic "Peaceful Revolution." In sum, we were able to conclude that the intervention involving the oral history approach was successful at improving students' historical competence and their topic-related knowledge across all intervention groups.

Comparison of Live Presentations Versus Video/Text

We next turned to the analyses of differences across the three intervention groups. Because we were most interested in comparing the live group with the video and text groups, we computed adjusted mean differences by applying contrast coding. Estimates of differences between the intervention groups were based on the intervention subsample ($N = 769$) with contrast-coded variables for the comparison of the live versus both other intervention groups (video and text conditions) as well as the comparison of those two groups (video vs. text condition). Again, we included all of the covariates described previously. No statistically significant differences were found between the video and text groups for students' achievement on the competence and factual knowledge tests and in the student reports on the

Table 6
Intervention Groups Versus Control Group Comparison, Based on Adjusted Effects at t_2 and t_3

	Intervention Groups Versus Control Group (t_1 to t_2)		Intervention Groups Versus Control Group (t_1 to t_3)	
Tests	*b*	*p* (Two-Sided)	*b*	*p* (Two-Sided)
Test factual knowledge	1.46	<.001	1.14	<.001
Test understanding reconstruction	0.31	.005	0.35	.008
Test understanding deconstruction	0.01	.879	0.03	.772
Test understanding oral history	0.30	<.001	0.42	<.001
Test concepts of sources and accounts	0.19	.020	Not included at t_3	

Note. Intervention groups versus control group (dummy coding: 0 = control group, 1 = intervention). N = 900. The multiple regression analyses included robust standard error estimation (cluster sampling, multiple imputation) and covariates: achievement at t_1 on the competence tests and self-evaluation of knowing topic-related terms; cognitive and reading ability; socioeconomic status and family's interest in history; grades in math, German, history lessons; history as one's favorite topic in school; self-concept; effort and interest in history lessons; general interest in history and the topic of the GDR; self-evaluation of previous topic-related knowledge; talking with eyewitnesses of the GDR outside of school; and effort in completing the test at all three measurement points.

effectiveness of the oral history approach. Therefore, Table 7 presents only the results for the differences between the live group and the video/text group.

Intervention Effects on Achievement Indicators

We computed a set of regression analyses to test for the effects of the intervention on the five outcome variables on the posttest and follow-up test, again controlling for the pretest and the set of covariates. As reported in Table 7, on the posttest, students in the live group had statistically significantly lower scores on the subtests understanding oral history and understanding deconstruction. This result confirmed our second hypothesis, which predicted that the students in the live condition would score lower on the tests concerning their insight into the specifics of oral testimonies as well as in the need to deconstruct narratives. The differences between the interventions groups were no longer statistically significant on the follow-up test. Regarding the understanding reconstruction test, no

Table 7

Live Group Versus Video/Text Group Comparison, Based on Adjusted Effects at t_2 and t_3

	Live Group Versus Video/Text Group (t_1 to t_2)		Live Group Versus Video/Text Group (t_1 to t_3)	
Tests	*b*	*p* (Two-Sided)	*b*	*p* (Two-Sided)
Test factual knowledge	–0.29	.098	–0.11	.498
Test understanding Reconstruction	–0.11	.265	0.02	.883
Test understanding deconstruction	–0.19	.049	–0.05	.552
Test understanding oral history	–0.23	.022	–0.13	.208
Test concepts of sources and accounts	0.16	.021	Not included at t_3	
Content learning scale (student report)	0.74	<.001	0.43	<.001
Thinking history scale (student report)	0.38	<.001	0.23	.016
Interest in oral history scale (student report)	0.99	<.001	0.67	<.001

Note. Live group versus video/text group (contrast coding), $N = 769$. The multiple regression analyses included robust standard error estimation (cluster sampling, multiple imputation) and covariates: achievement at t_1 on the competence tests and self-evaluation of knowledge of topic-related terms; cognitive and reading ability; socioeconomic status and family's interest in history; grades in math, German, history lessons; history as one's favorite topic in school; self-concept; effort and interest in history lessons; general interest in history and the topic of the GDR; self-evaluation of previous topic-related knowledge; talking with eyewitnesses of the GDR outside of school; effort in completing the test at all three measurement points.

statistically significant difference between the live group and the video/text group was found on either the posttest or the follow-up test.

Students' understanding of central historical concepts (concepts of sources and accounts) was assessed only on the posttest. In line with our third hypothesis, again controlling for the other covariates, the students in the live group scored statistically significantly higher on this test than the students in the video/text group.

With regard to the fourth hypothesis, which examined differences with respect to students' factual knowledge, no statistically significant difference between the live group and the video/text group was found on either the posttest or the follow-up test.

Altogether, we found a pattern of differential intervention effects on the achievement outcomes. The results point to a better understanding of the concepts of sources and accounts in the live oral witness group, whereas statistically significantly higher scores in the video/text group were found for two out of three scales from the epistemological beliefs in history scales. Thus, there was some evidence that students in the "live" condition learned less than students in the other two conditions at least on some subtests. However, statistically significant group differences were found only on the posttest but not on the follow-up test.

Students' Evaluation of Learning Success and Student Interest

Marked differences in students' evaluations of the teaching unit with oral history interviews were observed at both measurement points after the intervention. In line with our expectations in the fifth hypothesis, students in the live group compared with those in the video and text groups considered the learning arrangement to be statistically significantly more helpful in helping them to understand important facts about the history of the GDR (content learning scale [student report]), in supporting their understanding of the basics of history (thinking history scale [student report]), and in offering an approach that is more alive and interesting (interest in oral history scale [student report]). These differences were also observed at a similarly high level on the follow-up test, indicating that students who worked with the eyewitness had "more fun" and experienced greater subjective learning success.

Discussion

The results from our cluster randomized controlled trial with which we examined intervention effects in history lessons confirmed both the opportunities and dangers that have been associated with the presentation of live oral history witnesses in high school classrooms. (a) Comparing the three intervention groups with the control group, the intervention groups scored

better than the control group on four of the five achievement tests. (b) Comparing the live group with the video and text groups, the students in the live condition scored lower on two tests (understanding deconstruction, understanding oral history), which measured their understanding that narratives, even those of eyewitnesses, have to be deconstructed carefully. (c) However, the students in the live condition understood better the difference between sources and accounts. This test was based on the students' assessments of whether the reasons for classifying three documents (one of these was an oral history account) as primary sources or as historical accounts were correct or not. (d) Whereas the control group showed statistically significantly lower student achievement on the factual test than the intervention groups did, no differences between the intervention groups were found on this test. The students in the live condition learned the same in terms of the facts and figures as those in the video and text conditions. (e) The students in the live condition evaluated the learning and motivational potential of the oral history approach considerably higher than the students in the video and text conditions.

Comparing the intervention groups with the control group and regarding the goal of the teaching unit, which was to foster students' historical thinking, the oral history intervention worked. The results on the follow-up test are very clear: The learning benefits of the oral history approach compared with the nontreated control group persisted from the posttest to the follow-up test. Looking at the differential effects between the intervention groups, we see that in line with arguments that have been made for using oral witness approaches in classrooms (e.g., official recommendations such as the Standing Conference, 2009), the live group reported markedly higher scores on interest and subjective learning than students in the alternative conditions (video/text). However, with regard to more objective indicators of learning, and confirming concerns about using the live oral history approach in regular classrooms (e.g., Henke-Bockschatz, 2014; Sabrow, 2012), the text and video groups outperformed the live oral witness group on two dimensions involving epistemological principles, whereas the live group showed higher scores on only one indicator.

Live Oral Witnesses: Both Helpful and Potentially Dangerous

It is well known that there are several challenges involved in measuring historical thinking skills over and above factual knowledge in large-scale assessments (e.g., Ercikan & Seixas, 2015; VanSledright, 2014) and developing interventions to foster students' historical thinking (e.g., Maggioni et al., 2004; Mandell, 2008). Regarding the oral history approach, both the advantages (e.g., Lanman, 1987; Lanman & Wendling, 2006b; Ritchie, 2011; Whitman, 2004) and the potential pitfalls (e.g., Plato & Burley, 2009; Portelli, 1981/1991; Sabrow, 2012) have frequently been discussed over

the past several decades, but there has been a surprising lack of empirical studies (for rare exceptions, see Galda, 2013; Lanman, 1987; Obens & Geißler-Jagodzinski, 2008) that have put the contradictory claims to a test.

In line with previous findings and our first hypothesis, students in all of the intervention conditions performed better than the control group on tests measuring their understanding of the reconstructive character of history, specifics of oral history, and central historical concepts. It is interesting to note that they did not score better than the control group on the test that assessed their understanding of deconstruction, which was partly due to the fact that the live group scored lower than the video and text groups on this subtest. Across all of the intervention groups, an increase in students' factual knowledge was evident on both the posttest and the follow-up test. Unfortunately, we could not control for many variables in the control group, the most important of which were the teacher and the topic. The comparison with the control group has to be regarded as a comparison between different intervention groups that involved a high degree of control with a control group that was not treated in order to confirm that the learning processes of the intervention groups were not a result of the usual development of youth of this age (e.g., with regard to their epistemic cognition in history). With respect to the first research question comparing the control group and the three intervention groups, we were thus able to conclude that the intervention worked: The students in the three intervention groups learned something new about the epistemological principles of history that they do not get in everyday life or in their "usual" history lessons. Taken together, on both the posttest and the follow-up test, the oral history intervention seemed to improve students' historical thinking competence and factual knowledge across all conditions.

Regarding the risks of conducting an interview live in the classroom, the "aura of authenticity" (Sabrow, 2012, p. 27) of an eyewitness of the past might be of particular concern. In the intervention unit, which was administered in an identical fashion in all intervention groups—except for the lesson in which the (live) interview took place—we wanted to demonstrate the need to deconstruct historical accounts and oral testimonies. Thus, the limitations of eyewitness accounts were explicitly discussed in the three intervention groups. Even so, as addressed by our second hypothesis, the students in the live group scored statistically significantly lower on the posttest than the students in the video and text groups on the subtests understanding oral history and understanding deconstruction (see Table 7). At least in the central parts of the outcomes, the students in the live condition learned less than the students in the video and text conditions. In view of the fact that the limitations of oral history accounts were also discussed explicitly in the live condition, the lower scores of the students in the live group on the posttest are remarkable. It is possible that the lower scores on these tests can be explained by the theoretically expected consequences of the "aura of

authenticity" (Sabrow, 2012, p. 27). The live eyewitnesses' authenticity and credibility might have made it difficult for students in the live group to maintain their distance from oral testimonies and historical accounts in general. However, the statistically significant differences concerning the tests understanding oral history and understanding deconstruction disappeared on the follow-up test. Further research is necessary to investigate the potential "dangers" of the live eyewitness account regarding students' ability to maintain their distance from such accounts.

Understanding the distinction between primary sources and historical accounts is a necessary condition for working like a historian (e.g., Rouet et al., 1996; VanSledright, 2014). In line with the third hypothesis, the students in the live group scored statistically significantly higher than the students in the video and text groups on the concepts of sources and accounts subtest (see Table 7). The hands-on experience with an interviewee speaking today about his or her experiences in the past might have made it easier for the students in the live group to understand the difference between the past and history and—moreover—between primary sources and historical accounts. As we explained before, this test was based on three documents—one of which was an oral history account (see Table 1 for some example items). The students in the live condition seemed to grasp the specifics of the oral history account, which is both a primary source and a historical account, more easily because the eyewitness told his or her story about the past in the real present. By contrast, for the students in the video and text groups, grasping the distinction between a primary source and a historical account may have been more difficult because they worked with a previously recorded interview.

In line with previous empirical studies (Galda, 2013; Lanman, 1987; Obens & Geißler-Jagodzinski, 2008), the students in the live condition were profoundly convinced of their learning progress, but they scored lower than the video and text groups on two subtests regarding epistemological principles (see Table 7), which was the main goal of the lesson unit. On the one hand, this might be due to the fact that speaking directly with an eyewitness differs more from the usual teaching methods that are usually used in history lessons than working with a video or a text does. Moreover, the reason for this finding might once again be related to the "aura of authenticity" (Sabrow, 2012, p. 27). It seems possible that the students were so impressed by the direct personal communication of an eyewitness of the past that they were convinced of their learning progress. But it is precisely this emotional experience that might dissolve the cognitive distance from the oral history account that constitutes a cornerstone of historical reasoning.

How can the opportunities of live oral witnesses be used while simultaneously overcoming the potential pitfalls of this encounter? The motivational potential of the oral history approach should be used in history classes. In

line with previous research, the literature, and the experience of history teachers, working with a live oral history offers great possibilities for engaging students in history lessons. Moreover, the oral history approach—embedded in a competence-oriented unit lesson—could be used to foster students' insights into their understanding of central terms such as *sources* and *accounts* and improve their insight into epistemological principles of history.

However, teachers have to be aware of the risks involved in using the oral history approach regarding eyewitnesses' credibility for the students and should turn the eyewitnesses' account into a topic of further discussion in the classroom. The live eyewitness seems to be very impressive so that the students have difficulties maintaining their distance, which is necessary for analyzing and interpreting primary sources as well as historical accounts. Inviting two eyewitnesses to the classroom (together or successively) and comparing their accounts directly would be preferable, but this arrangement is usually not possible in real-life classrooms because of instruction time, a lack of "good" eyewitnesses, and resource restrictions. Therefore, we propose that the live interview be compared with additional textual accounts. Taking a critical look at the live witness account is especially important. If the live account is analyzed like every other source by using sourcing as well as contextualization and corroboration heuristics, there is a better chance to avoid developing an understanding of the oral history interview that lacks complexity (Schreiber & Árkossy, 2009). It should be reiterated however that despite using such elements in the last lesson of the teaching unit, we still found differences between the live and video/text conditions in our study.

Beyond Narratives of Successful Projects: Conducting a Randomized Controlled Field Trial

In previous studies, researchers have engaged in small-scale classroom interventions that focused on improving historical thinking (e.g., Swan, Hofer, & Locascio, 2008; VanSledright, 2002). Although the authors carefully considered students' thinking, the generalizability of these studies is limited by the small sample, lack of a control condition, and absence of baseline measures of historical thinking (Monte-Sano & Reisman, 2016). To the best of our knowledge, our study was the first cluster randomized controlled intervention study—not only on the effects of the oral history approach but moreover in probing for the effectiveness of history lessons—to use a large sample and a repeated-measures design and control for the clustered design of the data set.

One reason for the lack of empirical studies in large-scale formats with regard to history lessons is the lack of standardized assessment instruments for measuring the baseline and learning progress of the students' historical

thinking competences. Therefore, as part of our study, we developed psychometrically sound instruments for measuring several different outcomes, including students' historical thinking competence, factual knowledge, and evaluation of the oral history approach, in order to obtain reliable quantitative results regarding the intervention's outcomes. In the past few years, there have been some promising attempts to capture core concepts of historical thinking with standardized instruments (Körber & Meyer-Hamme, 2015; Trautwein et al., in press), which will help to evaluate the effectiveness of history lessons in large-scale assessments in the future. The current study shows one way to combine the expertise of educational psychology and a history lesson didactic to design, conduct, and evaluate an intervention study with the goals of improving historical competences and the teaching of such competences. Such goals seem important in a postmodern globalized world.

Limitations and Outlook

The present study has several strengths, including a research question that is of both theoretical and practical relevance, an adequate research design, the multidimensional assessment of intervention effects, and the use of elaborated statistical analysis methods. At the same time, a number of limitations and the need for further studies must be mentioned.

First, although we administered a relatively broad set of outcome variables, additional outcome variables should be used in future studies. For instance, it would be desirable to determine the effects of the oral history approach on additional historical competencies such as the ability to reconstruct historical accounts out of primary and secondary sources and the ability to deconstruct given historical accounts, such as those required in the "Historical Thinking Standards" (National Center for History in the Schoools UCLA, 1996). Are students able to create their own narratives that are based on the thoughtful reading of sources and accounts? However, this would be possible only if students' results on time-consuming open-ended tasks such as short interpretive essays were used (cf. VanSledright, 2014).

Second, to ensure the study's internal validity, all intervention classes were taught by the same teacher. This design, which uses an expert teacher, is well suited for controlling for the well-known differences in teaching competences across teachers (e.g., Hattie, 2009). Therefore, we decided to emphasize internal validity in the first study, but much care was taken to produce a teaching unit that could be directly incorporated into the normal curriculum with regard to the underlying goals of the learning arrangement, materials covered and needed by the teacher, and time that was needed to cover the materials. Moreover, we used readily available videos and transcripts of eyewitness accounts (available on the website: www.jugendopposition.de).

Finally, the complete intervention materials are available in a fully documented and scripted form, which allows for easy dissemination. Hence, there is reason to believe that the intervention can very easily be implemented in real-life classrooms. However, more studies are needed to evaluate the stability of these effects across different teachers (see Borman, Gamoran, & Bowdon, 2008).

Third, in the current study, we worked with an "untreated" control group because we wanted to ensure that we would be able to (a) identify effects of the interventions relative to no treatment and (b) gauge the effects by controlling for test-retest effects (which should also show up in the control condition). However, a control group taught about the same topic in the "usual way" would be desirable. Therefore, in upcoming studies, we (a) plan to explore the intervention in "real" conditions with trained teachers teaching a unit that is based on the complete intervention materials and (b) will include a "treatment as usual" group. In such a study, teachers in the intervention groups (working with oral history in active and passive ways) will be given the training immediately, and teachers in the control group will be given the training in the following school year. This design will allow us to explore (a) whether the intervention could be implemented and replicated in given "real" conditions as well as (b) differential effects between the oral history intervention conditions compared with the effects of the usual history lessons on the same topic.

Fourth, as is typical for all studies on the education of history, the content that is taught cannot be easily exchanged for other content. History education is bound to specific content, and this content varies across times and places. In Germany, the topic "Peaceful Revolution in the GDR" is fixed in the curricula. Therefore, speaking with an opponent of the GDR system is customary in Germany. By studying the effect of oral histories that address only one topic, it is not possible to conclude that oral histories of other topics would have similar effects. The effects of working with oral histories could be examined with regard to U.S. history (e.g., participants of the civil rights movement would be suitable for interviewing in the classroom). But besides these obligatory politically related topics, "history from below" (Thompson, 1978) would be important for history classes. Just as the Foxfire project focused on preserving the heritage of their region for nearly 50 years (Stein & Scatena, 2006), eyewitnesses of the past could be invited to classes to talk about the past of the local region. Although we believe that the general characteristics of oral witness accounts can be generalized across situations, there is a clear need for more studies with different content and a larger number of oral witnesses.

Fifth, the sample of ninth-grade school classes from the highest school track in Germany can be viewed as a limitation of the study. Regarding the problem of maintaining a distance from the living interviewee, the same or even stronger effects could be assumed for younger students or

students from a lower school track. Conversely, older students or adults might find it easier to keep their distance from the live witness because their epistemic beliefs are likely to be more developed (cf. P. Lee & Ashby, 2000). Thus, in future studies, the effectiveness of using oral history interviews with younger students or students who attend a lower school track should be explored. Because of the popularity of oral testimonies offered at memorial sites (A. Assmann, 2006), it would be desirable to examine older students' or adults' receptions of live oral history accounts.

Finally, it is likely for an interviewee's personality to have a marked impact on student outcomes. In our study, the power to examine such effects was restricted because we used a total of four different eyewitnesses. However, it can be assumed that witness-specific experiences play an important role in students' reception of an oral history interview. Charismatic eyewitnesses may make it even harder for the students to maintain their distance from the account. Hence, interviewing a charismatic eyewitness of the past would be desirable, but their persuasiveness should be taken into account—and might underscore the necessity for additional empirical studies.

Conclusion

The use of oral history interviews—live, video, or text—embedded in a learning environment that was geared toward promoting historical thinking as well as topic-related knowledge proved to be effective. The live classroom interview seems to bear the risk that the students take the oral reports about the past for granted, and this might lead them to learn less about history-related epistemological principles. However, at the same time, conducting a live oral history interview in school seems above all to be fun. A careful practice is needed—and so are studies that can shed more light on the opportunities and constraints offered by the live oral witness approach.

Notes

This research was supported by the Federal Ministry of Education and Research (BMBF), Germany, Project Number 01JG0913, and a grant by the Ministry of Science, Research, and the Arts of the State of Baden-Württemberg.

[1]To increase the database for imputation and scaling, we tested two more classes in an 11th school. Furthermore, before randomization, one teacher in one of the 10 schools where the interventions took place refused to participate in the intervention but allowed his or her class to be tested. We used the data from these three classes for imputation and scaling (total $N = 962$) but excluded them from the analyses. These data are not included in the description of the sample or in the results ($N = 900$).

[2]Because one expert teacher taught all the lessons in the 30 intervention classes, the intervention period covered nearly one school year from November 2011 to July 2012. In order to have similar measurement points for the assessments in the control group, the assessments of the control group classes also covered nearly the whole school year (from November 2011 to July 2012). Due to this long time period, the topics in the control group classes varied from the Weimar Republic in October to the end of the Cold War in July. Also, the teaching methods varied across the five control group classes.

[3]After the first round of item development, we became aware of Maggioni's Beliefs in History Questionnaire (BHQ) and recognized important parallels. However, in line with the goal of our intervention unit, we focused on students' insights into the need to reconstruct the past, to deconstruct history, and specifically for oral history interviews, whereas Maggioni was interested in exploring three types of historical cognition. Thus, we decided to proceed with our instrument and evaluated it in several surveys with high school students as well as with university students before administering it in the present study.

[4]Concerning the fill-in-the-blank questions on the pretest and posttest, missing data that resulted from the multimatrix design were imputed. Regarding the item that measured the frequency with which students had spoken with an eyewitness of the GDR, 36% missing data were counted because many of our students living in the west part of Germany did not know any people from the former GDR.

[5]The use of multiple imputation and maximum likelihood–based procedures (full information maximum likelihood; FIML) usually converge—if the same variables are used in both models—to identical results, which are unbiased if the missing at random (MAR) assumption holds, which means that the missing process is "explained" by variables included in the imputation (MI) or estimation (FIML) process (Enders, 2010).

[6]All effects that were based on the adjusted models (i.e., models including covariates) reported in this study were also estimated in the unadjusted models (i.e., models without any covariates) to check the sensitivity of the results with regard to the adjustment. Only minor differences between the effect estimates (i.e., regression coefficients) from the two types of models were found, although the significance level was affected in some cases, especially for the comparison between the treatment groups and the small control group. However, given the low power of the unadjusted analyses, these analyses are likely to underestimate the differences between groups and are not reported here in detail.

References

Angvik, M., & Borries von, B. (Eds.). (1997). *Youth and history: A comparative European survey on historical consciousness and political attitudes among adolescents*. Hamburg: Körber-Stiftung.

Assmann, A. (2006). History, memory, and the genre of testimony. *Poetics Today*, *27*(2), 261–273. doi:10.1215/03335372-2005-003

Assmann, J. (2010). Communicative and cultural memory. In A. Erll & A. Nünning (Eds.), *Cultural memory studies. an international and interdisciplinary handbook* (pp. 109–118). Berlin: De Gruyter.

Assmann, J., & Czaplicka, J. (1995). Collective memory and cultural identity. *New German Critique 65*, 125–133. Reprinted from J. Assmann & T. Hölscher (Eds.). (1988). *Kultur und Gedächtnis* [Culture and memory] (pp. 9–19). Frankfurt: Suhrkamp. doi:10.2307/488538

Barber, S., & Peniston-Bird, C. (Eds.). (2009). *History beyond the text. A student's guide to approaching alternative sources*. London: Routledge.

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*(4), 948–955. doi:10.1093/biomet/86.4.948

Bernard-Donals, M., & Glejzer, R. (2001). *Between witness and testimony. The holocaust and the limits of representation*. Albany, NY: State University of New York.

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). Randomized trial of teacher development in elementary science: First-year achievement effects. *Journal of Research on Educational Effectiveness*, *1*(4), 237–264. doi:10.1080/19345740802328273

Boyd, D. a., Fernheimer, J. W., & Dixon, R. (2015). Indexing as engaging oral history research: Using OHMS to "Compose History" in the writing classroom. *Oral History Review*, *42*(2), 352–367. doi:10.1093/ohr/ohv053

Breakstone, J. (2013). *History assessment of thinking. Design, interpretation and implementation*. (Doctoral dissertation). Stanford University. Retrieved from https://purl.stanford.edu/nt301xp3169

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

Breakstone, J., Wineburg, S., & Smith, M. (2015). Formative assessment using Library of Congress documents. *Social Education*, *79*(4), 178–182.

Bridge, D. J., & Paller, K. A. (2012). Neural correlates of reactivation and retrieval-induced distortion. *The Journal of Neuroscience*, *32*(35), 12144–12161. doi:10.1523/JNEUROSCI.1378-12.2012

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods*, *5*, 159–172.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods*, *2*, 49–68.

Christensen, L. (2012). Types of designs using random assignment. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (pp. 469–488). Washington, DC: American Psychological Association.

Clark, A. (2008). *A comparative study of history teaching in Australia and Canada. MONASH university education*. Retrieved from http://www.historyteacher.org.au/files/200804_HistoryTeachingReport.pdf

Common Core State Standards Initiative. (2010). *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/ELA-Literacy/RH/9-10/

De, La, Paz, S., Malkus, N., Monte-Sano, C., & Montanaro, E. (2011). Evaluating American history teachers' professional development: Effects on student learning. *Theory and Research in Social Education*, *39*(4), 494–540. Retrieved from http://www.education.umd.edu/Academics/Faculty/Bios/facData/CHSE/sdelapaz/TRSE.pdf

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Ercikan, K., & Seixas, P. (2011). Assessment of higher order thinking: The case of historical thinking. In G. Schraw (Ed.), *Assessment of higher order thinking skills* (pp. 245–261). Scottsdale, AZ: Information Age Publishing.

Ercikan, K., & Seixas, P. (Eds.). (2015). *New directions in assessing historical thinking*. New York, NY: Routledge.

Erdmann, E., & Hasberg, W. (2011). *Facing—mapping—bridging diversity. Foundation of a European discourse on history education*. Schwalbach: Wochenschau-Verlag.

Freund, A. (2015). Under storytelling's spell? Oral history in a neoliberal age. *Oral History Review 42*(1), 96–132. doi:10.1093/ohr/ohv002

Frost, J., de Pont, G., & Brailsford, I. (2012): Expanding assessment methods and moments in history. *Assessment and Evaluation in Higher Education*, *37*(3), 293–304.

Galda, M. (2013). *Geschichtsbewusstsein, historisches Wissen und Interesse. Darstellung von Zusammenhängen und Repräsentationen in semantischen Netzwerken*. [Historical consciousness, historical knowledge, and interest. Description of relations and representations in semantic networks.] (Doctoral dissertation). University of Frankfurt a. M., Frankfurt a. M. Retrieved from http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/28712

Goald, J. G., & Gradowski, G. (2014). Using online video oral histories to engage students in authentic research. *Oral History Review*, *41*(2), 341–350. doi:10.1093/ohr/ohu031

Hattie, J. A. C. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Havekes, H., Arno-Coppen, P., Luttenberg, J., & van Boxtel, C. (2012). Knowing and doing history: A conceptual framework and pedagogy for teaching historical contextualization. *International Journal of Historical Learning, Teaching and Research*, *11*(1), 72–93. Retrieved from http://www.cumbria.ac.uk/Public/ResearchOffice/Documents/Journals/IJHLTRNov12.pdf

Henke-Bockschatz, G. (2014). *Oral History im Geschichtsunterricht* [Oral History in history lessons]. Schwalbach: Wochenschau Verlag.

King, P. M., & Kitchener, K. S. (2002). The reflective judgment model: Twenty years of research on epistemic cognition. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 37–61). Mahwah, NJ: Lawrence Erlbaum Associates.

Klieme, E., Pauli, C., & Reusser, K. (2009). the pythagoras study. In T. Janík & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster: Waxmann.

Körber, A., & Meyer-Hamme, J. (2015). Historical thinking, competencies and their measurement: Challenges and approaches. In K. Ercikan & P. C. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 89–101). New York, NY: Routledge.

Köster, M., Thünemann, H., & Zülsdorf-Kersting, M. (2014). *Researching history education. International perspectives and disciplinary traditions*. Schwalbach: Wochenschau-Verlag.

Kuhn, D., & Weinstock, M. (2002). What is epistemological thinking and why does it matter? In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 121–144). Mahwah, NJ: Lawrence Erlbaum Associates.

Lanman, B. A. (1987). Oral history as an educational tool for teaching immigration and Black history in American high schools: Findings and queries. *International Journal of Oral History*, *8*, 122–135.

Lanman, B. A., & Wendling, L. M. (2006a). Foxfire and the Foxfire approach: Excerpts from the publications of The Foxfire Fund, inc. editors' introduction. In B. A. Lanman & L. M. Wendling (Eds.), *Preparing the next generation of oral historians: An anthology of oral history education* (pp. 5–14). Lanham, MD: AltaMira Press.

Lanman, B. A., & Wendling, L. M. (Eds.). (2006b). *Preparing the next generation of oral historians: An anthology of oral history education*. Lanham, MD: AltaMira Press.

Lee, C. R., & Nasstrom, K. L. (1998). Practice and pedagogy: Oral history in the classroom. *Oral History Review, 25*(1/2), 1–7. doi:10.1093/ohr/25.1.1

Lee, P. (2004). Understanding history. In P. Seixas (Ed.), *Theorizing historical consciousness* (pp. 129–164). Toronto: University of Toronto Press.

Lee, P., & Ashby, R. (2000). Progression in historical understanding among students ages 7–14. In P. N. Stearns, P. Seixas, & S. Wineburg (Eds.), *Knowing, teaching & learning history. National and international perspectives* (pp. 199–222). New York, NY: New York University Press.

Lee, P., & Shemilt, D. (2003). A scaffold, not a cage: Progression and progression models in history. *Teaching History*, *113*, 13–24.

Loewen, J. W. (1996). *Lies my teacher told me. Everything your American history textbook got wrong*. New York, NY: New Press.

Loewen, J. W. (2004). Foreword. In G. Whitman (Ed.), *Dialogue with the past: Engaging students and meeting standards through oral history* (pp. ix–x). Walnut Creek, CA: AltaMira Press.

Loftus, E. F. (1999). Lost in the mall. Misrepresentations and misunderstandings. *Ethics & Behavior*, *9*(1), 51–60.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*, 585–589. doi:10.1016/S0022-5371(74)80011-3

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment—A reanalysis of TIMSS data. *Learning Environments Research*, *9*, 215–230.

Lutteroth, J. (2011, November 18). Affäre um Hitlers Nachfolger. Dönitz erzählt vom Krieg. [Scandal about Hitler's successor. Dönitz is telling about the War] *Spiegel Online*. Retrieved from http://einestages.spiegel.de/static/topicalbumbackground/23961/doenitz_erzaehlt_vom_krieg.html

Maggioni, L. (2010). *Studying epistemic cognition in the history classroom: Cases of teaching and learning to think historically* (Doctoral dissertation). Retrieved from http://drum.lib.umd.edu/bitstream/1903/10797/1/Maggioni_umd_0117E_11443.pdf

Maggioni, L., Alexander, P., & VanSledright, B. (2004). At a crossroads? The development of epistemological beliefs and historical thinking. *European Journal of School Psychology, 2*(1–2), 169–197.

Maggioni, L., VanSledright, B., & Alexander, P. (2009). Walking on the borders: A measure of epistemic cognition in history. *The Journal of Experimental Education*, *77*(3), 187–213. doi:10.3200/JEXE.77.3.187-214

Mandell, N. (2008). Thinking like a Historian: A framework for teaching and learning. *OAH Magazine of History*, *22*(2), 55–63.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison Perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

Monte-Sano, C., & De La Paz, S. (2012). Using writing tasks to elicit adolescents' historical reasoning. *Journal of Literacy Research*, *44*(30), 273–299.

Monte-Sano, C., & Reisman, A. (2016). Studying historical understanding. In L. Corno & E. M. Anderman (Eds.), *Handbook of educational psychology* (3rd ed., pp. 281–294). New York, NY: Routledge.

National Center for History in the Schools UCLA. (1996). *National standards for history basic edition*. Retrieved from http://www.nchs.ucla.edu/Standards/historical-thinking-standards-1

Obens, K., & Geißler-Jagodzinksi, C. (2008). *"Dann sind wir ja auch die letzte Generation, die davon profitieren kann". Erste Ergebnisse einer empirischen Mikrostudie zur Rezeption von Zeitzeugengesprächen bei Jugendlichen/jungen Erwachsenen* ["Then we are the last generation that will benefit from this." The first results of an empirical microstudy of the reception of oral history interviews by the young]. Retrieved from http://www.migration-online.de/data/forschungsbericht_zeitzeugengesprche.pdf

Peniston-Bird, C. M. (2009). Oral history: The sound of memory. In S. Barber & C. M. Peniston-Bird (Eds.), *History beyond the text. A student's guide to approaching alternative sources* (pp. 105–121). London: Routledge.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*, 109–119. doi:10.3102/0013189X09332374

Plato, A. von, & Burley, E. (2009). Contemporary witnesses and the historical profession: Remembrance, communicative transmission, and collective memory in qualitative history. *Oral History Forum d'histoire orale*, *29*, 1–27. Retrieved from http://www.oralhistoryforum.ca/index.php/ohf/article/view/56

Portelli, A. (1991). The death of Luigi Trastulli: Memory and the event. In A. Portelli, *The death of Luigi Trastulli and other stories. Form and meaning in oral history* (pp. 1–26). Reprinted from Portelli, A. (1981). La memoria e l'evento. L'assassinio di Luigi Trastulli. *Segno Critico*, *2*(4), 5–32.

Raghunathan, T. E., Solenberger, P. W., & Hoewyk, J. V. (2002). *IVEware: Imputation and variance estimation software. Installation instructions and user guide*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

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

Ritchie, D. A. (Ed.). (2011). *The Oxford handbook of oral history*. Oxford, UK: University Press.

Rouet, J. F., Britt, M. A., Mason, R. A., & Perfetti, C. A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology*, *88*(3), 478–493. doi:10.1037/0022-0663.88.3.478

Rüsen, J. (2005). *History. Narration—interpretation—orientation*. New York, NY: Berghahn Books.

Sabrow, M. (2012). Der Zeitzeuge als Wanderer zwischen den Welten [The contemporary witness as wayfarer between two worlds]. In M. Sabrow & N. Frei (Eds.), *Die Geburt des Zeitzeugen nach 1945* (pp. 13–32). Göttingen: Wallstein Verlag.

SAS Institute Inc. (2013). *SAS/STATÂ® 13.1 User's guide*. Cary, NC: Author.

Schreiber, W., & Árkossy, K. (2009). *Zeitzeugengespräche führen und auswerten. Historische Kompetenzen schulen* [Conducting interviews with contemporary witnesses. Teaching historical competencies]. Neuried: ars una.

Schulz, K. F., Altman, D. G., & Moher, D., for the CONSORT Group. (2010). Research methods & reporting. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. *British Medical Journal*, *340*. doi:10.1136/bmj.c332

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.

Shemilt, D. (2000). The Caliph's coin: The currency of narrative frameworks in history teaching. In P. N. Stearns, P. Seixas, & S. Wineburg (Eds.), *Knowing, teaching & learning history. National and international perspectives* (pp. 83–101). New York, NY: New York University Press.

Shopes, L. (2015). After the interview ends: Moving oral history out of the archives and into publication. *Oral History Review*, *42*(2), 300–310. doi:10.1093/ohr/ohv037

Smith, M., & Breakstone, J. (2015). History assessments of thinking. An investigation of cognitive validity. In K. Ercikan & P. Seixas, P. (Eds.), *New directions in assessing historical thinking* (pp. 233–245). New York, NY: Routledge.

Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. (2009). *Stärkung der Demokratieerziehung* [Reinforcement of education for democracy]. Retrieved from: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2009/2009_03_06-Staerkung_Demokratieerziehung.pdf

Stearns, P. N., Seixas, P., & Wineburg, S. (2000). *Knowing, teaching & learning history. National and international perspectives*. New York, NY: New York University Press.

Stein, A. H., & Scatena, M. (2006). Studs' place in oral history education. In B. A. Lanman & L. M. Wendling (Eds.), *Preparing the next generation of oral historians: An anthology of oral history education* (pp. 17–31). Lanham, MD: AltaMira.

Stoel, G., van Drie, J. P., & van Boxtel, C. A. M. (2016). The effects of explicit teaching of strategies, second-order concepts, and epistemological underpinnings on students' ability to reason causally in history. *Journal of Educational Psychology*. Retrieved from http://dx.doi.org/10.1037/edu0000143

Stricklin, D., & Sharpless, R. (Eds.). (1988). *The past meets the present. Essays on oral history*. Lanham, MD: University Press of America. Retrieved from http://www.baylor.edu/content/services/document.php/33242.pdf

Swan, K., Hofer, M., & Locascio, D. (2008). The historical scene investigation (HIS) project: Instruction in the fifth grade social studies classroom. *International Journal of Social Education*, *22*(2), 70–100.

Thompson, P. (1978). *The voice of the past: Oral history*. Oxford, UK: Oxford University Press.

Thomson, A. (2015). Anzac memories revisited: Trauma, memory and oral history. *Oral History Review*, *42*(1), 1–29. doi:10.1093/ohr/ohv010

Trautwein, U., Bertram, C., Borries, B., Brauch, N., Hirsch, M., Klausmeier, K., . . . Zuckowski, A. (in press). *Assessing historical thinking competencies: Conceptualization, operationalization, and results of the project "Historical Thinking – Competencies in History"*. Stuttgart: Waxmann-Verlag.

Van Drie, J., & van Boxtel, C. (2008). Historical reasoning: Towards a framework for analyzing students' reasoning about the past. *Educational Psychological Review*, *20*, 87–110. doi:10.1007/s10648-007-9056-1

VanSledright, B. A. (2002). *In search of America's past: Learning to read history in elementary school*. New York, NY: Teachers College Press.

VanSledright, B. A. (2004). What does it mean to think historically . . . and how do you teach it? *Social Education*, *68*(3), 230–233.

VanSledright, B. A. (2014). *Assessing historical thinking & understanding. Innovative designs for new standards*. New York, NY: Routledge.

VanSledright, B., & Maggioni, L. (2016). Epistemic cognition in history. In J. A. Green, W. A. Sandoval, & I. Braten (Eds.), *Handbook of epistemic cognition* (pp. 128–146). New York, NY: Routledge.

VanSledright, B., & Reddy, K. (2014). Changing epistemic beliefs? An exploratory study of cognition among prospective history teacher. *Revista Tempo e Argumento*, *6*, 26–68.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. doi:10.1007/BF02294627

Welzer, H., & Markowitsch, H-J. (2005). Towards a bio-psycho-social model of autobiographical memory. *Memory*, *13*(1), 63–78. doi:10.1080/09658210344000576

Westhoff, L. M. (2009). Lost in translation. The use of primary sources in teaching history. In R. G. Ragland & K. A. Woestman (Eds.), *The Teaching American History Project: Lessons for history educators and historians* (pp. 62–76). New York, NY: Routledge.

Whitman, G. (2004). *Dialogue with the past. Engaging students & meeting standards through oral history*. Walnut Creek, CA: AltaMira Press.

Whitman, G. (2011). Motivating the twenty-first-century students with oral History. In D. Ritchie (Ed.), *The Oxford handbook of oral history* (pp. 449–469). Oxford, UK: University Press.

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, PA: Temple University Press.

Wineburg, S. (2004). Crazy for history. *Journal of American History*, *90*(4), 1401–1414.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalized item response modelling software*. Camberwell: ACER Press.

Manuscript received January 29, 2016
Final revision received January 16, 2017
Accepted January 17, 2017