

Standardized assessment of historical thinking competencies in an intervention study using perspectives on German history

Katharina Totter* , Wolfgang Wagner , Christiane Bertram 

University of Tübingen, Tübingen

Abstract


To assess the efficacy of an intervention study on the German post-1990 transformation targeting historical thinking, this paper presents the development of a standardized test designed to measure epistemological understanding and methodological competencies. Following a validation study ($N = 354$ students), we employed a revised test in an intervention study with $N = 1,301$ high school students in Baden-Württemberg. The newly developed tests underwent analysis concerning their psychometric criteria. The final test contained 38 items with various stimuli (e.g., interview snippets, cartoons) utilizing closed-format responses. The methodological test exhibited sufficient reliability and extensive overlap with a selection of items from an established test. However, the epistemological test showed some limitations in both reliability and validity, suggesting a potential opportunity for improvement through revision. Students' grades in history and German, cognitive skills, and socioeconomic status predicted their ability scores based on two-parameter logistic (2PL) item response models for both tests.

Keywords

standardized achievement test, historical thinking, test development, historical competencies, assessment

1. Introduction

Germany's latest historical seismic event, the merging of two states—the East and the West—into the German Federal Republic, occurred nearly 35 years ago. To this day, the former divide is still noticeable when looking at not only the distribution of wealth or election results but also perspectives on the transformation that followed from 1989/1990 (see, e.g., Großbölting, 2020). A recent project (“Generation 1975” see Bertram, 2020) with eyewitnesses who grew up on opposite sides of the “Iron Curtain”, which divided Germany into two countries, the FRG and GDR, during the Cold War, revealed that views in the East and West are still fundamentally different. Whereas the interviewees in the East experienced a complete change in their everyday lives, little had changed for those in the West (Bertram, 2020). Due to their predispositions from how

* **Contact:** Katharina Totter  katharina.totter@uni-tuebingen.de
University of Tübingen, Hector Research Institute of Education Sciences and Psychology, Tübingen, Germany

they were socialized in the competing systems (Großbölting, 2020), their accounts of historical events were influenced by and remembered differently across cultural and social groups (Körber & Lenz, 2014). These differences have not necessarily smoothed out over the years but rather developed into contradictory narratives (Rensmann, 2019). For instance, the East's perspective of being "taken over" by the West and being treated as second-class citizens contrasts with the West's perspective, where East Germans "lament their fate instead of being grateful for the (...) political and economic opportunities" (Rensmann, 2019, p. 33) that the West offered them. An ongoing debate about the past, representations, and relevance for the present is a strength of democratic societies, provided that citizens can participate and contribute to the discussion (Körber & Meyer-Hamme, 2015). However, this means that they need to be aware of and equipped with epistemic beliefs to appropriately handle conflicting perspectives on complex issues (VanSledright & Maggioni, 2016), such as reunification. Rather than assuming there is only one objective truth or that all viewpoints are merely subjective opinions that can be accepted or rejected depending on an individual's worldview, they should recognize that while different narratives must be critically examined for validity, they can certainly exist side by side (VanSledright & Maggioni, 2016). This directly aligns with a core concept in history teaching: multiperspectivity (Körber & Lenz, 2014).

A recent large-scale randomized controlled field trial (RCT) in Baden-Württemberg, Germany, applied this approach to foster historical thinking by having students engage with different perspectives on the topic of the time of the transformation with eyewitnesses of the aforementioned "Generation 1975". Classes were randomized into three conditions: two that received the intervention and one waitlist control group. During a three-lesson unit, the intervention students first prepared for interviews with eyewitnesses, one from the East and one from the West. The classes then differed in the second lesson regarding whether the students worked with the accounts obtained from videos or in-person interviews. Afterward, students in both conditions drew connections to both recent and historical contexts from the statements made in the interviews to address the core question of the lesson unit: "Has what belongs together grown together?" The intervention was designed to expand students' knowledge about the topic, addressing motivational aspects and historical competencies. Although learning with in-person eyewitnesses holds great potential to motivate students, it also poses a risk for students' historical learning (Bertram et al., 2017). Therefore, one of the main goals of the project was to foster historical thinking. To answer the core question of the intervention, students were constantly asked to use their historical thinking abilities. They had to engage with a variety of historical sources and accounts to develop the questions they wanted the eyewitnesses to answer and to contextualize their answers later. Furthermore, they were not only confronted with two opposing perspectives on German reunification but also had to sharpen their understanding of what conclusions can (and cannot) be drawn from the materials studied, particularly the eyewitness accounts.

One challenge of this large-scale study with over 1,000 students was how to assess the impact on the acquisition of historical thinking competencies. We applied standardized tests in our study for two reasons. First, standardized measurement procedures, in general, enhance the "clarity of communication" (Gelman & Hennig, 2017, p. 973) about the study results because they are carried out and evaluated in a clearly specified manner. Second, standardized tests are typically quite time- and cost-efficient, at least with regard to the coding of correct versus incorrect responses, in particular when closed answer formats are used. Currently, only a limited number of standardized test items that capture historical thinking are available (e.g., the HiTCH test; Trautwein et al., 2017). Even though the HiTCH test can capture historical competencies independent of specific topics, additional instruments more closely related to the intervention's topic and aims were missing. Therefore, we developed two new historical thinking tests to assess specific competencies we aimed to foster during the intervention: students' understanding of epistemological principles and methodological competencies. After providing a theoretical background, we present and discuss the tests in two steps, focusing on their validity and reliability. First, we used the empirical results of the newly developed items ($k = 58$) from a small sample in a validation study to eliminate or refine those that did not perform well on psychometric or content criteria. Next, we examined the performance of the final tests in the large sample of ninth-grade students in the intervention study.

2. Theory

In Western democracies, it has been declared that the main goal of history education is to foster students' historical thinking competencies and historical consciousness (see Lévesque & Clark, 2018). Although definitions and models of historical thinking – sometimes also referred to as historical reasoning (Van Drie & Van Boxtel, 2008) – differ (see, e.g., Van Drie & Van Boxtel, 2008; Wineburg, 1991), there are huge commonalities in the literature in the Western world (Lévesque & Clark, 2018). Seixas (2017) condensed a broad consensus among historians about these epistemological underpinnings into three main principles: differentiating between the past and history, the coexistence of multiple historical narratives, and the necessity to critique them in terms of their plausibility. Based on these key aspects, the current test development focused on the epistemological understanding of the nature of history and the methodological implications that arise in approaching these narrations. The next section outlines the epistemological understanding and methodological competencies and discusses their measurement.

2.1 Understanding epistemological principles of history

Epistemology deals with the nature and justification of knowledge (Hofer & Pintrich, 1997). The essence of history is essentially its narrativity (Rüsen, 2005). The epistemological principles of history are formed by a theory of history consisting of construction and narratives (Rüsen, 2005). A narration connects the past and history by transforming one into the other (Rüsen, 2005). Inevitably, this means history must be constructed by assembling past fragments to create a meaningful narrative (Rüsen, 2005, 2017). Directly linked to these foundational aspects of history are the principles of retrospectivity, selectivity, and particularity (Rüsen, 1989, 2005; see also VanSledright, 2014). Furthermore, history is narrated from a particular perspective (Rüsen, 1989), and multiple narratives coexist based on, for example, the narrator's perspective on the events, and there is no such thing as one, true history (Rüsen, 2005; Seixas, 2017).

People's views about the nature of knowledge (i.e., what, how certain, and how interrelated knowledge is) and the process of acquiring knowledge (i.e. the sources from which knowledge comes and how to evaluate and justify knowledge claims) are often referred to as their "epistemological beliefs" or "understanding" (Hofer & Pintrich, 1997). They can be mapped in development models and influence the cognitive processes of learners (see Hofer & Pintrich, 1997). VanSledright (2014) considers epistemological misunderstandings about history to be the greatest obstacle to historical understanding.

It is evident that these epistemological principles constitute an essential part of students' understanding of history. In the model developed by the FUER group, an international consortium whose acronym stands for the promotion and development of a reflective and (self-) reflexive historical consciousness (in German: Körber et al., 2007; English translation: Körber & Meyer-Hamme, 2015), insights into epistemological principles act as the foundation of historical thinking (Sachkompetenzen). In the international discussion, epistemological principles are referred to as second-order concepts (Lévesque & Clark, 2018). Previous publications on the development of epistemological ideas (e.g., from Kuhn & Weinstock, 2002) were consolidated by Maggioni et al. into a staged model (Maggioni et al., 2009; overview in Stoel et al., 2017). In the initial stage, students hold novice beliefs, perceiving historical knowledge as "fixed and a singular copy of the past" (Stoel et al., 2017, p. 122). The Copier holds beliefs such as that history equals the past, and historians are mere chronologists (Maggioni, 2010). The Subjectivist, on the other hand, views knowledge as a matter of personal opinion and perspective (Maggioni, 2010). Students' beliefs can progress to a more advanced level (Maggioni et al., 2009) until they reach the final development stance, Criterialist (Maggioni, 2010). In an expert stage, historical knowledge is viewed as constructed, interpretative, and changeable, and claims about the past need to be approached with disciplinary criteria (Maggioni, 2010). Stoehl et al. (2017) emphasize this value that students place on the methodological approach to historical knowledge as an important part of advanced historical thinking (as opposed to naïve beliefs).

2.2 Using methodological competencies: reconstruction and deconstruction

Derived from an understanding of the epistemological principles of history, the necessity to work with and critique narratives becomes evident; therefore, operating with adequate disciplinary criteria is essential (Seixas, 2017). In such a context, Rösen's dimensions of plausibility are frequently utilized (Rösen's *Triftigkeiten* (1989, 2013), in English, see also Körber, 2016; Seixas, 2017).

The elaboration of the Historical Thinking Standards, published by the National Center for History in Schools UCLA (1996), describes students' abilities "to create historical narratives and arguments on their own" and "thoughtfully read the historical narratives created by others (...) with conceptual analysis drawn from all relevant disciplines" (Historical Thinking Standards section). Sometimes referred to as methodological, these competencies involve working with historical material to engage in a constructional process involving either reconstructing or deconstructing historical narratives (Körber, 2011). To reconstruct a historical narrative, one locates pieces of information about the past and assembles, interprets, and integrates them to synthetically construct historical statements. The process of deconstruction starts with a given historical account, finds its narrative structures, and locates and evaluates the pieces of information given and the relationships made. In comparison with other international works, great similarities can be found in works by, for example, Wineburg (1991) and Van Drie and Van Boxtel (2008).

In both constructional processes, Rösen's disciplinary criteria (1989) needs to be applied to assess the empirical, normative, and narrative plausibility of the historical narrative. When evaluating the empirical plausibility, the focus lies on the narrative's reliance on past information and the quality, quantity, and relevance thereof (Körber, 2016). Are verifiable facts mentioned, and if so, how many? Can the claims be fact-checked using evidence from the past? Both of these questions are answered by critically evaluating the source material the narrative holds (Rösen, 2017). Concerning the normative plausibility of a historical narrative, the assessment focuses on the values and norms conveyed (Körber, 2016). The perspective of the narrative holds meaning and orientation for the present, and multiple narratives can contradict each other on the basis of their perspectives (Rösen, 2017). In practice, questions about the validity of the perspectives should be answered along the lines of: What are the values and norms conveyed, and do they match the audience and beyond? Are they acceptable (to all)? (Körber, 2016). Lastly, narrative plausibility focuses on the structure of the narration. Are the construction and the elements used therein convincing and logical? (Körber, 2016).

In summary, when reconstructing the past in a historical narrative, students, or anyone who deals with history, should consider the three plausibility criteria. These criteria are also applied when deconstructing given narratives. It is evident that understanding the epistemological underpinnings of history and applying disciplinary criteria to both reconstruct and deconstruct narratives are complex historical thinking competencies. Ways to assess these competencies and their consequences for the development of the test on epistemology and methodology are discussed in the next section.

2.3 Assessment of competencies of historical thinking

Numerous studies have used qualitative data to assess aspects of historical thinking mostly involving the ability to reconstruct and deconstruct historical narratives (e.g., see Waldis et al., 2015; Wineburg, 1991) and students' epistemological beliefs (e.g., Iordanou et al., 2020). Most of these assessments were based on students' writings, where a detailed evaluation of their abilities regarding these complex historical competencies was conducted (Ercikan & Seixas, 2015). When experts evaluate students' answers, even though there is considerable potential for high measurement validity, there is a lack of objectivity (Radinsky et al., 2015), and disentangling the relationship with basic literacy poses a challenge (Ercikan & Seixas, 2015). As students are often required to read and produce full paragraphs in these assessments, the question that arises is to what extent the bycatch of students' literacy is being assessed versus their historical thinking competencies (Ercikan & Seixas, 2015). Moreover, the length of the texts, especially considering the time it takes to produce and evaluate them, poses an issue (Bertram et al., 2021). For the purpose of assessing average achievement, assessing the distribution of achievement within groups, and comparing achievement between groups, qualitative approaches are not (yet) suitable (Körber & Meyer-Hamme, 2015). Suitability might lie in the future possibility of analyzing qualitative data in large-scale samples via artificial intelligence (e.g., see Bertram et al., 2021), although there are still major disadvantages (e.g., unreliability and the length of the

test for the students). At the moment, it is not yet possible to assess the content accuracy of the students' answers without a differentiated manual rating (Bertram et al., 2021).

Therefore, large-scale assessments call for quantitative approaches (Körber & Meyer-Hamme, 2015). There are many merits of having a highly standardized measure capable of assessing historical thinking competencies with high reliability and validity. With the lack of large empirical studies in mind (Körber & Meyer-Hamme, 2015), the test would be easy to use and evaluate in these large-scale settings (Wagner et al., 2023). Such tests are usually objective with no difference between raters, and there is usually a clear interpretation and replicability of the results, in contrast to open formats (Radinsky et al., 2015). Moreover, highly standardized and closed formats are more time- and cost-effective for both students and researchers (Smith, 2017).

Standardized assessment and the development of quantitative measures regarding epistemic beliefs have been proposed using items that are bare of concrete historical context that capture the degree of which participants (teachers or students) agree or disagree with them (e.g., Maggioni et al., 2009; Stoel et al., 2017; Wiley et al., 2020). The results regarding aspects of validity and reliability seem promising (see, e.g., Wiley et al., 2020), but limitations arise with regard to the lack of historical context (Stoel et al., 2017), and in some studies with students, the scores that were obtained were not related to students' abilities (Wiley et al., 2020).

Utilizing a scenario-based approach with a concrete historical context, Barzilai and Weinstock (2015) employed multiple-choice items to assess epistemic thinking. Students responded to statements that mirrored epistemic perspectives derived from a model by Kuhn and Weinstock (2002), whose categories (absolutist, multiplist, evaluativist) map onto the stances Maggioni (2010) proposed. Barzilai and Weinstock (2015) aimed to assess students' application of their epistemic assumptions rather than asking specifically about their epistemic beliefs in the form of a self-report. Results revealed that measuring students' epistemic thinking was generally possible across, yet influenced by, the historical topics and problems presented.

Standardized assessments of a variety of students' historical thinking competencies and students' attitudes toward history have already been employed in large-scale studies such as the European-wide study in 1997 by Angvik et al. or by the National Assessment of Educational Progress (NAEP), which has been assessing students' historical thinking since the 1960s but has been criticized for mainly assessing declarative knowledge (VanSledright, 2014). In 2017, Smith encountered severe validity problems with multiple-choice items used in the NAEP to measure historical thinking processes. Think-aloud protocols revealed that the items were not able to represent the complex thinking processes they were constructed for. The conclusion was that students did not engage in historical thinking processes but instead relied on their abilities to read, answer strategically on tests, and recall factual knowledge. In a follow-up study, Smith (2018) showed that newly developed multiple-choice items focusing on Wineburg's (1991) historical thinking constructs (sourcing, contextualization, corroboration) outperformed sample items from established standardized U.S. tests (e.g., the NAEP) in terms of validity. Aiming for large-scale use, the FUER group (Körber et al., 2007) developed a standardized test in German to measure a variety of historical thinking competencies in 2017 (Trautwein et al.) with closed formats. The HiTCH test is one of the few standardized tests that captures the various facets of historical thinking competencies, although the facets are not yet equally represented in the inventory, and the test is under constant development (see Wagner et al., 2023).

In summary, there are several main challenges specifically related to the standardized measurement of historical thinking competencies. How can one introduce the topic/context without it influencing the results of the assessment too much? Does the test measure the specific competency, and is it able to reflect the historical thinking process, or does it assess only factual knowledge or reading skills, for example? Most importantly, the valid assessment of competencies such as historical thinking competencies is especially difficult due to their complexity (Smith, 2017). The inherent nature of history, consisting of (multiple coexisting) narratives (Rüsen, 2005), often challenges the notion of having a single correct answer (VanSledright, 2014). However, aligned with a focus on competence rather than factual knowledge, the goal is not to, for example, produce a given narrative as a correct answer but rather to demonstrate the ability to think historically (Körber & Meyer-Hamme, 2015). Smith (2018) also entertained this idea, suggesting that, in general, multiple-choice items can measure complex competencies when concrete skills are assessed. Moreover, tasks should typically have a historical context without assessing merely factual knowledge, and the competencies being assessed should be transferable to different topics (Körber & Meyer-Hamme, 2015). Given the disciplinary challenges, especially in striving for high validity of measurement, we focused on two main aspects during item development. Answering the items correctly should not heavily rely on students' applica-

tion of their reading skills or factual knowledge. Furthermore, the correct answer to each item should primarily depend on the application of a specific historical thinking skill rather than a complex set of competencies.

In addition to these content-related criteria, we designed the tests using methodological approaches from psychology. Alongside constructing items that meet the usual criteria for test quality (e.g., see De Leeuw et al., 2008), we employed models from Item Response Theory (IRT; see De Ayala, 2009). This enabled us to empirically test the relationship between a latent variable assumed to represent the specific construct and the items, assess their difficulty, and judge them based on their performance to distinguish between students with low and high competencies.

In the following paragraphs, we outline the approach of developing two new tests on specific aspects of historical thinking. Our aim was to develop tests that aligned with the topic and objectives of the intervention study. The tasks needed to be designed to measure epistemological understanding and methodological competencies precisely, validly, and in a manner that was as closely related to the topic as possible. Utilizing the HiTCH test's logic as a blueprint seemed most fitting because it already addressed some of the disciplinary challenges. Tasks were designed to be independent from the topic but linked to the material; therefore, finding the correct answer should not necessarily require factual knowledge. Whereas the test may occasionally provide extra information, its aim is to keep the student's time and effort focused on the historical thinking task rather than, for example, requiring excessive reading (Trautwein et al., 2017). Our aim was to construct tasks with short questions and prompts that mainly use concrete historical material in the form of pictures (e.g., cartoons) or texts that require little reading. The materials included all additional factual knowledge necessary to solve the task. We presented the students with several possible answers from which to choose to solve the task.

Our aim was to construct a test using only closed-response formats and to ensure that testing, scoring, and result interpretation adhered to rigorous standards of objectivity. Consequently, during the development process, the emphasis was on achieving high levels of reliability and validity. One way to minimize measurement errors and thus ensure high reliability would be to match the items' difficulties to the ability distribution of the sample, thus preventing items from being too difficult or too easy (De Ayala, 2009). Moreover, the items need to be able to distinguish between students with high or low ability (De Ayala, 2009). We further aimed to motivate students to put effort into answering this nonmandatory test by incorporating items with interesting tasks and materials. Missing responses are a recurring issue in educational and psychological assessments (Rose et al., 2016), and unmotivated students' answers are a threat to an assessment's validity (Eklöf, 2010).

With the validity of the test being the most discussed in the literature and representing the most challenging aspect of test construction, the central question is: Do these items measure what we think they measure? (Kaliski et al., 2015). Here, different methods can be used, one of which employs models from IRT to represent the relationship between the items and the construct (De Ayala, 2009). A second method that helps assess the validity of the construct involves a priori assumptions about the theoretical interrelationships that are being empirically tested (Kaliski et al., 2015). Regarding other ability measures, when exploring relationships between the newly developed tests and established tests, Kaliski et al. (2015) emphasized that the scores obtained by students on the new test should demonstrate a robust correlation with established tests assessing the same construct. Furthermore, the expected correlation between the test scores and factual knowledge should be "substantial" (Kaliski et al., 2015, p. 198), yet not overly pronounced, as this could imply that students' test performance is heavily intertwined with their knowledge of the factual context of the items.

Students' perception of and motivation toward the subject of history should overall also be positively linked to their performance regarding historical competencies. The relevance students attribute to history (i.e., what it has to do with themselves, society, and human existence as a whole) seems to be linked with their development of epistemological beliefs (see Van Straaten et al., 2018). Moreover, as part of students' relevance of history (Van Straaten et al., 2018), their self-reported approach to forming opinions and justifying judgments should positively relate to their performance in methodological tasks as well as their epistemological understanding.

Students' expectancy beliefs in how much they think they're able to succeed (which includes their self-concept of how they assess their own competency) and their reasons for engaging with a subject (i.e., value beliefs) are closely related to their academic performance (see Eccles et al., 1983; Gaspard et al., 2015). This also holds for the subject of history, where Arens et al. (2016) found that students' self-concept in history was only substantially positively related to their performance in history, in contrast to their performance in other subjects like math or German.

Moreover, it has been shown that students' ability to evaluate the trustworthiness of historical sources is predicted by the value they place on this competency (Van der Eem et al., 2023). Additionally, students with more advanced epistemic beliefs held higher value beliefs (Guo et al., 2022). Therefore, students' performance in the two newly developed tests should be positively related to their value beliefs and self-concept in history.

In essence, it is crucial to explore whether the test effectively measures a construct such as historical thinking, unintentionally measures another construct, or does both (Kaliski et al., 2015). Another more practical aspect of validity lies in the test's ability to predict a person's behavior (Wiley et al., 2020). Student performance on the new tests should, to some extent, be predictable from their history grade. Prior research by Stoel et al. (2017) showed a positive relationship for nuanced epistemological beliefs. Reading is known to play a role in tests on historical thinking, and the challenge of minimizing the extent to which the test assesses reading rather than the construct was highlighted by Ercikan and Seixas (2015). Therefore, we also anticipate that some additional variance in the test scores can be attributed to basic cognitive and reading abilities (Kaliski et al., 2015). The number of books at home is a commonly used indicator of students' socioeconomic status and has consistently been shown to relate to student achievement (for an overview, see Heppt et al., 2022). Student characteristics (gender and age) should not contribute significantly to further explanation of variance, as performance on a test designed for all 9th-grade high school students in Germany should not depend upon these characteristics.

In conclusion, the development of a standardized historical thinking measure presents numerous challenges that need careful consideration. We endeavored to address these challenges as an interdisciplinary team comprising educational researchers, psychologists, and history education researchers by utilizing an interdisciplinary approach. The subsequent sections delve into the test's development and its evaluation with psychological standards. The research questions focus on the reliability and validity of the newly developed epistemological and methodological tests.

2.4 Research questions

1. Do the newly developed tests measuring historical thinking competencies adhere to psychometric standards regarding item discrimination and score reliability?
2. How are the newly developed tests associated with the selection of items from an established standardized test instrument used for assessing historical thinking, a factual knowledge test, and an evaluation of the perceived relevance of history and motivational variables associated with the subject of history?
3. To what extent do student characteristics (grades, abilities...) predict students' ability as measured by the newly developed tests?

3. Methods

The intervention study the tests were developed for was preregistered at the Registry of Efficacy and Effectiveness Studies (REES) prior to analysis (#14881.1v1). The intervention study on eyewitnesses was funded by the Deutsche Forschungsgemeinschaft (DFG) and approved by both the Ethics Committee of the University of Tübingen, Faculty of Economics and Social Sciences, and the state ministry of culture in Baden-Württemberg. In Germany, the grade levels can differ by 1 year based on education time (8 years being the G8-track, 9 years being the G9-track) in high school. The grade level the intervention was designed for was ninth grade (G8-track) and 10th grade (G9-track) because the topic of the intervention connects to the educational plans at the respective level. For easier reading, we refer to both levels as the ninth grade. The same logic applies to the validation study that we carried out in grade 10 (G8-track) and grade 11 (G9-track) and refer to as the 10th grade.

3.1 Development of the test inventory

3.1.1 Development phase

The operationalization of the two constructs – epistemological understanding and methodological competencies – and the selection of tasks for the items were based on the format of the HiTCH test (Trautwein et al., 2017). Through an iterative process, experts in history didactics, psychology, and education sciences collaboratively selected a total of 58 items for the initial draft. The closed-format items for both tests had to adequately represent the constructs, be answerable in a reasonable amount of time, and be understandable and motivating to the students. These items were organized into seven tasks, all presented as either simple multiple-choice (MC) items (where only one of several answers needed to be selected for the correct solution) or in a complex multiple-choice (CMC) format (where one or more of several answers needed to be selected for the correct solution).

3.1.2 Validation study and analysis strategies under IRT

In a validation study, a total of $N = 354$ 10th-grade students answered the newly developed tests along with other items. At this time, they were mostly 15 to 16 years old ($M = 15.59$, $SD = 0.72$) and had already completed the topics of the 9th-grade curriculum, which deals with German division and unification. As preregistered, all test answers were coded dichotomously (0 = false, 1 = correct) or coded as missing for invalid answers or nonresponse. The CMC items were scored as correct only if the student selected the correct pattern of answers. We wanted to assess the difficulty of the items and the extent to which the items allowed to differentiate between students' ability level (RQ1). We therefore analyzed the items using two-parameter logistic (2PL) IRT models (De Ayala, 2009; for an introduction in German, see Wagner, 2020). In IRT, a latent variable (e.g., historical competence), which means a variable that cannot be observed directly, corresponds with observable behaviors or manifest variables (De Ayala, 2009). This relationship can be depicted with a logistic regression function, where individuals' abilities are mapped onto the probabilities of solving the item correctly. In the simplest model, the items vary on only one parameter (one-parameter logistic [1 PL] model), namely, the item difficulty (b), which is equal to the location on the ability dimension (i.e., the latent variable in a unidimensional model) where individuals would be expected to have a 50% chance of solving the item correctly. Commonly, the parameter's range is -3 to 3 with items with $b < -2$ being "easy" and $b > 2$ being "hard" (De Ayala, 2009). We aimed for items that were within this range of difficulty for our student population (i.e., not too difficult, not too easy) and have some variance among the range (i.e., items that we expect to be solved by most, and some by fewer students). Most importantly, we wanted to assess the items' performance to differentiate between students with different competency levels (i.e., students with higher abilities should solve the item correctly, whereas those with lower abilities should not). We therefore applied 2PL models, where items are assumed to vary not only in their difficulty but also in their discrimination (a). Discrimination is the slope of the item characteristic curve and translates into the ability of the item to differentiate between individuals with different abilities. Desirable values range from 0.8 to 2.5 (De Ayala, 2009), which refers to the steepness of the slope (i.e., how sharply the item draws the line between higher and lower-ability students). Values that approach zero translate to the ability of the students less and less playing a role in their probability of solving the item correctly and negative values indicate that individuals with lower ability have a higher probability of solving the item (De Ayala, 2009; Wagner, 2020). A statistically significant positive discrimination of $a \geq 0.5$ was defined as the psychometric selection criterion for the items. These items should all be helpful in determining a final score for students that reflects their latent ability (here: historical competencies). We also revised tasks that did not yet adhere to our criteria, but where we saw the potential to improve them by revising the task description or the wording of the item. The final inventory for the intervention study was therefore selected on the basis of both psychometric and content-related criteria.

The original 58 items were clustered in seven different task formats (e.g., analyze cartoons, work with eyewitness quotes). Among these, 42 items exhibited statistically significant positive discrimination (with 29 above the threshold of $a = 0.5$). However, one task format from the methodological test comprised 24 items, nearly half of which did not exhibit statistically significant positive discrimination (i.e., did not contribute to the ability score they intended to

measure). We extensively redesigned the task format, resulting in 12 items being revised and another 12 items being excluded. Overall, based on the results of the validation study (see the overview in Table 1), a total of 14 items required revision (one item from the epistemological test and 13 from the methodological test). Additionally, 19 items were excluded (three items from the epistemological test and 16 from the methodological test). Overall, 25 items remained unaltered. The final inventory consisted of 39 items and can be found in the appendix as supplementary material (graphical material and interviews not included). The details and examples of the items from both tests, including the modifications and exclusions, are elaborated upon further in the subsequent sections. For additional details on the 2PL models referred to below, see the Statistical Analysis section and the Appendix.

Table 1: Psychometric quality of the items in the validation study

Item discrimination a	Frequencies split for tasks that cluster the items "z": Work with eyewitness quotes	All other
A statistically significant positive $a > 0.8$	3	10
B statistically significant positive $0.5 < a < 0.8$	6	10
C statistically significant positive $a < 0.5$	4	9
D not statistically significant positive	11	5

Note. Item statistics were estimated with a 2PL model including all items ($N = 353$). Reported frequencies are split between one task format which exhibited very poor item discrimination, and all other task formats.

3.2 Test inventory on central epistemological principles

Students were tasked with selecting one or more suitable responses for a statement concerning an epistemological principle of history. The responses represented epistemological beliefs that ranged from naïve (i.e., historical knowledge is merely subjective opinions / pure facts) to advanced (i.e., historical knowledge is constructed, changeable, and interpretable, whereby the need for critical evaluation using disciplinary criteria is particularly important; see Stoel et al., 2017). A high score on the epistemological understanding test means that a student rather rejects naïve beliefs and chooses the advanced options. On the basis of the outcomes of the validation study, one item was revised, and three were removed entirely (like the subsequent item k201, see Figure 1).

It could be that the correct answers were too easy to recognize, as all subitems were answered correctly by a large majority of the students (86% to 96%). Additionally, the advanced

There is only one way to look at history.

I agree, because history happened exactly as it is written in the history book.
"Objectivist" (96%)

I don't agree, because history is always completely subjective.
"Subjectivist" (86%)

I don't agree, because there can be different perspectives on the same event.
"Criterialist" (98%)

Note: Explanation of epistemological belief in italics (not provided in original test).

Figure 1: Item k0201 with percentages of subitems solved correctly in the validation study.

option was not too different from the naïve options. In any case, the item did not allow to differentiate very well between the students in terms of their epistemological understanding ($a = 0.35$) and was therefore excluded from the test inventory.

The final inventory contained 12 CMC items, each offering three to four subitems for selection to indicate agreement or disagreement. One task with five CMC items consisted of statements about the nature of history itself, unrelated to a specific topic (k02); the other one had statements more closely related to the intervention’s topic of eyewitness accounts and transformation time (k03).

3.3 Test inventory on reconstruction and deconstruction abilities

The test assessing methodological competencies consisted of a selection of $k = 27$ items, organized into five different tasks. Unlike the epistemological test, this methodological test displayed more diversity in task formats, incorporating visuals, such as comic strips, cartoons, and interview-style sources. For an overview of the tasks, refer to Table 2.

Table 2: Overview of the methodological test’s tasks and items included in the final test inventory

Task	Description of task	Item description
k01	Three cartoons needed to be analyzed in terms of their core messages and subsequently matched with corresponding statements.	Eight MC items, each featuring a core message that needed to be associated with one of the cartoons or with none.
k04	A comic strip reflecting on a class trip to the GDR analyzed for its underlying messages.	One CMC item was provided, offering five options for selection.
k05	Impacts of two magazine covers on stereotypes related to Bavaria and East Germany had to be explained on the basis of their covers.	Two CMC items with five to six options from which the correct one(s) had to be chosen.
k06	Two interviews containing opposing narratives about the GDR needed to be matched to corresponding messages.	Four MC items, each featuring a message that needed to be assigned to the correct interview or none.
z	Associated quotes from the oral history interviews of the “Generation 1975” project with either the East or West interview partner.	Twelve MC items with quotes that needed to be matched with the interviewee’s background.

The tasks required students to either deconstruct or reconstruct narratives using disciplinary criteria by Rösen (1989, 2017). For example, in task format “z”, students mostly needed to assess the empirical plausibility of statements to match the quotes of eyewitnesses to their background on the basis of an evaluation of the past information proposed. Task format k05 had students analyze magazine covers for the norms and values conveyed, targeting normative plausibility. In task format k06, students needed to derive and match narratives to the interviewees of interviews they read, assessing narrative plausibility. A high score means that the student applies the disciplinary criteria when dealing with historical narratives to a greater extent.

Based on the outcomes of the validation study, the items most significantly affected by revision originated from a task format “z”, where students needed to analyze quotes from eyewitnesses (initially $k = 24$ items during validation, subsequently reduced to $k = 12$ items). Table 3 presents a sample item from the original version of the task and the corresponding revised version thereof. The original version’s items exhibited negative item discrimination ($-1.72 \leq a \leq -1.50$; $SE = 0.21$, $N = 353$), which means students with higher methodological abilities were less likely to respond with the correct answer. The revised version of the item yielded a statistically significant positive item discrimination ($a = 1.98$, $SE = 0.08$, $N = 2.279$).

Table 3: Task format z with sample item z11 in its original and revised versions

Original Version (N = 12*2 items)	Refers to... (z111)		Prejudice experienced by... (z211)	
	Past	Present	Oneself	Others
Sometimes you get the impression that if you reveal yourself as an East German, you are still looked down upon by some West Germans.		X	X	
Revised Version (N = 12 items)	Statement made by interviewee from the...			
			East	West
Sometimes you get the impression that when you reveal your identity, you are still looked down upon by some people in the other Germany (z11)			X	

In the validation study, out of the 24 items, 15 displayed insufficient item properties ($\alpha < 0.5$), six had correct response rates of less than 10%, and eight had over 9% missing responses. Upon closer examination of students' answers, it became evident that the task's structure (two items to be answered for a single quote) and the task's ambiguous instructions contributed to these results. Therefore, the task's instructions were completely overhauled, resulting in a 50% reduction in the number of MC items.

3.4 Test employment in the intervention study

The aforementioned intervention study was structured as a randomized controlled field trial, where teachers and their respective students were randomly assigned to either the intervention or control groups. The teachers taught the intervention or business as usual, depending on the group (in-person, video, or control). The ninth-grade students completed the tests. Data were collected between May and July 2022, both before and after the three-lesson unit (or during the same period for the control groups), with each test session lasting 90 min, conducted by test administrators. Both a pretest and a posttest were administered, using the same set of outcome measures. In addition to the newly developed tests, the outcome measures encompassed additional tests, including one designed to assess historical knowledge pertaining to the transformation period and former East Germany (newly developed) and three sample tasks drawn from the HiTCH inventory item pool (hereafter: HiTCH tasks), chosen in accordance with psychometric criteria (Trautwein et al., 2017). The three tasks, two targeting methodological competencies and one focusing on the subject matter, spanned a total of 27 items. Whereas the two HiTCH tasks on methodological competencies were very similar to the newly developed ones on these competencies, the HiTCH tasks did not specifically target epistemological principles. Additionally, inventories querying the Relevance of History, Expectancy and Value of History were applied. The RHMS (Van Straaten et al., 2018), spanning 24 items, measures students' experience of the relevance of history for their own identity (Building a personal identity, e.g., "History affects the way I behave"), their own historicity and how people in the past have dealt with enduring problems of human life (Understanding the human condition, e.g., "History enables us to imagine what the world might look like later on"), and the society they live in (Becoming a citizen, e.g., "History makes me understand the news better."). We assessed students' emotional and volitional effects based on the expectancy-value theory of Eccles et al. (1983), employing three items each on value components utility, cost, intrinsic, and attainment value and four items on students' beliefs about their own academic ability in the subject of history (i.e., expectancy; see Table A2 in the Appendix for further information). Students' characteristics and demographics were collected with the pretest. Table 4 presents descriptive statistics for the tests. For all other measures, see Tables A1 and A2 in the Appendix. The final sample from the classes of the 50 participating teachers included 1,301 students who completed the test at least once (at pretest or

posttest). The average age of the students was 15 years ($SD = 0.63$, $n = 1,193$), with 49% identifying as female and 94% indicating German as their primary language spoken at home.

3.5 Statistical analysis

The psychometric quality of the final inventory with $k = 39$ items was evaluated using data gathered from $N = 1,301$ ninth-grade students in the intervention study, following a procedure analogous to the one used in the validation study. In line with the preregistered plan for the intervention study, data from all students who participated in a minimum of one measurement occasion were included in the analysis. For one class, data from the pretest and posttest could not be matched; therefore, data from students at pretest were missing at posttest and vice versa, resulting in $N = 1,317$ cases in total. Statistical significance was set at the $p < .05$ level. Following another model assessment (RQ1), where only items with sufficient psychometric properties were included in the models using the final sets of items, their correlations with other inventories used in the intervention study were explored (RQ2). Lastly, regression models were estimated with the scores from both newly developed tests predicted by student characteristics (RQ3). We briefly describe these analyses in the upcoming paragraphs. See the Appendix for additional details and the statistical software code.

To address RQ1, the data from both the pretest and posttest were initially subjected to an analysis using the 2PL models. The students were given the same set of items at pretest and posttest. Implying that the item parameters are identical across different time points, we applied the so-called virtual persons approach: treating items that were answered by the same individual at pretest and posttest as answered by two distinct “virtual” participants who responded to the items ($N = 2,634$). This allows us to assess the items in a much larger sample, deriving more precise item parameter estimates (see De Ayala, 2009). Items that did not exhibit positive and statistically significant discrimination were removed from the set of items used in the final model. This was done because such items do not seem to be adequate indicators of the latent variable representing the construct and are of minor relevance regarding the reliability of point estimates for person ability (see De Ayala, 2009). Estimates were based on the remaining sets of items that met the inclusion criteria ($a > 0$ and $p < .05$). Point estimates of person ability were obtained as weighted likelihood estimates (WLEs; Warm, 1989). We report WLE person separation reliability (WLE PSR; Andrich, 1982), representing the proportion of variance in the WLEs that could not be attributed to measurement error, as an indicator for score reliability (RQ1). For RQ2 and RQ3, models were estimated in the structural equation modeling (SEM) framework, based on the scores obtained from the unidimensional 2PL models from RQ1. This allows us to relate the scores (and covariates) to each other and to account for measurement error (see B. Muthén, 2002)—both of which are very important here as the scores and other variables are confounded and all tests are subject to a certain degree of measurement error. We accounted for missing data and the structure of the data (students nested in teachers) by relying on cluster-robust full information maximum likelihood (FIML) estimation. To examine latent correlation patterns, a four-dimensional 2PL model with all ability tests (methodology, epistemology, factual knowledge, and the HiTCH tasks) was computed. In both instances, correlation coefficients were estimated with cluster-robust standard errors. In order to assess potential differences in the correlations between the variables, we added model constraints to our SEM (i.e., we constrained the correlations to be equal). Chi-square tests of model fit were employed to determine whether the respective model with equality constraint showed a statistically significant lower model fit than the unconstrained model. For RQ3, we conducted blockwise multiple regression analyses to investigate the variance within the newly developed test scores explained by various types of student characteristics. We did so in a stepwise approach, where the three blocks were added to the equation one after another. Student grades were added first (first block), followed by reading speed and cognitive skills (second block), and finally age, gender, and the number of books at home (third block). All variables, except for those that were dummy-coded, were z-standardized, which means that the beta weights represented the change in standard deviations in the outcome variables if the predictor were to change by one standard deviation while controlling for all other variables (Kelley & Holden, 2013).

4. Results

4.1 Missing responses

The percentage of item nonresponse for methodological test ranged from 1.0% to 7.1%. In the case of the epistemological test, the range was from 2.1% to 5.7%, with the highest rates occurring toward the end of the respective set.

4.2 Test statistics

No items were excluded from the epistemological test, resulting in the final set comprising 12 items. The item difficulties ranged from -3.69 to 2.24 ($M = 0.25$). The most challenging items had a correct answer rate of only 10%, whereas the easiest items were answered correctly by 96% of respondents ($M = 44%$). The WLE PSR was .56. In task k02, which contained abstract statements, only two out of five items exhibited desirable item discrimination values of $a \geq 0.8$ (cf. De Ayala, 2009). In task k03, which offered statements closely related to the intervention's subject, five of the seven items had such item discrimination values. Regarding the methodological test, one item (z07, $a = -0.18$) was excluded due to predefined criteria. Overall, tasks that contained less lengthy texts (k01, k06, and z) tended to contain items with better discrimination ability than tasks with lengthy texts (k05) or multi-panel comic strip (k04). The final model included the remaining 26 items, with item difficulties ranging from -3.14 to 2.77 ($M = -1.01$). The percentage of correct answers to these items ranged from 6% to 92% ($M = 67%$), and the WLE PSR was .70. For a summarized view, see Table 4. The distribution of the ability scores is presented in the Appendix.

Table 4: Item and scale statistics for the achievement tests from the intervention study

Tests	k of inventory		Item and scale statistics				
			Solved (in %)		Discrimination a		WLE PSR
	Tasks	Items	Min	Max	Min	Max	
Methodological test	5	26	6	92	0.22	1.97	.70
Epistemological test	2	12	10	96	0.23	1.66	.56
HiTCH tasks	3	27	36	93	0.36	1.57	.73
Test of factual knowledge	3	22	15	95	0.17	1.74	.68

Note. Estimates are based on 2PL models including the final item selection ($2,273 \leq n \leq 2,279$).

4.3 Best-Performing items

Selected by psychometric criteria, a few items in both inventories deserved a closer look. For each test, we present three items (see Table 6) that ranged from easy to hard and exhibited good discrimination ($0.8 \leq a \leq 2.5$; De Ayala, 2009). Test information can be optimized by selecting items with rather large discriminations and difficulties that align with the expected ability distribution of the target population (De Ayala, 2009), in our case, 9th graders. Larger test information corresponds to increased reliability of point estimates for person ability (e.g., WLE PSR). For the

methodological test, there was an interesting pattern. Except for the easy item, the other two dealt with how the caricature mocked a narrative, where the correct answer was that none of the caricatures conveyed the messages provided in k0105 and k0108.

4.4 Correlations with other measures

Correlations (all p s < .001) are depicted in Table 5 and refer to the association between the two ability tests in each case, while the remaining ones were controlled for. The association between the manifest ability scores from the two newly developed tests was relatively moderate ($r = .38$). Notably, the methodological test ($0.56 \leq r \leq 0.57$) exhibited stronger correlations with both the HiTCH tasks and the knowledge test than the epistemological test ($0.31 \leq r \leq 0.38$) did ($\chi^2(1) = 57.079$, Scaling Correction Factor (SCF) = 0.767, and $\chi^2(1) = 88.943$, SCF = 0.867, both $p < .001$). Correlations between the methodological test and the HiTCH tasks regarding factual knowledge were not statistically significantly different ($\chi^2(1) = 2.084$, SCF = 1.283, $p = .149$). Correlations between the new tests and motivational variables (see Table A3 in the Appendix) were generally positive and low ($.15 \leq r \leq .36$). Regarding the RHMS, both tests had the highest positive correlation with the subscale “becoming a citizen” ($.29 \leq r \leq .36$). Both newly developed tests demonstrated small positive correlations with the dimension of perceived low cost of history (e.g., “History lessons in school cost me a lot of energy”, recoded to match the interpretation of the other scales). All correlations were statistically significant ($p < .001$).

Table 5: Correlations between the ability tests at posttest

	Epistemological test	Methodological test	HiTCH tasks
Methodological test	.38 (.55)		
HiTCH tasks	.38 (.54)	.57 (.79)	
Test of factual knowledge	.31 (.48)	.56 (.85)	.53 (.75)

Note. Correlation coefficients for manifest scores and latent variables (in parentheses) at posttest ($n = 1,071$) with robust standard errors. All coefficients were statistically significant ($p < .001$).

Table 6: Selection of the best items from both tests with difficulty level category

Test	Difficulty level	Item	a	b	Translated item as employed in the intervention study with the correct solution								
Epistemological test	Hard	k0307	1.53	1.24	<p>The history of the GDR and divided Germany is still highly relevant today.</p> <p><input checked="" type="checkbox"/> I agree because one can learn about today's conditions from the history of division.</p> <p><input type="checkbox"/> I disagree because the history of division does not help to explain current conditions. Knowledge about the past does not lead to a better understanding of the present.</p> <p><input checked="" type="checkbox"/> I agree because many contemporary societal conflicts today still have to do with or are justified by Germany's past.</p> <p><input type="checkbox"/> I disagree because one cannot solve today's problems with knowledge about the past.</p>								
	Medium	k0303	1.66	0.10	<p>In class, one should learn that there are many different but entirely justified perspectives on a historical event such as the German reunification.</p> <p><input type="checkbox"/> I disagree because there is only one truth.</p> <p><input checked="" type="checkbox"/> I agree because, for example, East Germans have had different experiences than West Germans.</p> <p><input type="checkbox"/> I disagree because I believe that East and West German students of my age should think similarly about the German reunification.</p> <p><input checked="" type="checkbox"/> I agree because one's own view of history depends greatly on how, where, and when one grew up.</p>								
	Easy	k0203	1.14	-3.69	<p>One should consider various historical sources before forming a judgment.</p> <p><input type="checkbox"/> I disagree because one usually has enough information from just one source to form a judgment.</p> <p><input type="checkbox"/> I disagree because different sources always report the same thing for the same event because everyone experienced the same thing.</p> <p><input checked="" type="checkbox"/> I agree because different sources can illuminate different aspects of an event.</p>								
Methodological test	Hard	k0105	1.03	0.41	<p>Central message</p> <p>Caricature...</p> <p>West Germans are appreciative of East Germans and acknowledge their life achievements.</p> <table border="1" style="width: 100%; text-align: center;"> <tr> <td>1</td> <td>2</td> <td>3</td> <td>none</td> </tr> <tr> <td></td> <td></td> <td></td> <td>X</td> </tr> </table>	1	2	3	none				X
	1	2	3	none									
				X									
Medium	k0108	1.18	-0.19	<p>Central message</p> <p>Caricature...</p> <p>In the GDR, there were good Spreewald pickles.</p> <table border="1" style="width: 100%; text-align: center;"> <tr> <td>1</td> <td>2</td> <td>3</td> <td>none</td> </tr> <tr> <td></td> <td></td> <td></td> <td>X</td> </tr> </table>	1	2	3	none				X	
1	2	3	none										
			X										
Easy	z11	1.97	-3.09	<p>Statement made by interviewee from the...</p> <p>East</p> <p>West</p> <p>Sometimes you get the impression that when you reveal your identity, you are still looked down upon by some people in the other Germany.</p> <table border="1" style="width: 100%; text-align: center;"> <tr> <td>East</td> <td>West</td> </tr> <tr> <td>X</td> <td></td> </tr> </table>	East	West	X						
East	West												
X													

4.5 Prediction of ability scores

Table 7 provides an overview of the blockwise regression for both new tests (i.e., to what extent the score can be explained by students' grades, their basic cognitive abilities, and their background characteristics). Regarding magnitude, the strongest statistically significant predictors of both test scores were the students' grades in history and German ($-0.260 \leq \beta \leq -0.095$) and their cognitive ability ($0.178 \leq \beta \leq 0.260$). Moreover, the number of books at home predicted both scores ($0.089 \leq \beta \leq 0.115$), and reading speed predicted the methodological test score ($\beta = 0.103$).

Table 7: Blockwise regression on ability scores from newly developed tests at posttest

Dependent variable: tests on...		Epistemological test (1)			Methodological test (2)		
Blocks and predictors contained		β	SE	<i>p</i>	β	SE	<i>p</i>
I: Student achievement in the form of grades	History	-0.122**	0.045	.007	-0.260***	0.042	.000
	German	-0.148***	0.042	.000	-0.095*	0.043	.026
	Math	0.017	0.041	.685	0.030	0.033	.357
Variance explained							
(1) $R^2 = 9\%$							
(2) $R^2 = 16\%$							
II: Students' basic abilities	Reading speed	0.037	0.031	.236	0.103**	0.033	.002
	Cognitive ability	0.178***	0.037	.000	0.260***	0.036	.000
Variance explained							
(1) $\Delta R^2 = 4\%$							
(2) $\Delta R^2 = 9\%$							
III: Students' characteristics	Age	-0.050	0.039	.199	-0.031	0.036	.387
	Female ^a	0.074	0.075	.322	-0.092	0.059	.119
	Diverse ^a	0.082	0.179	.649	0.180	0.214	.400
Variance explained							
(1) $\Delta R^2 = 1\%$	Books	0.089**	0.027	.001	0.115***	0.025	.000
(2) $\Delta R^2 = 2\%$							

Note. Standardized regression coefficients from multiple regression analyses containing all predictor variables ($n = 1,315$), including robust standard error estimation. The percentage of explained variance refers to the variance that is explained by adding variables from each respective block.

^aGender was dummy coded for the three categories (female, male, and diverse).

Overall, the model explained 26% of the variance on the methodological test and 14% on the epistemological test. Because the model was estimated based on manifest scores, the outcome included measurement error that could not be explained by any given predictor. Taking the reliability of the scores into consideration, the model explained 37% of the "true score" variance on the methodological test and 25% on the epistemological test.

5. Discussion

In this study, we empirically tested newly developed standardized items to measure aspects of historical thinking with over 1,600 students. One test aimed to capture students' methodological competencies (i.e., their abilities to either deconstruct or reconstruct historical narratives using disciplinary methods), and the other test assessed students' epistemological understanding (i.e., their views on the nature and justification of historical knowledge). In Step 1, we conducted a validation study with 354 students to test the initial set of 58 items. In Step 2, the finalized tests with 12 CMC items on epistemological understanding and 27 MC and CMC items on methodological competencies were employed in an intervention study on transformation time involving 1,301 students. We wanted the test to be engaging for the students, leading them to respond to all items. Our main research questions concerned the psychometric performance and reliability of the tests (RQ1) and aspects of validity: testing a priori assumptions about the correlation of the tests with other measures of student ability (RQ2) and the extent to which student background characteristics predict test scores (RQ3).

Results from this large-scale study demonstrate the strengths of both, the methodological test and the epistemological test. First, we attribute the high response rates in both tests (1.0% to 7.1% item nonresponse) to their context, German post-1990 transformation, and the materials used, as we believe that students related to them. Moreover, the items' abilities to distinguish between students in terms of the competencies based on their performance met the predefined criteria for 38 items, with only one item excluded due to insufficient item discrimination. Both tests showed significant, mostly low, positive correlations with the value that students attribute to the subject of history and their self-concept in history. This also applies to the students' perceived relevance of history, where the highest correlations were obtained for the "Becoming citizen" scale. As it included a self-assessment of their approach to forming opinions and justifying judgments, this could be related to aspects in both, the methodological test and the epistemological test. Both tests also exhibited similar performance when regressed on students' performance and characteristics. Concerning school grades, both the history and German grades predicted the test scores, whereas math grades did not. Whereas basic cognitive abilities appeared to be a strong predictor of both test scores, reading speed only predicted the methodological test score. To a lesser extent, the number of books at home also predicted the results of both tests.

Differences between the tests were observed in the item difficulties, which appeared to be higher on the epistemological test than on the methodological test, mirroring the higher rate of CMC items on the epistemological test compared with the methodological one. The tests also differed in measurement precision, with the epistemological one exhibiting low reliability, whereas acceptable values were obtained for the test on methodological competencies. Moreover, correlational patterns suggested that constructs measured in the small selection of items from the HiTCH test were more strongly related to the methodological test than the epistemological one. The relationships of the methodological test and the HiTCH tasks with other tests were similar, while the epistemological test showed weaker correlations with both factual knowledge and the HiTCH tasks. The 27 items selected from the HiTCH test for this study leaned more toward assessing methodological competencies than epistemological principles. Therefore, it was not surprising that the correlation between the HiTCH tasks and the epistemological test was rather weak.

Overall, both tests faced the challenging goal of adequately representing complex constructs, incorporating the historical context without overly influencing the assessment, and limiting the extent to which the test assesses reading skills. In the following paragraphs, we examine the results of both tests with regard to these challenges.

The methodological test adopted a promising approach in its tasks, primarily focusing on one aspect of disciplinary criteria to approach narratives from a specific methodological perspective. It individually targeted the students' methodological competencies regarding normative, narrative, or empirical plausibility. The complexity of methodological competencies was represented by dividing them into small portions that the students had to address in the tasks. Correlational patterns indicated that there was still a considerable relationship with factual knowledge, a pattern also observed for the HiTCH tasks (that worked with historical contexts other than that of the intervention in which factual knowledge was tested). Regarding reliability, the test exhibited acceptable values with potential for improvement. Positive evidence of the validity of the construct was provided by the moderate correlation with the HiTCH tasks, considering

the overlap in methodological tasks. Furthermore, history grades and cognitive ability emerged as the strongest predictors of the test score, whereas reading and text comprehension skills played subordinate roles. The items that performed well in terms of item discrimination in the methodological test were mostly those in which non-complex material (i.e., shorter texts or single-panel visuals) had to be analyzed. Additionally, in two of the best-performing items, the correct solution was not to select any of the given materials (i.e., the caricatures).

Assessing epistemological understanding introduced additional challenges, underscoring that it is challenging to adequately represent the complexity of the construct (Stoel et al., 2017). In tackling these challenges, the test aimed to elicit specific conceptions from students, such as distinguishing between the past and history, addressing the high level of abstraction by contextualizing items in concrete topics for most of the items. Results indicated better performance for items containing more concrete topics compared with abstract ones. In comparison with the methodological test, weaker correlations with factual knowledge were obtained, which could mean that the historical context in the statements was not overly influential. However, problems were noted in terms of psychometric quality, especially the low reliability. Furthermore, the final model predicting the test scores, incorporating grades, reading speed, cognitive ability, and students' characteristics could account for only one quarter of the true score variance, leaving much of the variance unexplained. One reason could be that students' performance in specific scenarios may be influenced not only by their overall epistemic beliefs but also by the interplay of controversial topics like the German post-1990 transformation and their family background. In their investigation of another controversial topic, Iordanou et al. (2020) found differences in students' epistemological processing based on the side taken by a historical account: Less mature epistemic beliefs led students to write summaries that only considered the perspective of their own ethnic group.

Considering the notable correlation of the epistemological test with students' self-reported ratings of the relevance of history, the test could be refined further by aligning it with Wiley et al.'s (2020) recommendations for a more direct approach. They proposed that items containing more self-reports about students' explicit epistemic beliefs measure epistemic understanding rather indirectly and require students to have already formed an epistemological understanding. Students' epistemological understanding would get measured more directly by way of, for example, having them act on concrete scenarios (VanSledright & Maggioni, 2016; Wiley et al., 2020). Students' actions (i.e., how they solve the task) should expose how they view and justify their knowledge (VanSledright & Maggioni, 2016) rather than a measure based mainly on self-reported explicit epistemic ideas (Wiley et al., 2020). In our measure, students were primarily confronted with concrete claims but had only a very limited number of response options. Furthermore, there was no impact from their decision described (e.g., that they would have to justify the claim in front of other people). If one decides to keep the items closed-ended, options could still be expanded, or scenarios could be created that resemble natural situations, such as a discussion among five friends on the way home from school, prompting students to select the argument they would contribute to the conversation.

Keeping in mind the challenges that came with the competencies' complexity, the strengths of both tests lay in specifically targeting the students' deeper thinking operations one at a time when tapping into aspects of their methodological competencies and epistemological understanding. Considering the results of this study, task formats and items that worked well in both newly developed tests should therefore be reassessed given their strengths: The students showed high response rates in the newly developed items, which are contextually embedded in a topic that remains highly relevant today: the German post-1990 transformation. Moreover, the psychometric properties of the items were investigated based on IRT with a large sample. Our (empirical) results regarding the difficulty of the items and their ability to differentiate between students with different competencies provide clear indications of which of the items and task formats have potential for further development. The task materials and assignments can contribute to and diversify existing standardized tests on specific aspects of these competencies or could each be extended on more aspects that they currently underrepresent. Both could contribute to forming more psychometrically sound standardized tests that are able to measure these complex competencies adequately in the future.

The complex process of historical thinking is neither part of natural psychological development nor easy to learn (Wineburg, 2010). However, it is essential for a democracy to provide students (citizens) with the tools to navigate the challenges the present and future hold and enable them to adequately deal with the multiple perspectives they will encounter (Körber & Lenz, 2014). With history education being a central subject for the cultivation of these his-

torical competencies, it is paramount to empirically investigate what goes on in the classroom. Large-scale interventions, such as the outlined study employing eyewitnesses' perspectives, are equipped to assess how and whether students benefit from the lesson units and improve their historical thinking. Considering the gap in valid and robust assessments of historical competencies that these settings and, ultimately, research on students' historical learning need (Körber & Meyer-Hamme, 2015), the present study on newly developed standardized measures attempted to present items that can contribute to closing this gap in the long run.

To cite this article

Totter, K., Wagner, W., & Bertram, C. (2024). Standardized assessment of historical thinking competencies in an intervention study using perspectives on German history. *Historical Thinking, Culture, and Education*, 1(1), 50–99. <https://doi.org/10.12685/htce.1382>

Peer review

This article has been peer reviewed through the journal's standard double-blind peer review, where both the reviewers and authors are anonymized during review.

Acknowledgements

The study would not have been possible without the help of the dedicated teams at the Universities of Tübingen and Konstanz. We extend our gratitude to all research and administrative staff, student helpers, and the teachers and students who participated. Special thanks go to the ZSL (Zentrum für Schulqualität und Lehrerbildung Baden-Württemberg), where Stefan Schipperges and Carsten Arbeiter played pivotal roles in designing the intervention and providing teacher training. We also thank our English language editor, Jane Zagorski. During the preparation of this work, the author(s) used ChatGPT for English language editing only. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The intervention study on eyewitnesses was funded by the Deutsche Forschungsgemeinschaft (DFG) under project number 416879869 (see <https://gepris.dfg.de/gepris/projekt/416879869?language=en>). The applicants' respective numbers are TR553/11-1 (Ulrich Trautwein), BE6291/2-1 (Christiane Bertram), and WA3160/2-1 (Wolfgang Wagner).

ORCID iD

Katharina Totter  <https://orcid.org/0009-0004-6078-5142>

Wolfgang Wagner  <https://orcid.org/0000-0001-9781-4630>

Christiane Bertram  <https://orcid.org/0000-0003-4520-8692>

References

- Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, 9(1), 95–104. <https://www.rasch.org/erp7.htm>
- Angvik, M., Von Borries, B., & Körber, A. (Eds.). (1997). *Edition Körber-Stiftung. Youth and history: A comparative European survey on historical consciousness and political attitudes among adolescents*. Körber-Stiftung.
- Arens, A. K., Möller, J., & Watermann, R. (2016). Extending the internal/external frame of reference model to social studies: Self-concept and achievement in history and politics. *Learning and Individual Differences*, 51, 91–99. <https://doi.org/10.1016/j.lindif.2016.08.044>
- Auer, M., Gruber, G., Mayringer, H., & Wimmer, H. (2005). *SLS 5-8: Salzburger Lese-Screening für die Klassenstufen 5-8*. Hogrefe.
- Barzilai, S. & Weinstock, M. (2015). Measuring epistemic thinking within and across topics: A scenario-based approach. *Contemporary Educational Psychology*, 42, 141–158. <https://doi.org/10.1016/j.cedpsych.2015.06.006>
- Bertram, C. (2020). “Generation 1975 – Mit 14 ins neue Deutschland” – Blick vom Osten und Westen in die deutsche Teilungsgeschichte. *Bürger & Staat*, 1(2), 81–89.
- Bertram, C., Wagner, W., & Trautwein, U. (2017). Learning Historical Thinking With Oral History Interviews: A Cluster Randomized Controlled Intervention Study of Oral History Interviews in History Lessons. *American Educational Research Journal*, 54(3), 444–484. <https://doi.org/10.3102/0002831217694833>
- Bertram, C., Weiss, Z., Zachrich, L., & Ziai, R. (2021). Artificial intelligence in history education. Linguistic content and complexity analyses of student writings in the CAHist project (Computational assessment of historical thinking). *Computers and Education: Artificial Intelligence*, 100038. <https://doi.org/10.1016/j.caeai.2021.100038>
- De Ayala, R. J. (2009). *The theory and practice of item response theory. Methodology in the social sciences*. The Guilford Press.
- De Leeuw, E., Hox, J., & Dillman, D. (Eds.). (2008). *International Handbook of Survey Methodology*. Psychology Press, Taylor & Francis.
- Eccles, J., Adler, T. E., Futtermann, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, Values and Academic Behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 76–138). W.H. Freeman.
- Eklöf, H. (2010). Skill and will: test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, 17(4), 345–356. <https://doi.org/10.1080/0969594X.2010.516569>
- Enders, C. K. (Ed.). (2010). *Methodology in the social sciences. Applied missing data analysis*. Guilford Press.
- Ercikan, K. & Seixas, P. (2015). Issues in Designing Assessments of Historical Thinking. *Theory into Practice*, 54(3), 255–262. <https://doi.org/10.1080/00405841.2015.1044375>
- Gaspard, H., Brisson, B. M., Häfner, I., Dicke, A.-L., Flunger, B., Parrisius, C., Nagengast, B., & Trautwein, U. (2019). *Motivationsförderung im Mathematikunterricht (MoMa 1.0): Skalendokumentation Schülerfragebogen*. Universität Tübingen / Hector-Institut für Empirische Bildungsforschung.
- Gaspard, H., Dicke, A.-L., Flunger, B., Schreier, B., Häfner, I., Trautwein, U., & Nagengast, B. (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology*, 107(3), 663–677. <https://doi.org/10.1037/edu0000003>

- Gelman, A. & Hennig, C. (2017). Beyond Subjective and Objective in Statistics. *Journal of the Royal Statistical Society Series a: Statistics in Society*, 180(4), 967–1033. <https://doi.org/10.1111/rssa.12276>
- González, J., Tuerlinckx, F., De Boeck, P., & Cools, R. (2006). Numerical integration in logistic-normal models. *Computational Statistics & Data Analysis*, 51(3), 1535–1548. <https://doi.org/10.1016/j.csda.2006.05.003>
- Goßmann, F. (2018). *Measuring Cultural Capital in the NEPS*. <https://doi.org/10.5157/NEPS:SP48:1.0>
- Großbölting, T. (2020). Wiedervereinigungsgesellschaft: Aufbruch und Entgrenzung in Deutschland seit 1989/90. *Schriftenreihe / Bundeszentrale für politische Bildung: Vol. 10610*. BpB.
- Guo, J., Hu, X., Marsh, H. W., & Pekrun, R. (2022). Relations of epistemic beliefs with motivation, achievement, and aspirations in science: Generalizability across 72 societies. *Journal of Educational Psychology*, 114(4), 734–751. <https://doi.org/10.1037/edu0000660>
- Heller, K. & Perleth, C. (2000). *Kognitiver Fähigkeits-Test für 5.-12./13. Klassen, Revision KFT 5-12+R*. Beltz Test.
- Heppt, B., Olczyk, M., & Volodina, A. (2022). Number of books at home as an indicator of socioeconomic status: Examining its extensions and their incremental validity for academic achievement. *Social Psychology of Education*, 25(4), 903–928. <https://doi.org/10.1007/s11218-022-09704-8>
- Hofer, B. K. & Pintrich, P. R. (1997). The Development of Epistemological Theories: Beliefs About Knowledge and Knowing and Their Relation to Learning. *Review of Educational Research*, 67(1), 88–140. <https://doi.org/10.3102/00346543067001088>
- Iordanou, K., Kendeou, P., & Zembylas, M. (2020). Examining my-side bias during and after reading controversial historical accounts. *Metacognition and Learning*, 15(3), 319–342. <https://doi.org/10.1007/s11409-020-09240-w>
- Kaliski, P., Smith, K., & Huff, K. (2015). The Importance of Construct Validity Evidence in History Assessment: What Is Often Overlooked or Misunderstood. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 195–205). Routledge.
- Kelley, K. & Holden, J. (2013). Multiple Regression. In T. Teo (Ed.), *Handbook of Quantitative Methods for Educational Research* (pp. 71–102). Brill.
- Körber, A. (2011). German History Didactics: From Historical Consciousness to Historical Competencies - and Beyond? In H. Bjerg, C. Lenz, & E. Thorstensen (Eds.), *Time, meaning, culture. Historicising the uses of the past* (pp. 145–164). transcript.
- Körber, A. (2016). Translation and its discontents II: a German perspective. *Journal of Curriculum Studies*, 48(4), 440–456. <https://doi.org/10.1080/00220272.2016.1171401>
- Körber, A. & Lenz, C. (2014). Introduction. In H. Bjerg (Ed.), *Neuengammer Kolloquien: Vol. 4. Teaching historical memories in an intercultural perspective: Concepts and methods: experiences and results from the TeacMem project* (pp. 21–30). Metropol.
- Körber, A. & Meyer-Hamme, J. (2015). Historical Thinking, Competencies, and their Measurement: Challenges and Approaches. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 89–101). Routledge.
- Körber, A., Schreiber, W., & Schöner, A. (2007). *Kompetenzen historischen Denkens: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik. Kompetenzen: Band 2*. ars una.
- 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). Routledge Taylor & Francis Group.
- Lévesque, S. & Clark, P. (2018). Historical Thinking. In S. A. Metzger & L. M. Harris (Eds.), *The Wiley handbooks in education. The Wiley international handbook of history teaching and learning*, Vol. 12 (pp. 117–148). Wiley Blackwell. <https://doi.org/10.1002/9781119100812.ch5>
- Maggioni, L. (2010). *Studying epistemic cognition in the history classroom: Cases of teaching and learning to think historically* [PhD Thesis]. University of Maryland. <https://drum.lib.umd.edu/bitstreams/b48079d2-159f-451f-974c-00b3c7c8a029/download>
- Maggioni, L., VanSledright, B., & Alexander, P. A. (2009). Walking on the Borders: A Measure of Epistemic Cognition in History. *The Journal of Experimental Education*, 77(3), 187–214. <https://doi.org/10.3200/JEXE.77.3.187-214>
- Muthén, B. (2002). Beyond SEM: General Latent Variable Modeling. *Behaviormetrika*, 29(1), 81–117. <https://doi.org/10.2333/bhmk.29.81>
- Muthén, L. K. & Muthén, B. (1998-2017). *Mplus user's guide: Statistical analysis with latent variables* (8^a ed.). Muthén & Muthén.

- National Center for History in Schools UCLA. (1996). *National standards for history basic edition*. <https://phi.history.ucla.edu/nchs/historical-thinking-standards/>
- Pan, J. & Thompson, R. (2007). Quasi-Monte Carlo estimation in generalized linear mixed models. *Computational Statistics & Data Analysis*, 51(12), 5765–5775. <https://doi.org/10.1016/j.csda.2006.10.003>
- R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.0) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Radinsky, J., Goldman, S. R., & Pellegrino, James, W. (2015). Commentary: Historical Thinking: In Search of Conceptual and Practical Guidance for the Design and Use of Assessments of Student Competence. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 132–142). Routledge.
- Rensmann, L. (2019). Divided We Stand. *German Politics and Society*, 37(3), 32–54. <https://doi.org/10.3167/gps.2019.370304>
- Revelle, W. (2022). *psych: Procedures for Psychological, Psychometric, and Personality Research* (Version R package version 2.3.3) [Computer software]. <https://CRAN.R-project.org/package=psych>.
- Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test Analysis Modules* (Version R package version 4.1-4) [Computer software]. <https://CRAN.R-project.org/package=TAM>
- Rose, N., Von Davier, M., & Nagengast, B. (2016). Modeling Omitted and Not-Reached Items in IRT Models. *Psychometrika*. Advance online publication. <https://doi.org/10.1007/s11336-016-9544-7>
- Rüsen, J. (1989). *Historische Vernunft: Die Grundlagen der Geschichtswissenschaft. Grundzüge einer Historik*. Vandenhoeck & Ruprecht.
- Rüsen, J. (2005). *History: Narration, Interpretation, Orientation. Making Sense of History*. Bergahn.
- Rüsen, J. (2013). *Historik: Theorie der Geschichtswissenschaft*. Böhlau.
- Rüsen, J. (2017). *Evidence and meaning: A theory of historical studies* (D. Kerns & K. Digan, Trans.). *Making Sense of History: Volume 28*. Bergahn.
- Seixas, P. (2017). Teaching Rival Histories: In Search of Narrative Plausibility. In H. Å. Elmersjö, A. Clark, & M. Vinterek (Eds.), *International Perspectives on Teaching Rival Histories* (pp. 253–268). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-55432-1_12
- Smith, M. D. (2017). Cognitive Validity: Can Multiple-Choice Items Tap Historical Thinking Processes? *American Educational Research Journal*, 54(6), 1256–1287. <https://doi.org/10.3102/0002831217717949>
- Smith, M. D. (2018). New Multiple-Choice Measures of Historical Thinking: An Investigation of Cognitive Validity. *Theory & Research in Social Education*, 46(1), 1–34. <https://doi.org/10.1080/00933104.2017.1351412>
- Stoel, G., Logtenberg, A., Wansink, B., Huijgen, T., Van Boxtel, C., & Van Drie, J. (2017). Measuring epistemological beliefs in history education: An exploration of naïve and nuanced beliefs. *International Journal of Educational Research*, 83, 120–134. <https://doi.org/10.1016/j.ijer.2017.03.003>
- Trautwein, U., Bertram, C., Borries, B. von, Brauch, N., Hirsch, M., Klausmeier, K., Körber, A., Kühberger, C., Meyer-Hamme, J., Merkt, M., Neureiter, H., Schwan, S., Schreiber, W., Wagner, W., Waldis, M., Werner, M., Ziegler, B., & Zuckowski, A. (2017). *Kompetenzen historischen Denkens erfassen: Konzeption, Operationalisierung und Befunde des Projekts „Historical Thinking – Competencies in History“ (HiTCH)*. Waxmann.
- Van der Eem, M., Van Drie, J., Brand-Gruwel, S., & Van Boxtel, C. (2023). Students' evaluation of the trustworthiness of historical sources: Procedural knowledge and task value as predictors of student performance. *The Journal of Social Studies Research*, 47(1), 64–76. <https://doi.org/10.1016/j.jssr.2022.05.003>
- Van Drie, J. & Van Boxtel, C. (2008). Historical Reasoning: Towards a Framework for Analyzing Students' Reasoning about the Past. *Educational Psychology Review*, 20(2), 87–110. <https://doi.org/10.1007/s10648-007-9056-1>
- Van Straaten, D., Wilschut, A., & Oostdam, R. (2018). Measuring students' appraisals of the relevance of history: The construction and validation of the Relevance of History Measurement Scale (RHMS). *Studies in Educational Evaluation*, 56, 102–111. <https://doi.org/10.1016/j.stue-duc.2017.12.002>
- VanSledright, B. (2014). *Assessing historical thinking and understanding: Innovative designs for new standards*. Routledge. <https://doi.org/10.1016/j.ssr.2014.02.001>

- VanSledright, B. & Maggioni, L. (2016). Epistemic cognition in history. In J. A. Greene, W. A. Sandoval, & I. Bråten (Eds.), *Educational psychology handbook series. Handbook of epistemic cognition* (pp. 128–146). Routledge.
- Wagner, W. (2020). Item-Response-Theorie (IRT): Messung nicht direkt beobachtbarer Fähigkeiten anhand kategorialer Itemantworten. In G. Weißeno & B. Ziegler (Eds.), *Handbuch Geschichts- und Politikdidaktik* (pp. 1–17). Springer Fachmedien Wiesbaden.
- Wagner, W., Bertram, C., & HiTCH-Konsortium. (2023). HiTCH II: Was wir aus der Weiterentwicklung des HiTCH-Tests über historische Kompetenzen lernen. In M. Waldis & M. Nitsche (Eds.), *Geschichtsdidaktik heute: Vol. 13. Geschichtsdidaktisch intervenieren* (pp. 238–258). hep.
- Waldis, M., Hodel, J., Thünemann, H., Zülsdorf-Kersting, M., & Ziegler, B. (2015). Material-Based and Open-Ended Writing Tasks for Assessing Narrative Competence among Students. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 117–131). Routledge.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427–450. <https://doi.org/10.1007/BF02294627>
- Wiley, J., Griffin, T. D., Steffens, B., & Anne Britt, M. (2020). Epistemic beliefs about the value of integrating information across multiple documents in history. *Learning and Instruction*, 65, 101266. <https://doi.org/10.1016/j.learninstruc.2019.101266>
- Wineburg, S. 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. <https://doi.org/10.1037/0022-0663.83.1.73>
- Wineburg, S. S. (2010). Historical Thinking and other Unnatural Acts. *Phi Delta Kappan*, 92(4), 81–94. <https://doi.org/10.1177/003172171009200420>

Appendix

A1. Additional tables and figures

Table A1: Item and scale statistics for sample demographics and student characteristics at pretest

Variable	Additional information	Descriptives	N
Grade History		$M = 2.37, SD = 0.93$	1,085
Grade German	Received in the last school year, ranging from 1 (very good) to 6 (insufficient), with 7 (I don't know) set to missing	$M = 2.46, SD = 0.83$	1,090
Grade Math		$M = 2.57, SD = 1.08$	1,092
Cognitive ability	KFT 5-12+R (Heller & Perleth, 2000). Visual (figurative) thinking, based on 2PL model deriving ability scores for each student analogous to all other 2PL models described in this paper	$M = -0.04, SD = 1.01$	1,189
Reading speed	SLS-A1/SLS-A2 (Auer et al., 2005). Reading and distinguishing between meaningful and nonsensical sentences, sum score	$M = 46.21, SD = 9.17$	1,193
Age	Answered in open format	$M = 14.82, SD = 0.63$	1,193
Number of books at home	See, e.g., Goßmann (2018). Ranging from 1 to 6 (< 11, 11-25, 26-100, 101-200, 201-500, > 500)	$M = 4.67, SD = 1.33$	1,157
Gender	Closed-response format	49.3% female, 48.7% male, 2.0% diverse	1,182

Note. Saturated SEM with FIML and manifest variables for RQ3 (all possible correlations estimated) taking the cluster structure of the data into account (TYPE=Complex) in Mplus ($n = 1,315$).


```

(...)
MISSING=.;
CLUSTER=idgleh;
USEV = t3mewle, t3epwle, t1d01, t1d05, t1v0301, t1v0302, t1v0303, t1fwle,
t1l, female, divers;
  DEFINE:
    IF (t1d03 EQ 1) THEN female = 1;
    IF (t1d03 EQ 2) THEN female = 0;
    IF (t1d03 EQ 3) THEN female = 0;
    IF (t1d03 EQ 3) THEN divers = 1;
    IF (t1d03 EQ 2) THEN divers = 0;
    IF (t1d03 EQ 1) THEN divers = 0;
  ANALYSIS:
    TYPE=COMPLEX;
  MODEL:
t3mewle t3epwle t1d01 t1d05 t1v0301 t1v0302 t1v0303 t1fwle t1l female
divers WITH
t3mewle t3epwle t1d01 t1d05 t1v0301 t1v0302 t1v0303 t1fwle t1l female
divers
  OUTPUT: SAMPSTAT STANDARDIZED;

```

Mplus code for table A1

Table A2 : Descriptive statistics for the subscales relevance of history and expectancy and value of history

Scale outcome measures	Short description	N _{items}	M	SD	α^a
Relevance of history (RHMS, translated from Van Straaten et al., 2018)	Building identity	7	2.24	0.56	.78
	Understanding the human condition	5	2.55	0.53	.75
	Becoming a citizen	12	2.88	0.44	.82
Expectancy-value beliefs in the subject of history ^b	Attainment value	3	2.68	0.72	.76
	Utility	3	2.56	0.66	.77
	Low cost	3	2.82	0.68	.81
	Intrinsic value	3	2.87	0.79	.94
Expectancy (Self-concept)		4	2.84	0.71	.90

Note. Responses from 1 (does not apply at all) to 4 (fully applies). Means and standard deviations based on saturated SEM with FIML and manifest variables of RQ2 (all possible correlations estimated) taking the cluster-structure of the data into account (TYPE=Complex) in Mplus ($n = 1,071$).

a Cronbach's α based on combined pretest and posttest data using the virtual person approach ($2,193 \leq n \leq 2,247$).

b Shortened and adapted for the subject of history from Gaspard et al. (2015) & Gaspard et al. (2019).

Table A3: Correlation coefficients with motivational variables at posttest

		Epistemological test	Methodological test
Relevance of History	Building identity	.21	.15
	Human condition	.25	.19
	Becoming a citizen	.36	.29
Motivation (Value)	Attainment value	.24	.23
	Utility	.21	.17
	Intrinsic value	.21	.26
	Low cost	.20	.23
(Expectancy)	Self-concept	.21	.32

Note. All correlation coefficients estimated with robust standard errors were statistically significant ($p < .001$), $n = 1,071$. For the Mplus code, see the section below.

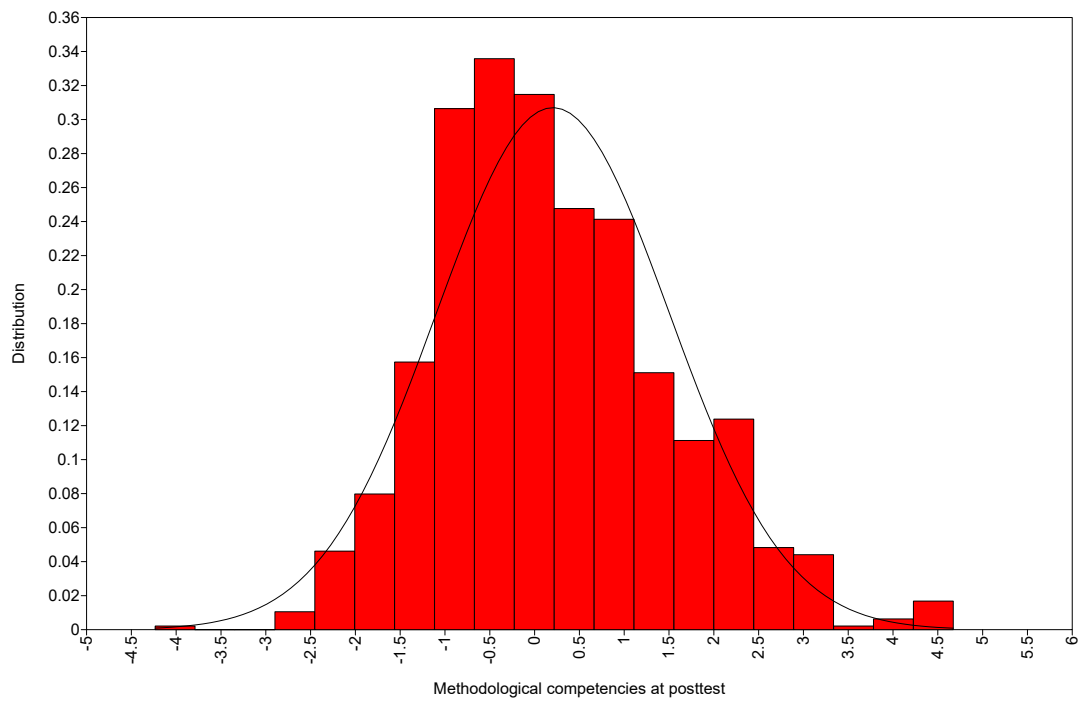
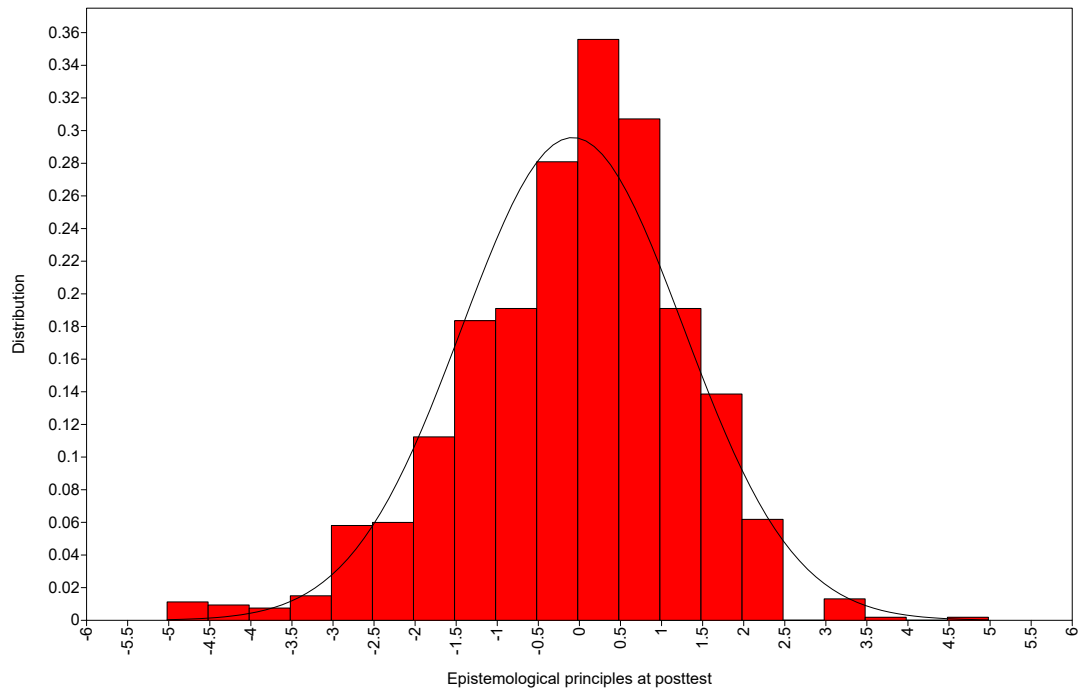


Figure A1: Distribution of point estimates of person ability, obtained as weighted likelihood estimates in the intervention sample at posttest

Note. Graphic derived from Mplus via TYPE = PLOT2 (1,068 ≤ n ≤ 1,070).

A2. Statistical analysis description with abbreviated statistical software code

For the validation study, answers to all items were examined in one 2PL model. The model was computed with “tam.mml.2pl” with the default settings, including Quasi Monte Carlo integration (González et al., 2006; Pan & Thompson, 2007), and the maximization steps (M-steps) for item parameter estimation were set to 10. The calculations were performed with R (v4.3.0, R Core Team, 2023) along with the tam package (version 4.1-4, Robitzsch et al., 2022).

```
# Multidimensional Item Response Model in TAM
# IRT Model: 2PL
TAM::tam.mml.2pl(resp = d[, hkompall], control = list(QMC = TRUE, MSteps
= 10, seed = 12345,
progress = FALSE))
```

R code for the 2PL model in the validation study

To compute descriptives for the scales Relevance of History, Expectancy and Value of History, results were computed with the R-package “psych” (version 2.3.3, Revelle, 2022). Mean scores were computed across the respective set of items for each subscale for all cases with valid answers for at least two thirds of the items from the respective set (otherwise the mean score was set to missing). Cronbach’s alpha was computed based on covariances of subscales.

```
# Reliability analysis
psych::alpha(x = dt[, vector])
```

R code for scale assessment (sample)

For RQ1, the models were computed using “tam.mml.2pl” with the default settings, including Quasi Monte Carlo integration (González et al., 2006; Pan & Thompson, 2007), and the maximization steps (M-steps) for item parameter estimation were set to 10. The calculations were performed with R (v4.3.0, R Core Team, 2023) along with the “tam” package (version 4.1-4, Robitzsch et al., 2022).

```
# Multidimensional Item Response Model in TAM
# IRT Model: 2PL
# Methodological Test
TAM::tam.mml.2pl(resp = dt[, vec_menew], pid = dt$idlv1,
control = list(QMC = TRUE, MSteps = 10, seed = 12345, progress = FALSE))
# Multidimensional Item Response Model in TAM
# IRT Model: 2PL
# Epistemological Test
TAM::tam.mml.2pl(resp = dt[, vec_ep], pid = dt$idlv1, control = list(QMC
= TRUE, MSteps = 10, seed = 12345, progress = FALSE))
```

R code for the methodological and epistemological tests’ models

To investigate RQ2 and RQ3, all (regression) models were computed in Mplus (version 8.6, L. K. Muthén & Muthén, 1998-2017) with person ability scores, accounting for the residual covariances between all variables. The issue of missing data was addressed with full information maximum likelihood estimation (FIML; Enders & Bandalos, 2001; L. K. Muthén & Muthén, 1998-2017). Given the clustered structure of the data (students nested within teachers), robust maximum likelihood estimation was employed, incorporating a design-based correction (using the command Type=COMPLEX) to ensure accurate standard error estimates and account for the cluster sampling design.

INPUT INSTRUCTIONS

```
(...)  
! t3mwle refers to methodological test  
! t3epwle refers to epistemological test
```

```

! other variable names refer to subscales in Relevance of History, Value
and
! Expectancy of History
CLUSTER=idgleh;
(...)
ANALYSIS:
  TYPE=COMPLEX;
MODEL:
  t3mewle WITH
  t3iin t3iwi t3inu t3ico t3iex t3ride t3rhum t3rcit t3hiwle t3wwle
t3epwle;
  t3epwle WITH
  t3iin t3iwi t3inu t3ico t3iex t3ride t3rhum t3rcit t3hiwle t3wwle
t3mewle;
  t3iin t3iwi t3inu t3ico t3iex t3ride t3rhum t3rcit t3hiwle t3wwle
WITH
  t3iin t3iwi t3inu t3ico t3iex t3ride t3rhum t3rcit t3hiwle t3wwle;
OUTPUT: SAMPSTAT STANDARDIZED;

```

Mplus code for manifest correlation model

To examine latent correlation patterns, a multidimensional 2PL model was computed in which factor variances were fixed to 1 and all factor loadings (i.e., item discriminations) were freely estimated for each ability test (methodology, epistemology, factual knowledge, and HiTCH tasks).

```

(...)
MISSING=.;
CLUSTER=idgleh; (...)
CATEGORICAL ARE (...);
ANALYSIS:
  TYPE=COMPLEX;
  ESTIMATOR = MLR;
  INTEGRATION = MONTECARLO;
MODEL:
  hitch by t3h0101* t3h0102 (...);
  hitch@1;
  meth by t3k0101* t3k0102 (...);
  meth@1;
  epis by t3k0202* t3k0203 (...);
  epis@1;
  wisse by t3w0101* t3w0102 (...);
  wisse@1;
  meth WITH hitch epis wisse;
OUTPUT: SAMPSTAT STANDARDIZED;

```

Mplus code for latent correlation model

In order to compare correlations, each model constraint was added to the model separately. For technical reasons in Mplus, the constraint sets the difference between two correlation coefficients to zero which is equivalent to constraining both coefficients to be equal.

```

MODEL:
(...)
! Check if Correlations are unequal
t3mewle(vary);
t3epwle(varz);
t3wwle(varx);
t3wwle WITH t3mewle(covxy);
t3wwle WITH t3epwle(covxz);

```

MODEL CONSTRAINT:

$$0 = \text{covxy}/\text{sqrt}(\text{varx}*\text{vary}) - \text{covxz}/\text{sqrt}(\text{varx}*\text{varz});$$

Mplus code for comparing correlations via Model Constraint (sample)

Regarding RQ3, all variables, except for those that were dummy-coded, were z-standardized beforehand. Means and standard deviations were obtained by running a saturated SEM with FIML and the manifest variables of RQ3 (all possible correlations estimated) taking the cluster-structure of the data into account (TYPE=Complex) in Mplus. Gender was dummy-coded with female = 1 if gender was reported as female, otherwise 0, and diverse = 1 if gender was reported as diverse, otherwise 0. Therefore, the reference category was male (if both dummies = 0).

(...)

```
! t3mwle refers to methodological test
! t3epwle refers to epistemological test
! t1v0301..3 refer to the grades reported by the students
! t1d01 refers to age
! t1d03 refers to gender
! t1d05 refers to number of books
! t1fwle refers to the ability test score in cognitive ability
! t1l refer to the test score in Reading Speed
MISSING=.;
CLUSTER=idg|eh;
```

DEFINE:

```
female = _MISSING;
divers = _MISSING;
IF (t1d03 EQ 1) THEN female = 1;
IF (t1d03 EQ 2) THEN female = 0;
IF (t1d03 EQ 3) THEN female = 0;
IF (t1d03 EQ 3) THEN divers = 1;
IF (t1d03 EQ 2) THEN divers = 0;
IF (t1d03 EQ 1) THEN divers = 0;
```

```
T3MEWLEs = (T3MEWLE-0.180)/SQRT(1.693);
T3EPWLEs = (T3EPWLE+0.096)/SQRT(1.822);
T1V0301s = (T1V0301-2.366)/SQRT(0.867);
T1V0302s = (T1V0302-2.463)/SQRT(0.691);
T1V0303s = (T1V0303-2.571)/SQRT(1.155);
T1D01s = (T1D01-14.824)/SQRT(0.392);
T1D05s = (T1D05-4.668)/SQRT(1.767);
T1FWLEs = (T1FWLE+0.040)/SQRT(1.026);
T1Ls = (T1L-46.210)/SQRT(83.986);
```

ANALYSIS:

```
TYPE=COMPLEX;
```

MODEL:

```
t1v0301s t1v0302s t1v0303s t1fwles t1ls t1d01s female divers t1d05s
```

WITH

```
t1v0301s t1v0302s t1v0303s t1fwles t1ls t1d01s female divers t1d05s;
t3mewles t3epwles ON
t1v0301s t1v0302s t1v0303s t1fwles t1ls t1d01s female divers t1d05s;
t3mewles t3epwles WITH
t3mewles t3epwles;
```

```
OUTPUT: SAMPSTAT STANDARDIZED;
```

Mplus code for the model with all three blocks as a sample

A3. Test used in the intervention study (Original German version and translated version in English)

Note: Original version of the test used in the intervention study (without external material), labeled with **item names** and codes for correct answers (1) and incorrect answers (0). Sources of the graphical materials (M...) are included in the workbook, attached after the test.

Aufgabe k01

Hier siehst du drei Karikaturen. Schau sie dir bitte an und bearbeite dann die Aufgaben dazu.
Die Karikaturen kannst du auch vergrößert und in Farbe im Materialheft unter M1-M3 auf den Seiten 1-3 finden.

Karikatur 1 / M1 (aus dem Jahr 1998)



Karikatur 2 / M2 (aus dem Jahr 2020)



Anmerkung zu Karikatur 2:

In dieser Karikatur geht es um die "Treuhandanstalt". Die Treuhandanstalt hatte die Aufgabe, die DDR-Wirtschaft "konkurrenzfähig" zu machen. Dabei gingen viele Arbeitsplätze im Osten verloren. Bis heute wird der Treuhandanstalt vorgeworfen, manche Ost-Firmen, wie die Leuna-Werke oder Carl Zeiss Jena, regelrecht "verscherbelt" (also viel zu günstig verkauft zu haben) zu haben.

Karikatur 3 / M3 (aus dem Jahr 2019)



Task k01

Here you see three caricatures. Please take a look at them and then complete the tasks related to them.
You can also view the caricatures enlarged and in color in the workbook under M1-M3 on pages 1-3.

Caricature 1 / M1 (year 1998)

Caricature 2 / M2 (year 2020)

Note on caricature 2:

This caricature is about the "Treuhandanstalt" (Trust Agency).

The Treuhandanstalt had the task of making the East German economy "competitive". In the process, many jobs in the East were lost. To this day, the Treuhandanstalt is accused of having "sold off" some East German companies, such as Leuna Works or Carl Zeiss Jena, at ridiculously low prices.

Caricature 3 / M3 (year 2019):

Karikaturen wollen etwas hervorheben und kritisieren; sie wollen eine Art Botschaft vermitteln.
Bitte entscheide dich für die Karikatur, die am besten zu der zentralen Botschaft passt.

Bitte kreuze an, welche der drei Karikaturen die jeweilige Aussage als zentrale Botschaft enthält.

Item	Zentrale Botschaft	Karikatur 1	Karikatur 2	Karikatur 3	keine der Karikaturen
k0101	Es ist (noch) nicht wieder zusammengewachsen, was zusammengehört.	1	0	0	0
k0102	Die DDR war ein Unrechtsstaat.	0	0	0	1
k0103	Bei der Treuhandanstalt wurden DDR-Betriebe zu billig verkauft.	0	1	0	0
k0105	Westdeutsche sind wertschätzend gegenüber Ostdeutschen und erkennen deren Lebensleistung an.	0	0	0	1
k0106	In der DDR konnte man billig Lebensmittel einkaufen.	0	0	0	1
k0107	Westdeutsche nehmen Ostdeutsche oft immer noch nicht ernst.	0	0	1	0
k0108	In der DDR gab es gute Spreewaldgurken.	0	0	0	1
k0109	Manche Westdeutsche haben sich nach der Wende in der früheren DDR bereichert.	0	1	0	0

Caricatures aim to highlight and criticize something; they intend to convey a kind of message. Please choose the cartoon that is best aligned with the central message.

Please check which of the three caricatures contains the respective statement as the central message.

Item	Central message	Caricature 1	Caricature 2	Caricature 3	None of the Caricatures
k0101	It (still) hasn't grown back together, what belongs together.	1	0	0	0
k0102	The GDR (German Democratic Republic) was an unjust state.	0	0	0	1
k0103	At the Treuhandanstalt, GDR companies were sold too cheaply.	0	1	0	0
k0105	West Germans are appreciative of East Germans and acknowledge their life achievements.	0	0	0	1
k0106	In the GDR, one could buy groceries inexpensively.	0	0	0	1
k0107	West Germans still often don't take East Germans seriously.	0	0	1	0
k0108	In the GDR, there were good Spreewald pickles.	0	0	0	1
k0109	Some West Germans enriched themselves in the former GDR after the reunification.	0	1	0	0

Aufgabe k04

Schau dir bitte den folgenden Comic an und bearbeite danach die Aufgaben dazu.
Den Comic kannst du auch vergrößert und in Farbe im Materialheft unter M4 auf Seite 4 finden.

Comic / M4



k04 Welche Botschaft will der Comic vermitteln?

Mehrere Möglichkeiten können richtig sein.

- k041** 0 Der Comic kritisiert die Willenlosigkeit der Ostdeutschen.
k042 1 Der Comic sagt etwas über die Vorurteile der Westdeutschen aus.
k043 0 Der Comic sagt aus, dass die Menschen in den DDR gefügig waren.
k044 0 Der Comic macht sich über die Vorstellungen von Jugendlichen damals lustig.
k045 1 Der Comic regt dazu an, sich mit den eigenen Vorstellungen über die DDR zu beschäftigen.

Task k04

Please take a look at the following comic and then complete the tasks related to it. You can also view the comic enlarged and in color in the workbook under M4 on page 4.

Comic / M4

k04 What message is the comic trying to convey?

Multiple options may be correct.

- k041** 0 The comic criticizes the passivity of East Germans.
k042 1 The comic reflects the prejudices of West Germans.
k043 0 The comic suggests that people in the GDR were compliant.
k044 0 The comic mocks the ideas of young people at that time.
k045 1 The comic encourages reflecting on one's own perceptions of the GDR.

Aufgabe k05

Lies dir zuerst die Hintergrundinformationen durch und beantworte dann die Frage.

1) Im August 2018 kam es auf einer Demonstration der Pegida (= Patriotische Europäer gegen die Islamisierung des Abendlandes, eine rechtsextreme Organisation) anlässlich des Besuchs der Kanzlerin Angela Merkel in Dresden zu einer Auseinandersetzung zwischen einem Teilnehmer und dem Kamerateam des Magazins "Frontal 21".

Der Mann, der einen **Anglerhut in Schwarz-Rot-Gold** trug, warf dem Kamerateam vor, eine "Straftat" zu begehen, als sie ihn filmten – eine Anschuldigung, die nicht stimmte, insbesondere, da der Mann selbst sich der Kamera aus eigenem Antrieb näherte. Später stellte sich heraus, dass der Mann ein Beamter des Landeskriminalamts war. Der Fall wurde in den Medien viel diskutiert und der Mann wurde als "**Hutbürger**" bekannt. Er wurde oft als Sinnbild für rechte Strömungen in Ostdeutschland herangezogen.

Im Jahr 2019 veröffentlichte der SPIEGEL dann ein Heft mit einem solchen Anglerhut Cover. Daraufhin kam es zu einem öffentlichen Aufschrei in den sozialen Medien, besonders von Ostdeutschen, die sich empörten.

Das Cover kannst du auch vergrößert und in Farbe im Materialheft unter M5 auf Seite 5 finden.

M5



- k0501** Warum wurde das Spiegel-Cover mit dem Anglerhut von vielen als provokativ empfunden?
- Mehrere Möglichkeiten können richtig sein.
- k05011** 1 Weil Ostdeutschen das Gefühl vermittelt wird, dass Westdeutsche auf sie herabblicken.
- k05012** 1 Weil nahegelegt wird, dass Ostdeutsche der Pegida nahestehen und die AfD wählen.
- k05013** 0 Weil hier von männlichen Ostdeutschen (auf dem Spiegel Cover steht "der Ossi") die Rede ist.
- k05014** 1 Weil viele Ostdeutsche das Gefühl hatten, dass Westdeutsche immer nur in Klischees über Ostdeutschland denken.
- k05015** 1 Weil nahegelegt wird, dass Ostdeutsche ein anderes Demokratieverständnis als Westdeutsche haben.

Task k05

First, read the background information and then answer the question.

1) In August 2018, during a Pegida demonstration (Patriotic Europeans Against the Islamization of the West, a far-right organization) on the occasion of Chancellor Angela Merkel's visit to Dresden, there was a confrontation between a participant and the camera team from the magazine Frontal 21.

The man, wearing a **fishing hat in black, red, and gold**, accused the camera team of committing a "crime" by filming him – an accusation that was untrue, especially since the man himself approached the camera voluntarily. It later turned out that the man was an officer of the State Criminal Police Office. The case was widely discussed in the media, and the man became known as the "**Hutbürger**". The person was often used as a symbol for right-wing currents in East Germany.

In 2019, DER SPIEGEL then published an issue with a cover featuring such a fishing hat. This led to a public outcry on social media, especially from East Germans who were outraged.

You can also find the enlarged and color version of the cover in the workbook under M5 on page 5.

M5

- k0501** Why was the Spiegel cover with the fishing hat perceived as provocative by many?
- Multiple options may be correct.
- k05011** 1 Because it conveys the feeling to East Germans that West Germans look down on them.
- k05012** 1 Because it suggests that East Germans are close to Pegida and vote for the AfD.
- k05013** 0 Because it refers to male East Germans (the Spiegel cover says "der Ossi", male version).
- k05014** 1 Because many East Germans felt that West Germans always think in stereotypes about East Germany.
- k05015** 1 Because it suggests that East Germans have a different understanding of democracy than West Germans.

2) Im Jahr 2015, also vier Jahre zuvor, veröffentlichte DER SPIEGEL dieses Cover. Das Cover kannst du auch vergrößert und in Farbe im Materialheft unter M6 auf Seite 5 finden.

M6



k0502 Auch hier wird pauschalisierend geurteilt, diesmal über die Bayern. Doch damals gab es keinen Aufschrei. Warum wurde dieses Spiegel-Cover im Jahr 2015 nicht als provokativ empfunden?

Mehrere Möglichkeiten können richtig sein.

- k05021** 0 Weil die Menschen in Bayern nicht so empfindlich sind wie die Ostdeutschen.
- k05022** 1 Weil die Bayern zwar auch mit Vorurteilen zu kämpfen haben, diese aber oft nicht so negativ sind.
- k05023** 0 Weil die abgebildeten Bayern repräsentative Kopfbedeckungen tragen (z.B. eine Krone).
- k05024** 1 Weil die hier angesprochenen Ansichten den Bayern gegenüber auch positive Aspekte haben.
- k05025** 0 Weil hier sowohl männliche als auch weibliche Bayern angesprochen wurden.
- k05026** 1 Weil Bayern vielfältig und reich an Kultur dargestellt wird, mit Merkmalen, auf die sie auch stolz sind.

2) In the year 2015, four years earlier, DER SPIEGEL published this cover. You can also view the enlarged and color version of the cover in the workbook under M6 on page 5.

M6

k0502 Here, too, judgments are made in a generalizing way, this time about Bavarians. However, there was no outcry back then. Why was this Spiegel cover in 2015 not perceived as provocative?

Multiple options may be correct.

- k05021** 0 Because people in Bavaria are not as sensitive as East Germans.
- k05022** 1 Because Bavarians do face prejudices, but these are often not as negative.
- k05023** 0 Because the depicted Bavarians wear representative headgear (e.g., a crown).
- k05024** 1 Because the views addressed here also convey positive aspects of Bavarians.
- k05025** 0 Because both male and female Bavarians are addressed.
- k05026** 1 Because Bavaria is depicted as diverse and rich in culture, with characteristics they are proud of.

Aufgabe k06

Im Folgenden wirst du zwei kurze Texte von Menschen lesen, die in der DDR gelebt haben und ihre Meinung über die DDR und die Wiedervereinigung Deutschlands sagen. Lies dir bitte zuerst die beiden Texte durch und beantworte danach die Fragen.

Ilko-Sascha Kowalcuk, geboren am 4. April 1967 in Berlin-Friedrichshagen (d.h. in Ostberlin), war beim Mauerfall 22 Jahre alt. Herr Kowalcuk gab das Interview im Jahr 2010; damals war er Projektleiter bei der Behörde zur Verwaltung der Stasi-Akten in Berlin:

<

Auszüge aus einem Interview von Kai Pfundt mit Ilko-Sascha Kowalcuk auf den Seiten 78 und 79 enthalten Antworten auf die erste, vierte, sechste und siebte Frage.

Quelle: Bundeszentrale für politische Bildung/bpb (2011). Geschichte der DDR. *Informationen Zur Politischen Bildung*, 312(3).

>

Simone Fall, eine 1939 geborene Rostockerin (d.h. zum Zeitpunkt des Mauerfalls 50 Jahre alt), die seit 1965 auf einer Werft gearbeitet hat, berichtete im Jahr 2009:

<

Auszüge aus dem Interview von Joachim Gebhardt und Wolfgang Hammer mit Simone Fall am 12.01.2009 auf Seite 36 sind von Zeile 15 bis 23, 30 bis 34, 35 bis 38 und 47 bis 53.

Quelle: Gebhardt, J. & Hammer, W. (2009). LebensWENDEn: "Es war nicht alles schlecht!" - "Es war nicht alles gut" (UE Sek I/Sek II). *Praxis Geschichte*, 5, 32–37.

>

Task k06

Below, you will read two brief texts from individuals who lived in the GDR, expressing their opinions on the GDR and the reunification of Germany. Please first read the two texts and then answer the questions.

Ilko-Sascha Kowalcuk, born on April 4, 1967, in Berlin-Friedrichshagen (i.e., in East Berlin), was 22 years old when the Berlin Wall fell. Mr. Kowalcuk gave the interview in 2010; at that time, he was a project manager at the agency responsible for managing Stasi files in Berlin:

<

Excerpts from an interview by Kai Pfundt with Ilko-Sascha Kowalcuk on pages 78 and 79 include responses to the First, Fourth, Sixth, and Seventh Question.

Bundeszentrale für politische Bildung/bpb (2011). Geschichte der DDR. *Informationen Zur Politischen Bildung*, 312(3).

>

Simone Fall, a native of Rostock born in 1939 (i.e., she was 50 years old when the Berlin Wall fell), who had been working in a shipyard since 1965, reported in the year 2009:

<

Excerpts of interview by Joachim Gebhardt and Wolfgang Hammer with Simone Fall on 12 January 2009 on page 36 included from lines 15 to 23, 30 to 34, 35 to 38, and 47 to 53.

Gebhardt, J. & Hammer, W. (2009). LebensWENDEn: "Es war nicht alles schlecht!" - "Es war nicht alles gut" (UE Sek I/Sek II). *Praxis Geschichte*, 5, 32–37.

>

Welcher der Texte drückt die folgenden Aussagen in anderen Worten aus?

Kreuze an. Setze bitte nur ein Kreuz pro Zeile.

Item		Ilko-Sascha Kowalcuk	Simone Fall	zu keinem Text
k0601	Die DDR war kein Rechtsstaat.	1	0	0
k0602	In der DDR ging es einem gut.	0	1	0
k0603	Die Führung der DDR war alt und verrückt.	0	0	1
k0606	In der DDR wurde den Menschen die Freiheit genommen.	1	0	0

Which of the texts expresses the following statements in different words

Check the appropriate boxes. Please only mark one box per line.

Item		Ilko-Sascha Kowalcuk	Simone Fall	None of the texts
k0601	The GDR was not a state governed by the rule of law.	1	0	0
k0602	Life was good in the GDR.	0	1	0
k0603	The leadership of the GDR was old and eccentric.	0	0	1
k0606	Freedom was taken away from the people in the GDR.	1	0	0

Aufgabe z

Die Aussagen in der Aufgabe sind vereinfachte Zitate aus einem Interviewprojekt, in dem Menschen im Osten und Westen Deutschlands zu ihrer Erinnerung an die DDR und BRD und zu ihrer heutigen Einschätzung der deutschen Einheit befragt.

Bitte überlege dir bei den folgenden Aussagen, von wem sie wohl stammen: von einem oder einer Zeitzeug*in aus dem Osten oder aus dem Westen?

Vielleicht wirst du dich bei manchen Aussagen zuerst nicht recht entscheiden können. Wähle bitte trotzdem eine Antwort aus, die du am wahrscheinlichsten findest.

Setze bitte für jede Aussage ein Kreuz.

Item	Aussage	Zeitzeug*in aus dem Osten	Zeitzeug*in aus dem Westen
z01	In der Schule haben sie uns erklärt, wie die Gesellschaftsordnung drüben funktioniert: Dass es sehr egoistisch drüben ist, dass es da eine Ellenbogengesellschaft gibt, Drogen, Kapitalismus, Ausnutzung.	1	0
z02	Wenn die Pakete an die Verwandten im anderen Deutschland gepackt wurden, war ich dann manchmal so ein bisschen neidisch. Weil da schöne Süßigkeiten in die Pakete gewandert sind, die ich nie bekommen hätte. Und dachte so, wie schlecht muss es denen gehen.	0	1
z03	Meine Eltern haben die einseitige Darstellung des anderen Deutschlands als kapitalistische Ausbeutergesellschaft in der Schule nicht thematisiert.	1	0
z04	Man hat sich das andere Deutschland eigentlich so vorgestellt: Die waren alle blass. Die hatten nichts zu essen und es ist einfach nur Chaos.	0	1
z05	Und dann wurde man so behandelt als wäre man ein Befreiter.	1	0
z06	Die Ignoranz der anderen Deutschen ist in Berlin weniger ausgeprägt als hier in "Restdeutschland", weit weg von dem ganzen Geschehen zwischen Ost und West.	1	0

Task z

The statements in the task are simplified quotes from an interview project in which people in East and West Germany were interviewed about their memories of the GDR and FRG and their current assessment of German unity.

For the following statements, please consider from whom they might come: from an eyewitness in the East or the West?

You might not be able to decide definitively about some statements at first. Still, please choose an answer that you find most likely.

Place a check mark for each statement.

Item	Statement made by interviewee from the...	East	West
z01	In school, they explained to us how the societal order worked over there: That it's very selfish over there, that there's a cutthroat society, drugs, capitalism, exploitation.	1	0
z02	When packages were packed for relatives in the other Germany, I was sometimes a bit envious. Because beautiful sweets were going into those packages that I would never get. And I thought how bad it must be for them.	0	1
z03	My parents did not address the school's one-sided portrayal of the other Germany as a capitalist exploitation society.	1	0
z04	We actually imagined the other Germany like this: They were all pale. They had nothing to eat, and it's just chaos.	0	1
z05	And then you were treated as if you were a liberator.	1	0
z06	The ignorance of the other Germans is less pronounced in Berlin than here in "the rest of Germany", far away from all the events between the East and West.	1	0

	Aussage	Zeitzeug*in aus dem Osten	Zeitzeug*in aus dem Westen
z07	Ich kenn' auch viele, die auch heute noch über Deutschland, also unser Deutschland, schimpfen.	0	1
z08	Mein Eindruck ist, dass die anderen Deutschen das Gefühl haben: Die von drüben haben sich ihnen angeschlossen, deshalb sind sie selbst etwas Besseres.	1	0
z09	Ich sage auch, dass nicht jeder gelitten hat. Es gab auch Beispiele, wo es gut lief.	1	0
z10	Für mich war das andere Deutschland halt irgendwie ein anderes Land, wo die Leute auch deutsch sprechen, die arm dran sind, denen es schlecht geht. Aber warum das so ist, das wusste ich gar nicht.	0	1
z11	Also manchmal hat man so den Eindruck, dass man, wenn man sich zu erkennen gibt, dass man dann nach wie vor von manchen im anderen Deutschland vermeintlich so von oben herab angeschaut wird.	1	0
z12	Wir haben früher schon ein extrem gutes Leben auf hohem Niveau gehabt. Und dieses Leben hat das andere Deutschland überhaupt nicht gehabt.	0	1

	Statement made by interviewee from the...	East	West
z07	I also know many who still complain about Germany, our Germany, even today.	0	1
z08	My impression is that the other Germans feel: Those from over there have joined them, so they themselves are somewhat better.	1	0
z09	I also say that not everyone suffered. There were also examples where things went well.	1	0
z10	For me, the other Germany was just somehow another country where people also speak German, who are poor, who are not doing well. But why that is so, I didn't know at all.	0	1
z11	Sometimes you get the impression that when you reveal your identity, you are still looked down upon by some people in the other Germany.	1	0
z12	We already had an extremely good life at a high level before. And the other Germany did not have this life at all.	0	1

Aufgabe k02

Bitte lies die folgenden Aussagen durch.

Beispiel: Leonardo da Vinci war...

- ein italienischer Universalgelehrter.
- Ein Philosoph im antiken Griechenland.
- Maler des berühmten Gemäldes „Mona Lisa“.

Hier stimmen die erste und die dritte Option.

Entscheide bei den nächsten Aufgaben selbst, was stimmt und was nicht.

Kreuze jeweils die Aussage an, der du zustimmst.

Es können eine oder mehrere Möglichkeiten richtig sein.

- k0202** Geschichte, das ist einfach eine Reihe von Fakten.
- k02021** 0 Stimme ich nicht zu, weil es in Geschichte keine Fakten, sondern nur Meinungen gibt.
- k02022** 1 Stimme ich nicht zu, weil Geschichte auch beinhaltet, Zusammenhänge zwischen Ereignissen herzustellen.
- k02023** 0 Stimme ich zu, weil bestimmte Daten über wichtige historische Personen das Wichtigste an Geschichte sind.
- k0203** Man sollte verschiedene historische Quellen berücksichtigen, bevor man ein Urteil fällt.
- k02031** 0 Stimme ich nicht zu, denn man hat meist mit einer Quelle schon genug Informationen, um ein Urteil zu fällen.
- k02032** 0 Stimme ich nicht zu, weil verschiedene Quellen zum selben Ereignis immer dasselbe berichten, weil ja alle dasselbe erlebt haben.
- k02033** 1 Stimme ich zu, weil verschiedene Quellen unterschiedliche Gesichtspunkte eines Ereignisses beleuchten können.
- k0204** Geschichte kann niemals etwas Gesichertes über die Vergangenheit aussagen.
- k02041** 0 Stimme ich zu, weil es im Bereich der Geschichte immer so viele verschiedene Perspektiven gibt, dass man nie sagen kann, welche wahr ist und welche nicht.
- k02042** 0 Stimme ich nicht zu, weil man immer eindeutig weiß, was passiert ist.
- k02043** 1 Stimme ich nicht zu, weil es bei vielen Ereignissen der Vergangenheit Erkenntnisse gibt, die als weitgehend sicher betrachtet werden können.

Task k02

Read the following statements.

Example: Leonardo da Vinci was...

- An Italian polymath.
- A philosopher in ancient Greece.
- The painter of the famous painting the Mona Lisa.

Here, the first and the third options are correct.

Decide for yourself in the next tasks what is correct and what is not.

Check the statement(s) you agree with.

One or more options may be correct.

- k0202** History is simply a series of facts.
- k02021** 0 I disagree because, in history, there are no facts, only opinions.
- k02022** 1 I disagree because history also involves establishing connections between events.
- k02023** 0 I agree because certain data about important historical figures is the most important aspect of history.
- k0203** One should consider various historical sources before forming a judgment.
- k02031** 0 I disagree because one usually has enough information from just one source to form a judgment.
- k02032** 0 I disagree because different sources always report the same thing for the same event because everyone experienced the same thing.
- k02033** 1 I agree because different sources can illuminate different aspects of an event.
- k0204** History can never state something definite about the past.
- k02041** 0 I agree because, in the field of history, there are always so many different perspectives that one can never say which is true and which is not.
- k02042** 0 I disagree because we always know clearly what happened.
- k02043** 1 I disagree because many events in the past have insights that can be considered largely secure.

- k0207** Alle Perspektiven auf ein historisches Ereignis sind prinzipiell gleichberechtigt.
- k02071** 1 Stimme ich nicht zu, da manche Perspektiven Aussagen enthalten, die nicht überprüfbar sind oder sogar gesicherten Erkenntnissen widersprechen.
- k02072** 1 Stimme ich nicht zu, weil man zwar schon verschiedene Perspektiven berücksichtigen sollte, aber auch überprüfen muss, wie plausibel sie sind.
- k02073** 0 Stimme ich nicht zu, weil Berichte von Zeitzeug*innen eher der Wahrheit entsprechen als die von sonstigen Quellen und Historiker*innen.
- k0208** Wenn man ein historisches Thema verstehen möchte, sollte man sich mehrere Quellen anschauen.
- k02081** 0 Stimme ich zu, weil man dabei merkt, dass sich alle Quellen unterscheiden und man deshalb nie sagen kann, was wahr oder falsch ist.
- k02082** 1 Stimme ich zu, da Quellen oft helfen können, verschiedene Ansichten aus der Vergangenheit zu verstehen.
- k02083** 1 Stimme ich zu, da Quellen auch oft verfälscht oder voller Unwahrheiten sein können. Daher braucht man mehrere Quellen, um vergleichen zu können.

- k0207** All perspectives on a historical event are essentially equal.
- k02071** 1 I disagree because some perspectives contain statements that are not verifiable or even contradict established knowledge.
- k02072** 1 I disagree because while one should consider different perspectives, one must also assess how plausible they are.
- k02073** 0 I disagree because accounts from eyewitnesses are more likely to reflect the truth than those from other sources and historians.
- k0208** If one wants to understand a historical topic, one should look at multiple sources.
- k02081** 0 I agree because this reveals that all sources differ, and therefore, one can never say what is true or false.
- k02082** 1 I agree because sources often help people understand various views from the past.
- k02083** 1 I agree because sources can also be distorted or full of falsehoods. Therefore, one needs multiple sources for comparison.

Aufgabe k03

Lies die folgenden Aussagen durch.

Beispiel: Leonardo da Vinci war...

- ein italienischer Universalgelehrter.
- Ein Philosoph im antiken Griechenland.
- Maler des berühmten Gemäldes "Mona Lisa".

Hier stimmen die erste und die dritte Option.

Entscheide bei den nächsten Aufgaben selbst, was stimmt und was nicht.

Kreuze jeweils die Aussage an, der du zustimmst.

Es können eine oder mehrere Möglichkeiten richtig sein.

k0301 Ich bin überzeugt, dass ich anders über die DDR und die BRD denke, wenn ich mehr darüber lerne.

k03011 1 Stimme ich zu, weil man immer neue Erkenntnisse hat, wenn man sich mit Geschichte beschäftigt.

k03012 1 Stimme ich zu, weil man merkt, dass es viele Dinge in der heutigen Welt gibt, die stark mit der Vergangenheit zusammenhängen.

k03013 0 Stimme ich nicht zu, weil die Geschichte der DDR zwar schon interessant ist, aber die heutige Welt damit nichts wirklich zu tun hat.

k0302 Ich kenne die historischen Gründe für einige der heutigen Probleme in Ostdeutschland und denke, dass sich meine Meinung kaum noch ändern wird.

k03021 0 Stimme ich zu, denn auch neue Informationen werden meine Meinung nicht grundlegend verändern können.

k03022 1 Stimme ich nicht zu, weil es viele Perspektiven gibt und man dadurch nie objektiv sagen kann, was wahr ist und was nicht.

k03023 1 Stimme ich nicht zu, weil es immer historische Perspektiven oder Sachverhalte gibt, die man noch nicht kennt – diese könnten meine Meinung doch wieder ändern.

k03024 0 Stimme ich zu, weil meine Meinung auf Fakten beruht, so dass es gar nicht sein kann, dass ich meine Meinung nochmal ändere.

Task k03

Read the following statements.

Example: Leonardo da Vinci was...

- An Italian polymath.
- A philosopher in ancient Greece.
- The painter of the famous painting the Mona Lisa.

Here, the first and the third options are correct.

Decide for yourself in the next tasks what is correct and what is not.

Check the statement(s) you agree with.

One or more options may be correct.

k0301 I am convinced that I will think differently about the GDR and FRG if I learn more about them.

k03011 1 I agree because one always gains new insights when delving into history.

k03012 1 I agree because one realizes that many things in the present world are strongly connected to the past.

k03013 0 I disagree because, while the history of the GDR is interesting, it doesn't really have much to do with the present world.

k0302 I know the historical reasons for some of the current problems in East Germany and think that my opinion is unlikely to change significantly.

k03021 0 I agree because even new information is unlikely to fundamentally alter my opinion.

k03022 1 I disagree because there are many perspectives, making it impossible to objectively determine what is true and what is not.

k03023 1 I disagree because there are always historical perspectives or facts that one may not know yet—these could potentially change my opinion.

k03024 0 I agree because my opinion is based on facts, so it's unlikely that I would change my opinion.

- k0303** Im Unterricht sollte man lernen, dass es viele unterschiedliche, aber durchaus begründete Sichtweisen auf ein historisches Ereignis wie die Deutsche Einheit gibt.
- k03031** 0 Stimme ich nicht zu, denn es gibt nur eine Wahrheit.
- k03032** 1 Stimme ich zu, da beispielsweise die Ostdeutschen andere Erfahrungen als die Westdeutschen gemacht haben.
- k03033** 0 Stimme ich nicht zu, denn ich denke, dass ost- und westdeutsche Schüler*innen in meinem Alter gleich über die Deutsche Einheit denken sollten.
- k03034** 1 Stimme ich zu, da das eigene Geschichtsbild stark davon abhängt, wie, wo und wann man groß geworden ist.
- k0304** Die Beschäftigung mit der Vergangenheit kann uns dabei helfen, die Gegenwart besser zu verstehen und heutige Möglichkeiten zum Handeln besser einzuschätzen.
- k03041** 0 Stimme ich zu, weil die Kenntnis über die Ereignisse in der Vergangenheit eine klare Richtschnur für heute ist.
- k03042** 1 Stimme ich zu, doch man sollte die Vergangenheit und Gegenwart trennen. Die Geschichte wiederholt sich nicht und wir müssen uns schon selbst Lösungen für unsere Probleme einfallen lassen.
- k03043** 1 Stimme ich zu, weil man sich manchmal aus dem, was in der Vergangenheit passiert ist, ein Urteil bilden kann, das einem bei Entscheidungsfindungen heute helfen kann.
- k0305** Aus den persönlichen Berichten von Zeitzeug*innen kann man etwas darüber erfahren, wie unsere Gegenwart so geworden ist, wie sie ist.
- k03051** 0 Stimme ich nicht zu, weil unsere Gegenwart und die Berichte von Zeitzeug*innen nichts miteinander zu tun haben.
- k03052** 0 Stimme ich nicht zu, weil die Berichte von Zeitzeug*innen sich ständig widersprechen und sie deshalb keine Relevanz für die Gegenwart haben.
- k03053** 1 Stimme ich zu, weil die Erfahrungen der Zeitzeug*innen in der Vergangenheit neue Perspektiven auf die Gegenwart geben können.
- k0306** Wenn wir heute etwas über die Geschichte der DDR lernen, ist das dieselbe Geschichte, die meine Eltern damals in der Schule gelernt haben.
- k03061** 0 Stimme ich zu, denn die Vergangenheit hat sich ja nicht geändert.
- k03062** 1 Stimme ich nicht zu, weil man damals eine andere Gegenwart und deswegen auch eine andere Deutung der Vergangenheit hatte.
- k03063** 1 Stimme ich nicht zu, weil sich Geschichte immer verändern kann, je nach dem, in welcher Zeit man in die Vergangenheit zurückschaut.
- k03064** 0 Stimme ich nicht zu, weil man nie wirklich etwas Wahres über die Vergangenheit wissen kann, weder heute noch damals.

- k0303** In class, one should learn that there are many different but entirely justified perspectives on a historical event such as the German reunification.
- k03031** 0 I disagree because there is only one truth.
- k03032** 1 I agree because, for example, East Germans have had different experiences than West Germans.
- k03033** 0 I disagree because I believe that East and West German students of my age should think similarly about the German reunification.
- k03034** 1 I agree because one's view of history depends greatly on how, where, and when one grew up.
- k0304** Dealing with the past can help us better understand the present and better assess today's opportunities for action.
- k03041** 0 I agree because knowledge of past events is a clear guideline for today.
- k03042** 1 I agree, but one should separate the past from the present. History does not repeat itself, and we must come up with solutions to our problems ourselves.
- k03043** 1 I agree because sometimes, judgments about what happened in the past can help in decision-making today.
- k0305** From the personal accounts of eyewitnesses, one can learn something about how our present has become what it is.
- k03051** 0 I disagree because our present and the accounts of eyewitnesses have nothing to do with each other.
- k03052** 0 I disagree because the accounts of eyewitnesses constantly contradict each other, and therefore, they have no relevance to the present.
- k03053** 1 I agree because the experiences of eyewitnesses in the past can provide new perspectives on the present.
- k0306** If we learn something about the history of the GDR today, it is the same history that my parents learned in school back then.
- k03061** 0 I agree because the past has not changed.
- k03062** 1 I disagree because back then, there was a different present and, therefore, a different interpretation of the past.
- k03063** 1 I disagree because history can always change depending on the time from which one looks back into the past.
- k03064** 0 I disagree because one can never truly know something true about the past, neither today nor back then.

- k0307** Die Geschichte der DDR und des geteilten Deutschlands hat heute noch große Relevanz.
- k03071** 1 Stimme ich zu, da man etwas über heutige Zustände aus der Teilungsgeschichte lernen kann.
- k03072** 0 Stimme ich nicht zu, da die Teilungsgeschichte nicht dabei hilft, heutige Zustände zu erklären. Man kann die Gegenwart mit Wissen über die Vergangenheit nicht besser verstehen.
- k03073** 1 Stimme ich zu, da viele gesellschaftliche Konflikte heutzutage immer noch mit der Vergangenheit Deutschlands zu tun haben bzw. damit begründet werden.
- k03074** 0 Stimme ich nicht zu, da man die heutigen Probleme nicht mit dem Wissen über die Vergangenheit lösen kann.

- k0307** The history of the GDR and divided Germany is still highly relevant today.
- k03071** 1 I agree because one can learn about current conditions from the history of division.
- k03072** 0 I disagree because the history of division does not help explain current conditions. Knowledge about the past does not lead to a better understanding of the present.
- k03073** 1 I agree because many contemporary societal conflicts still have to do with or are justified by Germany's past.
- k03074** 0 I disagree because one cannot solve today's problems with knowledge about the past.

Materialheft

M1

Karikatur 1 (aus dem Jahr 1998)

Karikatur "Menschenmauer" von Barbara Henniger aus dem Jahr 1998 (<https://barbarahenniger.de/>).

Henniger, B. (1998). *Menschenmauer* [Bild]. Weser Kurier. Abgerufen am 1. August 2024 von https://www.weser-kurier.de/resources/0266-11874356229d-1ad62cceb3c0-1000/format/large/was_vom_schrecken_blieb_barbara_henniger_zeichnete_die_karikatur_mit_dem_titel_menschenmauer_im_jahr_1998_abbildung_barbara_henniger.jpeg

Workbook

M1

Caricature 1 (year 1998)

Caricature "Menschenmauer" by Barbara Henniger, year 1998 (<https://barbarahenniger.de/>).

Henniger, B. (1998). *Menschenmauer* [Image]. Weser Kurier. Retrieved August 1, 2024, from https://www.weser-kurier.de/resources/0266-11874356229d-1ad62cceb3c0-1000/format/large/was_vom_schrecken_blieb_barbara_henniger_zeichnete_die_karikatur_mit_dem_titel_menschenmauer_im_jahr_1998_abbildung_barbara_henniger.jpeg

M2

Karikatur 2 (aus dem Jahr 2020)

Karikatur "DDR-Ausverkauf" von Peter Butschkow aus dem Jahr 2020 (<https://www.butschkow.de/>).

M2

Caricature 2 (year 2020)

Caricature "DDR-Ausverkauf" by Peter Butschkow, year 2020 (<https://www.butschkow.de/>).

Karikatur 3 (aus dem Jahr 2019)

Karikatur "Gelingt die Vollendung der Einheit?" von Greser und Lenz aus dem Jahr 2019 (<https://www.greser-lenz.de>).

Greser & Lenz (2019). Gelingt die Vollendung der Einheit? [Bild]. FAZ. Abgerufen am 1. August 2024 von <https://www.faz.net/aktuell/feuilleton/greser-lenz-witze-fuer-deutschland-17815556/karikatur-greser-und-lenz-16382381.html#21>

Caricature 3 (year 2019)

Caricature "Gelingt die Vollendung der Einheit?" by Greser and Lenz, year 2019 (<https://www.greser-lenz.de>).

Greser & Lenz (2019). *Gelingt die Vollendung der Einheit?* [Image]. FAZ. Retrieved August 1, 2024, from <https://www.faz.net/aktuell/feuilleton/greser-lenz-witze-fuer-deutschland-17815556/karikatur-greser-und-lenz-16382381.html#21>

Comic

Flix. (2014). *Da war mal was: Erinnerungen an hier und drüben (Extended original version)*. Carlsen.

Comic "Moritz" von Flix auf den Seiten 24 bis 26.

Comic

Flix. (2014). *Da war mal was: Erinnerungen an hier und drüben (Extended original version)*. Carlsen.

Comic "Moritz" by Flix on pages 24 to 26.

Der Spiegel Cover (August 2019)

Spiegel. (2019, 23. August). *Der Spiegel*, Nr. 35. Abgerufen am 1. August 2024
<https://www.spiegel.de/spiegel/print/index-2019-35.html>

Der Spiegel Cover (August 2019)

Spiegel. (2019, August 23). *Der Spiegel*, No. 35. Retrieved August 1, 2024, from
<https://www.spiegel.de/spiegel/print/index-2019-35.html>

Der Spiegel Cover (August 2015)

Spiegel. (2015, 21. August). *Der Spiegel*, Nr. 35. Abgerufen am 1. August 2024
<https://www.spiegel.de/spiegel/print/index-2015-35.html>

Der Spiegel Cover (August 2015)

Spiegel. (2015, August 21). *Der Spiegel*, No. 35. Retrieved August 1, 2024, from
<https://www.spiegel.de/spiegel/print/index-2015-35.html>