

How to Examine the (Self-)Reflective Effects of History Teaching

BODO VON BORRIES

In many countries the public and the political authorities strongly insist on *standards* and *large scale evaluation* (using *high stake testing*). In this concept, a school subject without precisely controllable and self-evidently useful achievements – especially for future employment – does not merit learning time at school. This is fashionable, but merely a so-called *mechanism to please everybody*. Simple testing of historical competencies is an *illusion*, it promises more than can ever be effected.

In contrast, many progressive critics (e.g. teachers and educators at memorial sites) deny and refuse any time-saving, objective and efficient evaluation of learning processes in history. They express their deep conviction that the mania of objective tests is damaging (self-)reflective historical learning. This scepticism of evaluating historical learning processes is just as bad – and as risky for historical learning as standards and large-scale evaluation might be, as long as it means that more elaborated tests are not developed and tried out. If there are no measurable effects, history may vanish from the future obligatory syllabi. Any communication between supporters and critics of evaluation seem to have totally stopped in some countries. Here both parties have to return to the common table and learn together.

Obviously, learning history is everything but *copying and pasting* of information about the past into the students' minds. It is rather the individual interpretation of the relation between past, present and future within social frameworks, which are relevant for an individual's and a group's orientation in life. Teaching history in a (self-)reflective way

does not mean teaching (nor enforcing) special knowledge, conclusions, opinions and attitudes, but stimulating and promoting participation in the *culture of history* (using [re]presentations of history), building one's own *historical identity* (definition of self and groups in the course of time) and improving one's *historical competence* (developed abilities and willingness of thinking historically) (see Körber et al. 2007; Schreiber/Körber 2006)¹. The result will and legitimately may be different for every individual, for instance among thirty seventh-graders in a common classroom. This is obvious, but seldom seriously taken into account by history teachers.

How does one evaluate, examine, and differentiate such a complex and ambivalent learning process? *The long-term effects and the real behaviour in future life cannot be measured exactly, but only predicted hypothetically to a certain, rather low degree.* The limitations of *multiple-choice* questions and of *paper-pencil* tests in general in evaluating historical competencies are narrow, though potentially much wider than the actual practice of testing in many countries (comp. chapter 2). Indeed, differentiated tests employ more intelligent formats like allowing a choice of alternative answers, production of *mental maps* or writing short essays, which are able to diagnose special figures and logical steps of thinking historically.

Nevertheless, the combination of insights, emotions, conclusions, and activities, which are typical for the entire process of historical learning, can only be evaluated by observation of students' (re-)actions in similar situations or conflicts. Competences are transferable – but domain-specific – abilities of finding and solving problems in future life. *There is no possibility to evaluate them other than by simulating new (hypothetical) problems of a similar type to those the student has already solved and by making the student work and act in them.* In the following paragraphs, this will be explained by examples.

The argumentation will proceed in four steps:

- Explanation of the preconditions of testing (self-)reflective historical thinking,
- Examples of traditional (anti-)reflective history tests,
- Examples of alternative testing approaches to second order concepts in history,
- Examples of innovative testing methods for the evaluation of *dealing with history*.

1 For elaboration of the theoretical understanding of the development of (self-)reflective historical thinking see the contribution by Körber in this volume.

1. General approach

What is the aim of history learning? Since the 1970s, German tradition and theory has been pointing to *an elaborated historical consciousness* (Jeismann 1985; Rüsen 1994a; Rüsen 1994 b; Schreiber 2003). But what does that mean? Is there any chance of operationalizing the concept of historical consciousness, which will be a necessary precondition of diagnosing and measuring? Other concepts which seem to emerge with strength and conviction in the debate, are historical identity, historical competence and historical culture (or better: culture of history), as well as politics of history, identity management and places of remembrance. This requires a small theoretical remark: To me, “historical consciousness” remains the key concept and the general expression to be used (Borries 2008a:4ff; 2008b). Therefore, the crucial objective is to create “reflecting and (self-)reflexive historical consciousness” (see references above) of which three dimensions can and should be distinguished:

- “*Historical identity*” means the self-definition of an individual or a group (city, gender, class, profession, nation) between interpreted past, perceived present and expected future. It is the construction of *sense* and *continuity* (persistence in change) in the course of time and its *contingency* (more than accident, less than necessity). It needs to be added that knowing and understanding otherness (*alterity*) and *foreign* behavior and identities is part of the process of evolving one’s own identity.² Every identity has a narrative and a historical structure (Borries 2004. 259-287; Ricoeur 1988/89/91), but not all histories told by individuals or groups are sufficiently plausible or even plausible to a higher degree (Rüsen 1983: 85-136). The inclusion of verified primary sources from different sides and of different points of view of the present etc. improves the quality of historical narratives and the chances of communication and agreement.

“*Historical competence*” is the ability and willingness of thinking historically as one of the main accesses to the world (in addition to a mathematical, a linguistic, an aesthetic, a scientific one etc.). It is the capability and readiness: (1.) to ask historical questions and understand them, (2.) to investigate historical topics and produce historical narratives about them and/or analyze the already perfected histories of others, (3.) to orientate oneself through historical narrations within present and future (reorganizing ones own historical consciousness and the understanding of oneself, the others and the world) and

2 In order to know oneself, a narcissistic glance into the mirror is not sufficient; discovering the world and reflecting one’s relation to it helps to position oneself.

(4.) to store and organize concepts, categories, notions, structures of history (Körber et al. 2007; Schreiber/Körber et al. 2006).

- “*Historical culture*” is the public and private mode of handling history in presentations, consumption and discourse. Though the individuals decide to a large extent personally on the degree of participation – and non-participation – in historical culture, history culture is basically a social, societal, and communicative phenomenon. Today it is mostly determined by the mass media and their private and profit-oriented market organization. The type of interference of individuals and groups in this field of historical activities (e.g. use, neglect, protest, debate, partisanship, revision) is the third dimension of historical consciousness.

All three dimensions can be operationalized, observed (and sometimes measured) and evaluated (to some degree at least). Normally, test producers think of *knowledge* only,³ but *competences* are more important in modern concepts of historical learning (Borries 2008; Körber et al. 2007; Schreiber/Körber 2006). Living and showing *historical identity* and taking part in *historical culture* are activities too and can be observed and described. But the performance of historical identity and the use of historical culture always need and require *historical competence*, the second of the three dimensions of historical consciousness. Thus, the main problem will be the diagnosis of historical competences.

Additionally, we have to state that identities and cultures cannot simply be judged by their affiliations. A *nationalist* does not per definition have a better or worse historical identity than a *cosmopolitan*. The same is true for *conservatives* and *progressives*, rightists and leftists. Visiting museums or being member of a local history club does not have a higher status than watching television or reading historical novels, at least not in a self-evident way. Already thirty years ago, Jeismann (see 1985) distinguished three interdependent and necessary levels of historical learning in any learning process: “historical topic analysis”, “historical topic judgement” and “historical value judgement”.⁴ Jeismann clarified a peculiarity of the third layer (that is *historical value judgement* or *historical orientation*): In this field it is inadmissible to *grade* the direc-

3 Of course, we have to be careful because of the very wide meaning of the word “knowledge” in English. In German, we may distinguish between special and single information (“Kenntnisse”), the content knowledge in a field (“Wissen”) and the general system including methods, insights and reflections (“Wissensformen” in a very comprehensive sense).

4 Rüsen (1994b: 64-73, 164-170) similarly enumerates “historical perception”, “historical interpretation” and “historical orientation”.

tion of the decision in its content. Only the quality of argumentation and reflection may be evaluated, because it is important to compare and legitimize a decision contrasting it to other possibilities (Jeismann et al. 1987). This limitation of grading remains very important.

2. Exclusive focus on knowledge damages (self-)reflective learning

I will here provide two examples of how the testing of mere historical information can damage (self-)reflective learning. Since 2008, Germany has a “scientific” and “standardized” test – designed by a well-known testing institute in Berlin – for evaluating an obligatory orientation course for foreigners desiring naturalization. This is the direct preparation for a “naturalization test”, devised by the Ministry of the Interior. The whole test consists of “multiple choice” items (four possibilities, one correct and three incorrect).⁵ The test was commented by some experts as statistically *reliable*, but showing no *content validity* whatsoever. Many items were said to be wrong, ideological, unfair or unimportant. One question reads:

“When did Hitler become Chancellor of the ‘[German] Reich’?
1923 1927 1933 1936”

Undoubtedly, this is a technically correct topic – and it is rather simple. What does a correct answer actually measure? Knowing the information about the date – as opposed to the dates of the nearly fatal crises of the Weimar Republic (1923), the summit of the “Golden Years” (1927) and the Olympics in Berlin/Spanish Civil War (1936) – is useful, though not at all important. Without the knowledge of why this was important and what happened involving which actors with what consequences it is “pointless, dead information”. Choosing the correct date 1933 in the test does not say anything about the reflections on and the meaning of the date for the person answering the question. With such questions, absolutely no evaluation of the (self-)reflective effects of history teaching takes place.

All questions in the test under discussion are of this type. More than three items only refer to the dates of World War II, even more to the simple fact that the Nazi-regime was a “dictatorship”. Only one additional name (Stauffenberg, “assassination attempt”) and one additional

5 The whole test was published in many major newspapers, e.g. “Frankfurter Rundschau” (3.9.2008, p. 25-32).

date (9.11.1938, “destruction of synagogues”) are required. In these cases, all false solutions are simply funny “kangaroos”. Of course, such tests say nothing about understanding, connections, contexts or reflections. They are simply a control of a superficial – and superfluous – *naming of history items*, of an ability of pretending to know what the public discussion is about. It is mechanical memorizing (i.e. learning by rote – and not really “by heart”). Since such examinations determine the type of preceding instruction (“*teaching to the test*”) they are detrimental to learning history in a reflecting way.

In 2005, a major newspaper (“Die Welt”) presented a very similar questioning of average German citizens (Die Welt, 25.5.2005, p. 10-15). Questions were, for instance: “*What is meant by the term ‘Holocaust’?*”; “*What happened on 31 January 1933?*”; “*Who carried out an assassination attempt against Hitler on 20 July 1944?*”; “*Which was the first country that Hitler Germany attacked?*”⁶; “*What is the first German republic between 1918 and 1933 called?*”; “*Who was the last German Emperor?*” These are perhaps conventional historical questions (*are they really important?*), but they are obviously not designed to generate thinking or reflection, judgement or balancing, but a simple – possibly pointless – automatism of *stimulus and response*.

One can deduce from the publication that the questionnaire is implemented as an *open answer format*; at least, the percentage of correct answers is reported, but no frequency of particularly erroneous answers. The Holocaust question is answered correctly with “Annihilation of the Jews” by 81,7%, but the name “Weimar Republic” is known only to 52,0% and the “Take-over of Power by the Nazis”⁷ on 31 January 1933 to 48,4%. Poland is mentioned as the first victim of World War II by only 50,2%, Wilhelm II. by 50,5%, and the name of the would-be assassin Stauffenberg by 42,9%.

Nevertheless and astonishingly, most commentators (journalists, bishops, historians, teachers, politicians etc.) expressed great satisfaction about the citizens’ “good” level of knowledge. But in fact, the results of

6 In German, the question is phrased in such a way that it leaves open who attacked whom. The question is even worse, because it does not explicitly mention the context of World War II: Before that war, Germany had already launched military attacks against Spain in 1936, Austria in 1938, Czechoslovakia in 1938 and 1939 and Lithuania in 1939. Thus the so-called *right* answer of the test is quite *wrong*.

7 This is a very problematic expression – by no means innocent or neutral. Contemporaries and historians debated different concepts like “seizure of power”, “transference of power” or “take-over of power”. This controversy is not at all taken into account in the test question, which anticipates the solution in a dogmatic way and forbids or hinders reflection.

this asking for numbers, names, dates and facts without context are very poor (as in all other cases of such type of questioning). This had nothing to do with diagnosis of *historical identity* and *historical competence*.

3. Measuring competences

Let us look at an example of testing an application of historical methods or insights instead of mere reproduction of information. Hans-Jürgen Pandel (2005: 61-62) proposes a nonsense-caricature (“a drawn joke”) with two archaeologists in pith helmets inside an ancient Egyptian tomb, discovering a wall-painting of a man burning a computer. The purpose is testing “consciousness of temporality” and “consciousness of reality”. After being taught about ancient Egypt and its type of tomb paintings, the tasks for the pupils (fifth to seventh grade) are: *“What is incorrect with the picture? Identify the place where the two men are. In which country are they? Which profession do they have? Name the people who may have drawn the wall-painting. The picture is a joke. Why?”*

Pandel hoped that the topic would be adequate for testing “*consciousness of temporality*” (then versus now/in the future) and “*consciousness of reality*” (reality versus imagination/fiction). But he experienced – and complained – that the pupils were not really able to solve the task. However, this is not astonishing at all, since “counter-factual” suggestions and anachronistic errors are rarely, if at all, mentioned or analyzed in history lessons (to say nothing about pith-helmets and short pants). At best, our instruction is only oriented towards re-construction from – reliably true and pre-selected – *primary sources*; at worst, it does not go beyond the *drawing of information* from – unquestioned – *primary sources*. It is unfair to test second order skills or concepts which definitely have not been taught before.

This Pandel-assignment can be done with paper and pencil, as a short essay. Maybe it could even be transformed into a “multiple choice” form:

“Which sentence explains the picture and its information best:

For tourists in Egypt the use of computers is forbidden.

The picture has the character of a caricature and is a joke.

An ancient Egyptian king ordered the destruction of all computers.

Archaeologists found out that already the ancient Egyptians used computers.”

But what is measured by the marking with a correct cross or – more complicated for students and researchers – by a good short essay? There is a certain hint of a conventional methodical competence.⁸ Detecting a major anachronistic mistake – even in the form of a dumb joke – is an achievement. The insight that it is possible for a stupid error to appear in print – even in a serious book – may be even more important. This is a characteristic of critical thinking, though not really specific for the *cognitive domain of history*. But the capacity of narrating and understanding history is not directly examined by the Pandel-assignment. The impact on the self and its historical definition (“self-reflection”) is completely lost.

Another example of the attempt to evaluate historical thinking stems from the group “Chata” around Peter Lee, which attempts to test students’ competences in middle range concepts (*second order concepts*) in the course of a *method-orientated* curriculum (comp. History in the National Curriculum 1995) (Lee/Ashby 2000). In different grades, Lee presents two short versions of the “*End of the Roman Empire*”. He restricts himself to very short versions (30 lines) with comic-style drawings (not with two long chapters and complicated primary sources!). The meaning is rather simple: The first version dates the “fall” in 476, but mentions the existence of Byzantium, the other describes the “fall” in 1453, having spoken of the end in the West before and giving a map of the split Empire. Afterwards the students have to produce short texts or oral presentations about the relation between both narrations.

The test can be done by paper and pencil (Lee’s group used short essays), by an interview or – again – by the transformation into a *multiple choice format*.⁹ But then, more than one sentence can be correct and less than three incorrect.

In the two texts, the meaning of the term “The Fall of the Roman Empire” is a different one.

Both texts contradict each other: The Roman Empire can only have fallen in 476 or in 1453.

Different historians use different pieces of information and therefore draw different conclusions.

-
- 8 Demanding a “short essay” would cost students more time for writing and teachers more time for evaluation. But it would be analytically far superior, because it would show the structure and reasons of students’ insights and errors in detail.
- 9 In this format, the students do not write themselves, but think about solutions. Therefore the risk of omitting the questions – and missing data – is smaller. Nevertheless, the logic of false answers can be analyzed.

Decisions of controversies in history are not so easy; there are no sufficient arguments for 476 or 1453.

Since we weren't there and nobody is an eye-witness of 476 or 1453, we cannot know when the Roman Empire fell.

A "Roman Empire" has to control Rome; therefore "The Fall of the Roman Empire" was in 476.

The Byzantine State is a continuation of the Roman one; therefore "The Fall of the Roman Empire" was in 1453.

The situation is doubtful; we have to look for what the best and most famous historians (and the handbooks) say.

History is written from "primary sources" of contemporaries: We must examine what they say about 476 and what about 1453.

Some of these versions may be doubtful and some may be suggestive (additionally, the correct ones may promote the learning process, which had failed before, in the situation of the test itself). In our context, the other problem is much more important: What does this assignment test? It is focusing on the relation between the past and the historical interpretation/narration. The past was full of innumerable (billions of billions of) particles of reality, most of them unknown today, never having been written down, painted or sculptured – or having since been lost completely.

History is a hypothetical construction and narration about relevant parts of processes in the past with consequences for today. It requires words, notions, terms and concepts which are adequate to the past itself, to us and to our relation with the past. Those concepts have a huge impact on the histories told. To realize this and to apply this thinking to a case of historical interpretation is an important part of methodical competence and concept competence (See also the contribution by Andreas Körber in this volume). It may even have an influence on self-reflection.

Peter Lee's group used the assignment in a short essay version. We should have a brief look at their qualitative and quantitative results, which show the assignment to be a very powerful instrument for diagnosis. Three basic concepts in the answers are distinguished: "factual", "multiple past" and "criterial" (Lee/Ashby 2000: 208).

- In the first concept, the problem is *facts* and *errors*; the two reports about "the end of the Roman Empire" are contradictory. The vast majority of the students (more than three quarters) in second and fifth grade hold this conviction, but one third even in eighth grade.
- The second concept is a medium or transition solution (an example of "seeking" instead of clear "explanation"): History is more complicated and multiple than you think. Perhaps there is a chance to

combine both reports. This is written by tiny minorities in the second and fifth grade, but by larger numbers in sixth and eighth grade.

- The third concept includes a differentiation of the notion “end of the Roman empire” itself: Does it mean “Western Rome” only or “Eastern Rome” (“Byzantium”) as well? Both texts – with different though homonymous/homophone terms – are not at all contradictory. The insight that it is a question of the criteria used is inaccessible to pupils in second and fifth grade, but attainable for 25% in sixth and 40% in eighth grade.

Of course, this is a rather disappointing or discouraging result. What might seem a very simple and logic conclusion to experts suggestively offered by the short texts themselves, is not so obvious in the way lay people relate to history. The theoretical – but empirically substantiated – five levels of “progression in students’ ideas about accounts and their relation to the past” (Lee/Ashby 2000: 212) should be added. This is really a developmental model (among others of “Chata”) which can be applied (perhaps in a revised form) to other concepts of particular historical competencies of the middle range (second order concepts). The progression in students’ ideas about accounts (i.e. historical narrations, B.v.B.) and their relation to the past are listed as follows:

- **The past as given.** Stories are about the same thing: The story is equivalent to something “out there”.
- **The past as inaccessible.** We can’t know – we weren’t there. Nothing can be known. Differences in accounts are a result of lack of direct access to the past.
- **The past as determining stories.** Stories are fixed by the information available; there is a one-to-one correspondence. Differences in accounts are results of gaps in information and mistakes.
- **The past as reported in a more or less biased way.** Shift of focus from the story and reports to the author as active contributor. Differences in accounts are a result of distortion (in the form of lies, bias exaggeration, dogmatism); the problem is not just a lack of information.
- **The past as selected and organized from a viewpoint.** Stories are written (perhaps necessarily) from a legitimate position held by the author. Differences in accounts are a result of selection. Stories are not copies of the past.
- **The past as (re-)constructed in answer to questions in accordance with criteria.** Shift of focus from the author’s position and choice, to the nature of accounts as such. It is the nature of accounts to differ. (Lee/Ashby 2000: 212)

In fact, we do not really know the reasons for the poor outcome mentioned above: Lee/Ashby (2000) suppose that even in the English case of a method-oriented curriculum, the teachers fail to provide sufficient promotion, stimulation and exercise of *historical thinking* during the lessons. One could add: It seems that many teachers are fixed on contents and are sometimes confounding past, facts, interpretation, narrative and history themselves. But that's a supposition only; What to do if teachers have intensively tried to teach *historical thinking* and *reflection*, but pupils prefer *simplicity*, *security* and *un-ambiguity*?

This is only one example of the project "Chata", but it perfectly shows the type of diagnosis conducted by the British researchers and history didacticians. Though a paper-pencil method and partly designed for placing a tick only (see Lee et al. 1998: 237), the approach is vastly superior to the typical German testing, but of course it in no way covers the two main domains of personal identity and long-term biography. Tests cannot validly diagnose the anticipation of future positions and decisions. And they must not grade the normative convictions and existential conclusions of the pupils (as long as the *democratic constitution* and *human and civil rights* are loyally observed).

4. Testing "dealing with history"¹⁰

For the reflecting and (self-)reflective *dealing with history* some qualifications are turning points, i.e. elaborated use of *primary sources* (mostly texts), but also the critical use of historiography or narrations (mostly texts as well), and finally the application to new cases and situations (transfer of insights). At the same time, historical argumentation should be examined, e.g. the distinguishing of *occasion* – or *pretext* – (causality of surface) and *reason* – or *cause* – (causality of deep structure) and additionally the weighting of controversies (especially concerning the structural causes).

The topic "*Great Witch Persecution*"¹¹ has a high status of relevance in Germany, much higher than in other countries.¹² It is quite clear that

10 For the following propositions, I feel inspired and encouraged by two lectures in Goettingen (31.3.-2.4.2008) of the circle of Robert Selman on the one hand and Sam Wineburg on the other hand. Interestingly, the approaches included one project with closed items (Alan Stoskopf and Ulrike Hartmann) and another with open items (Avishag Reisman). The second one will be published soon in Martens e.a. 2010, the first one seems to remain unpublished. None of my proposals is a direct application or translation.

“Witch Hunt” is an excellent *parallel* or *similar* phenomenon to Nazi-crimes – and therefore adequate for an experimental test of transfer abilities.¹³

4.1 Analysis of primary sources

The knowledge and analysis of the perspective character and interest orientation of all *primary source texts* is a very basic point. Therefore, it is fruitful to present a short text of that type (or a picture as well) for evaluative analysis. Short essays are common, but “multiple choice” can be tried as well:

“A Document about the ‘Great Witch Hunt’

Petition of the municipality Elz near Limburg/Lahn in 1589 to the Archbishop and Elector of Trier: ‘ [...] we poor supplicants beseech Your Electoral Grace in the matter of the suspicious and disreputable sorceresses, as we have already at various times most humbly besought Your Electoral Grace for the sake of God, to order Lord Keller to question them [the witches] and execute, what will reveal itself as truth and be declared as justice. Because, most graceful Prince and Lord, should such great and pernicious harm as hath befallen us not be punished, they [the witches] would be strengthened in their designs, and we would be forced to take flight with our wives and children and to leave everything behind.’” (Schormann 1981: 56)

Of course, it is easy to make students evaluate this text with an open formulation, i.e. in a short essay. “Please analyze the position, the interests and the intentions of the authors and reflect which relevance the

-
- 11 Compare a teaching model “The ‘Great Witch Persecution’ (1555-1665) – Death Throes of Middle Ages or Birth Pangs of Modern Times?” in Borries 2003: 177-206.
 - 12 To give British North-America as an example: The *Witch Hunt* in Salem 1692, where the judge officially asked for forgiveness from the victims and the survivors some years later, has been a single case; in Germany, there have been at least fifteen thousand casualties during four centuries (although the *three hundred thousand* or *nine million* sometimes mentioned are apparently overestimated by far).
 - 13 Some years ago, large groups of students (6th, 9th and 12th grades) were questioned about the “Great Witch Hunt”, especially about their emotions, partisanships, abilities of empathy, moral judgements, and conclusions for today (see Borries et al. 1992: 58-63, 200f.). Some of the strategies of this study can be transformed into a testing of historical reflections as well.

quoted contemporary evidence can have in respect to the explanation and judgement of witch trials."¹⁴ The short interpretations could read:

- "The Elector of Trier puts the responsibility for his cruel acts against the supposed 'witches' at other people's door, like his council and his subjects."
- "The applicants are very sure on the subject of 'witches' and act rather decidedly and cunningly: By threatening to run away (emigrate) if the prince doesn't punish the witches they are in effect blackmailing him."
- "Hatred between Catholics and Protestants was one of the main reasons of the Great Witch Hunt in the 16th and 17th century."
- "Influential subjects themselves repeatedly demanded the persecution of 'witches' from their authorities. According to their conviction, these 'witches' caused them much harm through sorcery."
- "If such letters are frequent, one could conclude that there was a deep-rooted fear and bitter hatred among the normal people in the villages."
- "The rulers are by no means the only ones responsible for or guilty of the witch hunt. They have to take the convictions, wishes and feelings of their subjects into account."
- "The practice of accusing individuals of sorcery seems to have been something so frequent and normal that one can conclude: Attempts to perform witchcraft – at least making evil wishes or casting the evil eye – must have been common then."
- "Even in the dark, superstitious times of the 16th century, enlightened people often warned against witch hunts and the torture and cruel burning of innocent women."

Apparently, here we have the interesting case of complexity and controversy: Not one version is completely and definitely correct. And most of the others cannot be taken as undoubtedly erroneous or absurdly remote. Instead, different levels of understanding are possible, which may be linked to competences. But such items demand remarkable methodical

14 But it is also possible, to write a set of interpretations or commentaries about the text (more or less correct and complex) and to make the students rate these texts with a five step Likert scale:

Completely unplausible; Mainly unplausible; So-so; Mainly convincing; Completely convincing

Obviously, some critics will argue that it would be fair to add an additional answer "*don't know*"; but the interpretation of a primary source text is not a task to refuse legitimately. For the ignorant and anxious ones, the existence of a version "*so-so*" is already an offer (without loss of data for further calculations).

standards, first in the process of construction of the items by the researchers. It would be necessary to make small experimental groups write short essays, to analyze and classify those texts and to transform some typical remarks into test items. Without a qualitative pilot study such quantitative research is methodically (and normatively) impossible or at least unwelcome.

These experimental versions should be tested (tried) with larger samples. In this way, the earlier mentioned suggestions about levels of difficulties and levels of competences in the different versions can be verified: Have the students' articulations been predicted correctly? Thus, in the next phase, the evaluation of the students is less important than the verification or falsification – and revision – of precedent theories about levels of difficulties, learning processes and the logic of development of competencies.

4.2 Production of historiography

On the other hand, such items, even if answered correctly, do not guarantee *historical understanding in a comprehensive sense*, but only the *correct understanding of particular historical texts*. This is a very important difference. The mentioned achievements, which have the character of *criticism of primary sources*, could be introduced and exercised in a conventional source-oriented teaching. Of course, this is not meant negatively. But more is needed for historical understanding, i.e. the production of one's own historical description/narration and the analysis of divergent historical descriptions/narrations (this is called *competence of de-construction*).

What does that mean for our example? Undoubtedly, a small set of primary sources (two texts, two quotations from the Bible, two woodcuts) had to be combined. To remain realistic, the whole collection could not exceed two or three printed pages in a testing situation. Additionally, the learners should first formulate questions on the material and then try to give a written presentation. It is nearly impossible to imagine this in a format of closed questioning (e.g. multiple choice); but in view of the limited time available, requiring the learner to create a *mind map* could be a better solution than demanding a fully elaborated *short essay*.

It would be preferable to present a larger collection of diverse primary sources (pictures and texts), some theories, some transfers to the present (propositions of *historical orientation*), different forms of fiction (pictures and texts), different narrations (pictures and texts), and to give the students the task to produce their own sense-making narration (with legitimation and argumentation) on the basis of this collection. This

would of course take at least three hours. We have occasionally carried this out at university level (see Borries 2006, 2007). The results have been rather problematic, even discouraging. Apparently, university students interpret these demands as highly overburdening, but also as provocative.

Apparently, written history class tests and even essays under exam conditions at university usually follow the model of a simplifying analysis of primary sources and content problems. In fact, it seems altogether uncommon to ask for the production of a synthetic sense-making narration or a critically qualified judgement about other pre-fabricated narratives. Thus, the usual exam question seems just an unfortunate mixture of arbitrary – and relatively simple – parts falling between what have been termed competence of “re-construction” and the competence of “de-construction” (Körber et al. 2007; Schreiber/Körber et al. 2006; See also the contribution by Lenz in this volume).

4.3 Analysis of historiography

The next task is of course just the opposite one: short narrations of the “Witch Hunt” have to be compared and evaluated.¹⁵ Again, a series of pictures from the 19th century – or a memorial of the 20th – could be integrated and added to two popular texts. Surely, such a request for de-construction will first be arranged with open answers (short essays). But why not try to construct a set of prototypical stages from typical – more or less elaborated – articulations among experimental groups? These could be presented to learners for a rating via ticking boxes. One could also attempt to correlate particular positions to different political sympathies. As already said, it is especially important to distinguish “causes” and “occasions” and to combine and discuss several incidental and also structural interpretations.

As a substitute, I have listed six – basically contradicting – models of explanation of the “Great Witch Persecution” (see Borries 2003: 196-199). Since the list is shortened to half its length, they are probably inadequate in respect to their amount, but perhaps overburdening in respect to their degree of difficulty. The focus is not on the narration itself (it easily becomes too long), but on the theoretical ascription of causalities (background and deep-structure reasons). Here I quote only two – especially contradicting – examples:

15 In fact, because of the limitations on time in a testing situation, only contradicting short judgements can be presented.

Six Explanations of the “Great Witch Persecution”

Thesis 1: *The obsessive belief in witches is a typical spawn of medieval superstition and dark sadism. It is based on pathological notions in the heads of fanatic theologians, who were enemies of science, and men of law, who were blind to reality. Thus, deviant minorities and autonomously thinking people could be intimidated and eliminated. It took a very long time until modern Enlightenment drove away these haunting ghosts and replaced fear of the supernatural by rational scientific explanations: “Since I assume the vice of witchcraft to be a myth, I advise this one precaution only: A prince should never allow an investigation to be conducted [...] on the grounds of the vice of witchcraft, i.e. the alliance of human beings with the devil” (a philosopher of the Enlightenment, 1701).*

Thesis 6: *The witch trials primarily served religious and secular authorities to discipline and intimidate the rural subjects. Previously, priests and noblemen had tolerated all kinds of superstition and magic in the almost independent villages: “Often men and women came to see her (an old woman) in her house, begging her to release their friends, who had been bewitched, from the curse [...]. She can bring a husband back to a woman” (a witness from 1446). Henceforth, the soul was to belong exclusively to the church, the body unconditionally to the king. The persecution of witches was the most brutal phase of a long struggle, at the end of which man came to be shaped and moulded to the requirements of the modern state, conformist, manipulated and controlled. It is precisely this that constitutes the mental aspect of the long transition from medieval feudalism to modern capitalism. In that sense, the witch hunts are a spawn and a manifestation of modern rationality and of a strategy of modernization.*

Again, one could choose an open-question format, i.e. a list of questions like the instructions for working with textbooks. In this way, an evaluation of the theses can be promoted and demanded. Afterwards the outcomes can be rated/coded according to strict criteria, so that even quantitative studies are possible.

“Which questions do you formulate after reading the six theses? By which procedures do you try to answer them? Which additional information would you like to have and use? Where would you have a chance of accessing it? How do you provisionally assess the mutual connection of the six theses and their peculiar persuasiveness?”¹⁶

16 An alternative, more detailed order – with other focuses and more help – could read:

Such a qualitative preliminary stage with an experimental group is unavoidable in the case of larger investigations. Afterwards, an attempt with closed items – and therefore quantitative methods – can be undertaken.¹⁷ But for the moment it is not really clear how far it will be successful in testing the ability for historical reflection.

At the end, a request to rank the statements in an explicit order could be added:

“Please rank the six theses suggesting an explanation of the ‘Great Witch Hunt’ according to: which one you consider the most convincing(1), which one the second most (2), which one the least (6), which

“Please, discuss the six theses and decide – as far as possible – to which extent they are valid for you. The following questions may be helpful:

- *Where is the persecution of witches said to be typically ‘medieval’, where specifically ‘modern’? On which side is ‘rationality’, on which ‘superstition’, found in the respective explanations?*
- *To which degree are general social and mental mechanisms described, which can produce similar phenomena in all situations and epochs? To which degree are very specific circumstances of the late 16th and early 17th century mentioned and made responsible?*
- *Which theories assume ‘witchcraft’ to be a real phenomenon (though not a crime worthy of death sentences)? Which ones declare it to be an invention, a myth or a figment of the imagination of the persecutors or the persecuted?*
- *Which explanations remain general and neutral towards gender? Which give reasons for the fact that mainly women were accused and executed?*
- *Who is the main blame for the large-scale judicial murder attributed to? Please analyze in particular whether the activities are assigned to the ‘authorities’ or to the ‘subjects’ themselves!*
- *Which explanations can be combined with each other, which ones contradict others directly and unequivocally? Discuss the scope and the persuasiveness of each theory!”*

17 Every statement could be rated by the students according to several categories. Here is a sketchy draft (valid for all six theses):

“As an explanation of the ‘Great Witch Persecution’ thesis 1 [2, 3 etc.] is (answer only ‘Yes’, ‘Un-decided’ and ‘No’) [...]

a)[...] convincingly argued and consistent in itself

b)[...] well compatible with the other statements

c)[...] completely arbitrary and rather far-fetched

d)[...] intelligently concluded from the known facts

e)[...] strongly biased in the interest of certain involved groups

f)[...] irrelevant for us today - with our different rules and habits

g)[...] worked out perfectly from the contemporary conditions of those times

h)[...] much too lenient towards the crimes of those who committed them at the time

i)[...] perfectly uncommon and unexpected, but productive and enlightening”

one the second least (5) and which one you would put in the middle (3, 4).

Thesis 1 Thesis 2 Thesis 3 Thesis 4 Thesis 5 Thesis 6"

4.4 Transfer achievement

Obviously, the fourth step has to be the application (transfer). In order to show (and measure) the qualifications described above, it will however be necessary to use new topics instead of recurring to already familiar ones. Therefore it is important to rely on similar or equal categories (e.g. "cause" and "occasion", "perspectivity", "controversy"). If, for instance, the subject matter "witches" has already been discussed in the classroom, it would not be appropriate to use it for the following exam or test. Instead, the topic could perhaps be a xenophobic riot or an anti-Semitic pogrom, violent protests against the Huguenot immigrants in Germany around 1720 or the shooting of Belgian hostages by the Germans at the beginning of World War I in 1914. Nevertheless, in its structure, the questionnaire could, even should look like the one about the "witches" described above.

When measuring the transfer, new problems arise. What do researchers really measure when they expect insights into deep structure causes may be gained from the "Witch Hunt", but are used for "Persecution of Jews 1933/45" or "Xenophobic Riots 1993"? Is this really *historical thinking*? By generalizing a certain figure or mechanism and making it obligatory or transferable to other cases, we have to exclude changes and fix or prolong ongoing rules automatically (*standstill*). This suggests *purely exemplary sense-making* as Rüsen understands it (1994b: 37-41, 85-90, 150-155, 231-234) one might call it *social-psychological insight* as well, which implicitly assumes that the logic of social behavior has not changed in the course of long historical processes, not even during the process of modernization and in times of pressure and crisis.

As historians, we will find this restriction problematic (but that is a general dilemma of presence-related historical learning). Without a *genetic sense-making* – as well as a *critical* one, which means distancing oneself from an accepted interpretation – there is no chance at all (Rüsen 1994b). But how can these figures be tested and how can sentences of genetic character (e.g. "in fact at that time [...], but today [...]" or "already then [...], and even more today [...]") be classified as "correct", "so-so" or "false"? Admittedly, with this point, we are not only in the center of a sensitive question of diagnosis, but in a mine-field of theory itself: Which status can *exemplary* – and *traditional* – sense-making preserve in a world with unavoidably dominant *genetic* interpretations?

5. Final remarks

As already mentioned, the possibilities of testing by paper-pencil formats are much greater than the current practice in most countries. This paper was mainly about experiments which should be performed by supporters of reflecting and (self-)reflective history learning, before more and more boring, anti-reflective knowledge-tests (see above) flood the market. According to my experience, most teachers are critical towards tests (e.g. multiple choice formats), but not towards their mere content-based – and sometimes dogmatic unambivalent – logic itself. It is very important to know and to practice more intelligent and more elaborated paper-pencil-questions even of closed types (crossing and ranking), but of open types (mental maps, short essays) as well.

Apart from paper and pencil, we have other opportunities for diagnosis during the teaching itself, which are more suitable for goals that are not operationalized or cannot be operationalized. In projects, pupils can be observed while they are investigating, discussing, reflecting, presenting, evaluating, comparing, judging, transferring to other cases and personal life etc. Even (very cautious) assumptions can be made about their future behavior in a long-term perspective (life and biography), and their sense-making in private, communicative and public situations may be improved to some extent by these merely qualitative methods (which are often not compatible with exam *grades*). The simple reason is that the young people have already tried and executed such processes themselves and can transfer the competences to other situations and use them successfully. To describe these observations which cannot be quantified or objectified mechanically, in detail, would be the subject of another article.

References

- Borries, Bodo von (1992) (in cooperation with Dähn, Susanne, Körber, Andreas and Lehmann, Rainer H.): *Kindlich-jugendliche Geschichtsverarbeitung in West- und Ostdeutschland 1990. Ein empirischer Vergleich* (= *Geschichtsdidaktik. Studien, Materialien. Neue Folge* 8), Pfaffenweiler.
- Borries, Bodo von (1997): *Wendepunkte der Frauengeschichte II. Über Muttergöttinnen, Männeransprüche und Mädchenkindheiten. Modelle und Materialien zum Ausprobieren und Bessermachen*, Centaurus, Herbolzheim.

- Borries, Bodo von (2004): *Lebendiges Geschichtslernen. Bausteine zu Theorie und Pragmatik, Empirie und Normfrage* (= Forum Historisches Lernen), Wochenschau Verlag, Schwalbach/Ts.
- Borries, Bodo von (2006): “‘Fremdverstehen’ – ‘Empathieleistung’ – ‘Abenteuerfaszination’? Zu Chancen und Grenzen interkulturellen Geschichtslernens”. In: Manuela Boatcă et al. (eds.), *Des Fremden Freund, des Fremden Feind. Fremdverstehen in interdisziplinärer Perspektive*, Waxmann, Münster, pp. 65-84.
- Borries, Bodo von (2007): “‘Geschichtsbewusstsein’ und ‘Historische Kompetenz’ von Studierenden der Lehramter Geschichte”. In: Gerhard Henke-Bockschatz (Mod.), *Geschichtsdidaktische empirische Forschung* (= Zeitschrift für Geschichtsdidaktik 6. Jg. Jahresband 2007), Wochenschau, Schwalbach/Ts., pp. 60-83.
- Borries, Bodo von (2008a): *Historisch Denken Lernen – Welterschließung statt Epochenüberblick. Geschichte als Unterrichtsfach und Bildungsaufgabe*, Budrich, Opladen und Farmington Hills.
- Borries, Bodo von (2008b): “‘Orte’ des Geschichtslernens – Trivialität oder Schlüsselproblem?”. In: Saskia Handro/Bernd Schöнемann (eds.), *Orte historischen Lernens*, LIT, Berlin, pp. 11-35.
- Die Welt: “Was die Deutschen über ihre Geschichte wissen”, 25.5.2005, pp. 10-15.
- Frankfurter Rundschau: “Wie deutsch sind Sie”, 3.9.2008, pp. 25-32.
- Fußmann, Klaus/Grütter, Heinrich Theodor/Rüsen, Jörn (eds.) (1994): *Historische Faszination. Geschichtskultur heute*, Böhlau, Köln.
- History in the National Curriculum. England; London (HMSO) (1995).
- Jeismann, Karl-Ernst (1985): *Geschichte als Horizont der Gegenwart*, Schöningh, Paderborn.
- Jeismann, Karl-Ernst et al. (1987): *Die Teilung Deutschlands als Problem des Geschichtsbewußtseins. Eine empirische Untersuchung über Wirkungen von Geschichtsunterricht auf historische Vorstellungen und politische Urteile*, Schöningh, Paderborn.
- Körber, Andreas/Schreiber, Waltraud/Schöner, Alexander (eds.) (2007): *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik* (= Kompetenzen: Grundlagen – Entwicklung – Förderung, vol. 2), ars una, Neuried.
- Lee, Peter/Ashby, Rosalyn (2000): “Progression in Historical Understanding among Students Ages 7-14”. In: Peter N. Stearns/Peter Seixas/Sam Wineburg (eds.), *Knowing, Teaching, and Learning History*, New York University Press, New York, pp. 199-222.

- Lee, Peter/Dickinson, Alaric/Ashby, Rosalyn (1998): "Researching Children's Ideas about History". In: James F. Voss/Mario Carretero (eds.), *Learning and Reasoning in History* (= *International Review of History Education* 2), RoutledgeFalmer, London, pp. 227-251.
- Martens, Matthias et al. (eds.) (2010): *Interpersonal Understanding in Historical Context*, Sense Publishers, Rotterdam, Boston, Taipei.
- Pandel, Hans-Jürgen (2005): *Geschichtsunterricht nach PISA. Kompetenzen, Bildungsstandards und Kerncurricula* (= *Forum Historisches Lernen*), Wochenschau, Schwalbach/Ts.
- Ricoeur, Paul (1988/1989/1991): *Zeit und Erzählung*, 3 vol., Wilhelm Fink, München.
- Rüsen, Jörn (1983): *Historische Vernunft. Grundzüge einer Historik I: Die Grundlagen der Geschichtswissenschaft*, Vandenhoeck & Ruprecht (Kleine Vandenhoeck-Reihe; 1489), Göttingen.
- Rüsen, Jörn (1994a): *Historische Orientierung. Über die Arbeit des Geschichtsbewußtseins, sich in der Zeit zurechtzufinden*, Böhlau, Köln.
- Rüsen, Jörn (1994b): *Historisches Lernen. Grundlagen und Paradigmen*, Böhlau, Köln.
- Schormann, Gerhard (1981): *Hexenprozesse in Deutschland*, Vandenhoeck & Ruprecht (Kleine Vandenhoeck-Reihe; 1470), Göttingen.
- Schreiber, Waltraud (mod.) (2003): *FUER Geschichtsbewusstsein. Ein internationales geschichtsdidaktisches Forschungsprojekt zum Geschichtsunterricht* (= *Zeitschrift für Geschichtsdidaktik*, 2. Jg. 2003, Jahresband).
- Schreiber, Waltraud/Körber, Andreas/Borries, Bodo von/Krammer, Reinhard/Leutner-Ramme, Sibylla/Mebus, Sylvia/Schöner, Alexander/Ziegler, Béatrice (2006): *Historisches Denken. Ein Kompetenz-Strukturmodell* (= *Kompetenzen: Grundlagen – Entwicklung – Förderung*, vol. 1), ars una, Neuried.

