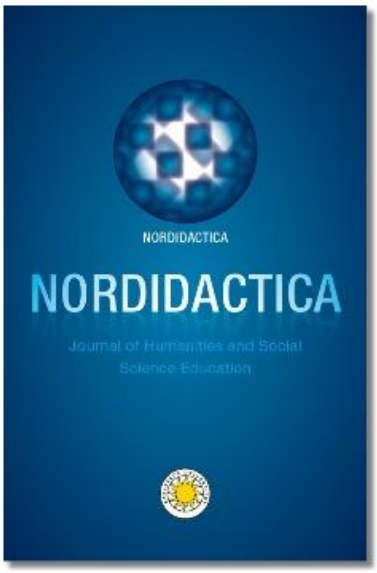
Designing assessment criteria for document-based history essays

Sari Sulkunen, Minna-Riitta Luukka, Anna Veijola & Johanna Saario



Nordidactica

- Journal of Humanities and Social Science Education

2026:2

Nordidactica – Journal of Humanities and Social Science Education
Nordidactica 2026:2
ISSN 2000-9879
The online version of this paper can be found at: www.kau.se/nordidactica

NORDIDACTICA – JOURNAL OF HUMANITIES AND SOCIAL SCIENCE EDUCATION
ISSN 2000-9879
2026:2 95-118

Designing assessment criteria for document-based history essays

https://doi.org/10.62902/nordidactica.v16i2026:2.27702

Sari Sulkunen, Minna-Riitta Luukka, Anna Veijola & Johanna Saario

University of Jyväskylä

Abstract: In Finnish history curricula, the paradigm shift from traditional teaching to teaching historical thinking and skills has been recently strengthened by the cross-curricular competence of multiliteracies which translates into teaching disciplinary literacies in all school subjects. Still, research shows that history teaching in Finnish schools is fairly content-oriented, and there is need to develop pedagogical approaches and tools for teaching historical literacy practices. In this study, we examine the process and product of design-based development of criteria for evaluating source-based history essays. The development process was realized as co-design with a group of history teachers in a mid-sized Finnish upper secondary school. First, we introduce the design procedure and compare different versions of the criteria to provide insight into the process. Second, using content analysis of teachers' interview data we examine how the collaborative effort worked for the teachers. Third, based on student questionnaire data (n=32), we study how accessible the criteria were to students. Findings show that the co-design process offered a valuable opportunity for the teachers to reflect on their assessment practices and the criteria were quite accessible for students. The value and limitations of the findings are discussed.

KEYWORDS: HISTORICAL LITERACY, ESSAY, ASSESSMENT, CRITERIA

About the authors: Sari Sulkunen, PhD, is a professor in Finnish language and mutiliteracies at the University of Jyväskylä, Finland. Her current research interests include teaching disciplinary literacies and blue-collar workers' professional multiliteracies in construction industry.

Minna-Riitta Luukka, PhD, is a professor in Finnish language at the University of Jyväskylä, Finland. Her current research interests include mother tongue teaching and learning, multiliteracies and assessment of writing skills.

Anna Veijola, docent, PhD, is a teacher of history, social studies, and philosophy, as well as a teacher educator at the University of Jyväskylä Teacher Training School. Her research interests include the pedagogy of history, historical literacy, and critical thinking.

Johanna Saario is a lecturer in Finnish as a second language at the University of Jyväskylä, Finland. Her current research interests include teaching disciplinary literacies and learning disciplinary literacies in a second language in a multilingual environment.

Introduction

In history teaching, there has been a paradigm shift from traditional history teaching to teaching historical thinking and reasoning. This has meant that in addition to studying factual knowledge students are to develop also their skills in historical inquiry (van Drie and van Boxtel, 2008; Nokes, 2013). This skill-based approach has been visible in Finnish history curricula since 1980's (Veijola, 2016), and more recently it has been further strengthened by the cross-curricular competence of multiliteracies which translates into teaching disciplinary literacies in all school subjects both in basic education and the Finnish general upper secondary school (Finnish National Agency for Education, 2014; 2019).

The curriculum of the general upper secondary school, for example, stipulates (Finnish National Agency for Education, 2019, p. 280) that history teaching aims at strengthening students' historical thinking which entails understanding the interpretative nature of historical knowledge and mastering historical literacy practices. Veijola (2016) notes that while the relationship between historical thinking and historical literacy can be defined in multiple ways, teaching historical thinking is usually realized by teaching literacy practices with primary sources. Indeed, one of the concrete aims of history teaching is that students can construct knowledge about the past using various source materials in a critical way (Finnish National Agency for Education, 2019, p. 280). Moreover, in the Matriculation Examination, a high-stakes assessment at the end of the general upper secondary school, the tasks require the ability to interpret and critically evaluate historical sources and compare interpretations and their validity (Virta, 2014; Ylioppilastutkintolautakunta, 2023) albeit to limited extent (Puustinen, Paldanius and Luukka, 2020). Previous research shows that history teaching is still content-oriented (Rantala, et al., 2020; Sulkunen, et al., 2019) and even when historical literacy practices are included, instruction for them is often scarce and implicit, and takes place as a response to students' queries or after assignments (Sulkunen and Saario, 2019). Thus, there is need to develop pedagogical practices in teaching historical literacy.

In Finnish history teaching a typical assignment is writing an essay. In a survey for general upper secondary school history teachers, the essay proved to be frequent assignment in many classrooms, and it was particularly frequently used in skill-based teaching (Sulkunen, et al., 2019) in which the essay is useful in showing both mastery of the content knowledge and demonstrating capabilities for historical inquiry: in essay assignments students are asked to interpret primary sources, reflect on the causes and consequences of historical events, or otherwise demonstrate historical thinking. Essays are also considered to be a good way of assessing the ability to reason and communicate historical knowledge. Thus, this type of assignment is often used in assessing students' performance. In such contexts, it is essential to make clear to students what is expected of them, and one way of doing that is by assessment criteria (Arter and McTighe, 2001; Reddy and Andrade, 2010). It has been noticed that the assessment guidelines of Matriculation Examination focus mainly on content and not as much on historical thinking and literacy skills (Löfström and Kaarninen, 2013; Puustinen, Paldanius and

Luukka, 2020). If the guidelines do not make these aspects sufficiently explicit, the student may be rewarded for describing the material and demonstrating only their mastery of content knowledge. This prompts the following questions: What kinds of assessment criteria would effectively communicate to students the expectations regarding not only content but also historical literacy, and how could these criteria be developed considering teachers' practical needs and expertise, as well as research on teaching historical literacy?

In this article we – researchers in a research project focusing on teaching historical literacy in general upper secondary school – address these questions by examining the process and product of a design-based development of assessment criteria for source-based history essays implemented as a co-design with a group of history teachers in a Finnish upper-secondary school. More specifically, our research questions are: 1) how this collaborative effort worked from the perspective of teachers, and 2) how accessible the criteria developed were to students. Our aim is to support developing skill-based history teaching stipulated in the curriculum and to provide teachers with concrete tools for teaching and assessment.

Developing the assessment criteria was implemented as a part of a research project "Engaging in disciplinary thinking: historical literacy practices in Finnish general upper secondary schools" (HisLit), which was a project funded by the Research Council of Finland in 2016–2021 (PI Minna-Riitta Luukka). The research team was a multidisciplinary consortium with linguistics and literacy scholars, historians, and educational scientists. Using multiple data and methods, such as classroom observation and video data, design-based tasks, interviews, and surveys, the project examined what kinds of texts students engage with while studying history, how adequately they master historical literacy practices, how historical literacy is taught, learned and assessed and what kinds of pedagogical and assessment designs can be developed to support the mastery of disciplinary literacy practices. This sub-study was implemented by the team of researchers with expertise in literacy studies and history teaching.

Assessment criteria for historical literacy

In this study, we focus on historical literacy, and more specifically literacy practices required in writing a source-based essay. Theoretically, we approach historical literacy in the framework of disciplinary literacy (Goldman, et al., 2016; Moje, 2015) in which literacy practices are negotiated by the disciplinary community and intertwined with discipline's knowledge construction practices. In history, teaching disciplinary literacy means engaging students with historical inquiry using various source materials as evidence and understanding that historical evidence is biased, incomplete and often conflicts with other sources (Goldman, et al., 2016, p. 234). Working with source materials involves certain literacy practices, namely sourcing evidence by identifying the author's intentions and biases, contextualizing source materials by considering their historical context and corroborating and verifying interpretations (Wineburg, 1991; see also Goldman, et al., 2016; Nokes, 2013, pp. 24–25). These literacy practices are

specific to the discipline of history, and we thus refer to them as historical literacy practices.

In the field of literacy education, teaching disciplinary literacy is a rather recent approach which has been introduced in the 2000s (Moje, 2015; Derlinger Kane, Ortlieb and Cheek, 2024). Prior to that, particularly in the U.S., there was a long tradition of teaching general literacy skills adaptable across the curriculum, often characterized as 'content area literacy'. In practice, this meant for example teaching reading comprehension strategies (Shanahan and Shanahan, 2012; Derlinger Kane, Ortlieb and Cheek, 2024). These two approaches – disciplinary literacy and content-area literacy – have often been seen as opposites. However, many scholars rather see them as continuum of approaches that complement each other (e.g. Brozo, et al., 2013; Moje, 2015). Shanahan and Shanahan (2008) present a three-stage model in which disciplinary literacy is the highest stage with most specialized literacy practices. The intermediate level prior to that include generic comprehension strategies and other generic literacy practices. At the bottom are basic literacy skills, such as decoding and vocabulary, as they are the foundation for the intermediate and high practices. In our data, we see traces of both the generic and disciplinary approaches.

Developing history teaching towards skill-based approach that includes historical literacy requires updating content and methods of teaching towards procedural skills and historical inquiry but also updating assessment criteria so that they meet the learning objectives and outcomes (Bearman and Ajjawi, 2018, p. 1; Zhao, 2022). From the perspective of assessing historical literacy, tasks that require historical thinking and historical literacy can be more challenging and subjective than assessing tasks that focus solely on content and right or wrong answers. Criteria for an assignment that integrates content knowledge and literacy practices need to explicate not only content-related (Breakstone, 2014) but also epistemological aims (Pitt and Quinlan, 2021) and aims related to disciplinary language use (Humphrey, et al., 2010; Pessoa, Mitchell and Miller, 2017).

Assessment criteria are an essential tool in criterion-referenced assessment that focuses on specific learning objectives and measures students´ performance based on predefined criteria (Lok, McNaught and Young, 2016). This method provides a picture of what a student has mastered and what areas need improvement. Criterion-referenced assessment has become more common in recent years in Finnish schools (Atjonen, et al., 2019). In Finnish general upper secondary education, there are no grade criteria defined in the curriculum. However, the curriculum does set objectives for history teaching that the teacher must assess. Since essay is a typical task format in Finnish history assessment, it is essential to consider the criteria for evaluating this type of assignment.

Using criteria as a basis of assessment has been shown to have many benefits. Firstly, they can help teachers make learning outcomes concrete and communicate to students what is expected from them in a certain task (Arter and McTighe, 2001; Reddy and Andrade, 2010). Secondly, criteria influence the quality of assessment since they help the teachers be more accurate, consistent, and unbiased in scoring (Monte-Sano and De La Paz, 2012; Radinsky, Goldman and Pellegrino, 2015). Thirdly, the criteria also

allow the teacher to justify his/her assessment to the student (Arter and McTighe, 2001). Typically, the assessment criteria include the *evaluation criteria* which depict the factors defining the quality of students' work, *quality definitions* describing each criterion on different performance levels and the *scoring rubrics* for the students' work (Reddy and Andrade, 2010). Together these three elements are often referred to as assessment rubrics that are analytical rather than holistic, and either generic or task-specific (Arter and McTighe, 2001). In this study, we mainly focus on developing the evaluation criteria that history teachers agree upon and find useful and suitable for communicating the task requirements to students. Also scoring rubrics were developed for the pilot study but quality definitions were beyond the scope of this study.

Based on the review of 75 articles, Jonsson and Svingby (2007) concluded that the use of analytic and topic-specific assessment rubrics improves the reliability of assessments. This is especially true when accompanied by examples of various levels of performance or when users receive training on how to use the rubrics. Rubrics have also been shown to improve student performance. However, according to the literature review by Panadero and Jonsson (2013), the research on their effectiveness is not consistent as some studies report significant improvements, while others show only minor or no changes. Particularly generic criteria used across tasks have shown inconsistent results as they are challenged by varying task-types (O'Donovan, Price and Rust, 2004; Rust, Price and O'Donovan, 2003; Zhao, 2022). Zhao (2022) presents that to cover multiple tasks generic criteria may include vague wordings which may result in challenges understanding them. This suggests the need for task-specific criteria. In our study, the criteria developed are task-specific as they aim at evaluating source-based essays in history.

To meet the communicative purposes mentioned above and the requirement of transparency (Bearman and Ajjawi, 2018), criteria need to be clear and comprehensible, i.e. accessible (Zhao, 2022). In Zhao's (2022) study examining undergraduate students' perceived accessibility of assessment criteria, accessibility was operationalized as how clear and easy to understand the criteria are. The results showed that the higher the level of clarity of the criteria and the easier they were to understand, the higher the students' understanding of assessment requirements, feedback and grades, as well as their ability to apply the feedback in future assignments. In this study, we examine students' perception of the accessibility of the criteria and operationalize it in terms of comprehensibility of communicating what is expected in the assignment (source-based essay).

The starting point for developing criteria is to consider the properties of historical thinking and to transform them into skills to be assessed. Hardy and Iwatani (2021) for example investigated several U.S. frameworks to operationalize criteria for assessing students' course work. They ended up with seven major elements: historical argumentation, causation, comparison, contextualization, continuity and change over time, and sourcing. Historical thinking skills often overlap with general literacy skills and therefore they did not include reading or writing as distinct categories in the rubrics. In Finnish context of history teaching, the primary goals of assessment are related to mastering the knowledge and skills specific to the discipline. These include the ability

to comprehend temporal relations, causal links and consequences, acquire information from various sources, distinguish between relevant and irrelevant information, and critically evaluate historical phenomena, interpretations, and the use of historical knowledge (Finnish National Agency for Education, 2019).

Using criteria is not a guarantee of valid, reliable assessment or pedagogically high-quality teaching. The use of criteria can yield reliable results only if they and performance-level descriptions are clear and focused (Brookhart and Chen, 2015). Moreover, short-time interventions in using assessment criteria alone do not seem to deepen understanding of assessment practices but teachers need training in using them (Panadero and Jonsson, 2013). Yet, criteria need to be introduced to students in multiple ways. In Zhao's (2022) study, students called for dialogue with tutors or teachers about the criteria, concrete examples on how to meet the criteria in the assignment and ways to contextualize the criteria in a written text. In this study, when the newly developed criteria were piloted, the teacher introduced the criteria to students in the context of presenting the assignment in class and explained verbally what each criterion meant.

Co-design process, data and methods

In the third year of our project on historical literacy (2019) we had collaborated with a medium-size upper secondary school's history teachers by collecting various types of data on several occasions, such as classroom observations, video and teacher interview data. At this point, we wanted to turn our attention to teachers' needs and do something with the teachers that would benefit them directly. Together with the history teachers, we decided to develop criteria for evaluating source-based historical essay. This collaboration was implemented as design-based research (Edelson, 2002), and more specifically as co-design in teacher-researcher partnership (TRP; Juuti, et al., 2021). Typically, the co-design is characterized by collaboration to solve practical problems or issues and developing design-based improvements for educational practice (Coburn and Penuel, 2016, p. 48; Henrick, et al., 2017), and the process starts with empirical problem analysis or needs analysis (Aksela, 2019).

The teachers participating in the co-design (n=6) worked in a mid-sized Finnish general upper secondary school. General upper secondary education provides students with knowledge and competencies for further studies (Finnish National Agency for Education, 2026). Studies usually last for three years, and they include 6 compulsory credits and 6 optional credits of history (of total of 150 credits). At the end of the studies students take the Matriculation Examination in which history is optional.

Juuti et al. (2021) emphasize that in co-design process teachers are to be equal partners in the research team with shared goals. In our partnership, teachers had a key role in the empirical problem analysis phase in determining which practical issue we start solving and thus the goals for our collaboration. Moreover, while the researchers organized the work, teachers' input was essential in developing and piloting the criteria, and they were considered experts in teaching historical literacy during the key phases of the process. This was reflected, for example, in the group discussions on the criteria

draft in phase 2 when teachers' suggestions were accepted even if they contradicted those of the researchers (see next chapter on co-designing the criteria). Furthermore, the group discussions were characterized by an informal atmosphere, fostering an environment conducive to open exchange of ideas. However, in design-based research it is challenging to achieve fully equal partnership as researchers' knowledge of theory, prior research and conducting research limits teachers' influence on the direction of the research process.

Design-based research provides information on the development process, the product developed, and the background theory (Edelson, 2002). While the whole process started with empirical needs analysis with the teachers, the co-design process was also informed by theory and previous research, as presented in the previous chapter. This was visible in the theory-driven analysis of teachers' group interviews and discussions in which researchers' knowledge on historical literacy and thinking (Goldman, et al., 2016) and general literacy skills (Shanahan and Shanahan, 2008) contributed to the analysis. Moreover, literature on accessibility of evaluation criteria guided their formulation of the criteria and the pilot study. During the co-design process, collaboration with researchers supports teachers' professional development (Coburn and Penuel, 2016), and in the process reflecting experiences is crucial for learning (Mansvelder-Longayroux, Beijaard and Verloop, 2007). In our collaboration, group interviews and discussions offered teachers time and place to reflect the collaborative development process. Next, we will outline the co-design process in detail (figure 1).

The co-design is an iterative process (Edelson, 2002; Aksela, 2019). In the first phase, history teachers evaluated and scored individually six document-based essays following their usual criteria which were based on their understanding of implicit learning objectives determined by curricular goals and expectations in the Matriculation Examination. These criteria did not exist in a written document shared with the students, and thus they could have had individual emphasis. The essays were delivered by one of the teachers in the group, and students had written them as part of a regular assignment in class. The source material in this assignment consisted of Martin Niemöller's poem written after WWII. Students were asked to consider, using the poem as a source, why Nazi party succeeded in getting to power in Germany, and to evaluate how holocaust implemented by the party should be seen today. After teachers scored the essays independently using their typical individual criteria, we organized two online group interviews (two researchers and three teachers in each) in which teachers discussed their scores and the strengths and weaknesses of the essays leading to the scores. Ethical guidelines of the Finnish National Board on Research Integrity (2023) were followed in the project, and teachers gave their written consent to all data gathered as did the students who gave their consent for using the essays in this project. We, the researchers, employed both content analysis of the teachers' group interview data and systemic-functional analysis (Halliday, 1973) to determine which features history teachers valued in source-based essays and how these features were realized linguistically in essays in question. Based on this analysis, we drafted the first version for the evaluation criteria (V1, see appendix 1). This phase of the process and the analysis have been reported (in Finnish) in Paldanius et al. (2021).

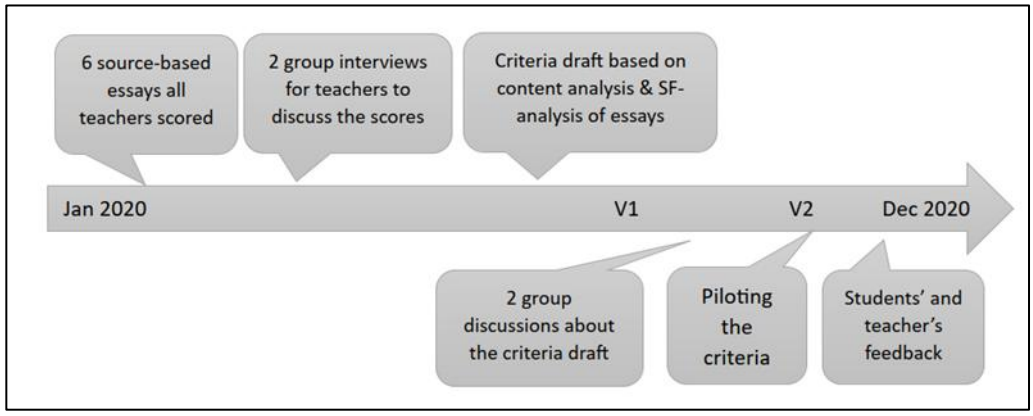


FIGURE 1

Timeline for developing the evaluation criteria

In the second phase, we organized two online group discussions in which teachers (two researchers and three teachers in each session) discussed the first draft of the criteria (V1). In these discussions, the focus was no longer on source-based essays written for a particular assignment as in phase 1, but in developing criteria that would apply to any source-based history essay. Also, during the process we moved from teachers' individual and implicit, quite holistic criteria to explicit criteria everyone could agree upon. Based on the lively discussion in both group discussions, many changes were suggested and implemented to the criteria, and a second version of the criteria was developed (V2, see appendix 1). The discussions were analyzed using discourse analysis (Gee, 2014) and reported (in Finnish) in Sulkunen et al. (2022).

Until this point in the process, the focus had been on drafting evaluation criteria that would cover key expectations of source-based essays. For teachers to be able to pilot and eventually use the criteria, also scoring rubrics was needed. After discussing different options with teachers, two alternative rubrics were created resulting in two versions of the criteria. In these versions, the evaluation criteria were practically identical with each other. The other, and the two alternative rubrics were the two-point dichotomy of "Room for improvement" and "Masters the skills", and the four-point "Fail – Satisfactory – Good – Very good". These verbal characterizations of the students' performance were considered more informative than numeric ones. In the pilot phase, teacher would pilot both versions with different groups of students, and tick one of these for each criterion. Based on these criterion-specific grades, teacher would give a score from 0 to 20 in each of the four categories (see appendix 1). The given score is not a sum or average of the criterion-specific grades but is the basis for teachers' assessment.

In the design-based research, testing a pilot model of the product is essential (Aksela, 2019). Thus, in the third phase, one of the teachers (Vappu, all names are pseudonyms) piloted the criteria with both alternative rubrics in assessing an end-of-course essay assignment. In presenting the assignment to students in class, she introduced and explained the criteria for the students and provided them for the students in an online

platform so that they would have access to them while writing the essay. After piloting, teacher was interviewed about the usability of criteria. We also gathered feedback from students about the accessibility of the criteria via an online questionnaire. Based on the feedback, the criteria were finalized. Only small edits were required at this point.

In this study, we use several types of data and methods to illustrate and examine both the co-design process and the product. The methods employed were selected to align with the objectives of the inquiry. Firstly, we provide insight into the co-design process by scrutinizing the changes made to the draft criteria by comparing the different versions of the criteria. We focus on versions V1 and V2 as the most significant changes were made between these two versions. Secondly, we use teachers' group interview data (conducted online in all phases in the spring and fall of 2020; see figure 1) and the online interview of the teacher piloting the criteria. These data were analyzed with content analysis (Mayring, 2015) to examine teachers' views on the joint development process in detail. First, expressions were identified from the interviews in which teachers expressed their views on how the collaborative effort worked for them. These expressions were reduced thematically through an iterative process, so that in the end, reduced statements were used to identify themes related to teachers' experiences of the co-design process. The analysis was carried out by the first author and verified by other co-authors. In this way, we aimed for researcher triangulation (Laine, Bamberg & Jokinen, 2007) to strengthen the reliability of the results.

Thirdly, we turn our focus on the product of the co-design using student online questionnaire data (n=32). The questionnaire included both multiple choice questions with Likert-scales and open questions with student-constructed responses. The multiple choice responses were analyzed using descriptive statistics (means, distributions) to examine how accessible the criteria were to students. The student-constructed responses were analyzed with content analysis to study the factors associated with the accessibility of the criteria. The analysis was carried out in similar manner as outlined above for teacher interviews.

Results

Co-design process: changes from V1 to V2

Several changes were made from the first draft of criteria to the second one. The number of the criteria changed a great deal, too. In the first version of the criteria, the number of individual criteria was 27. In the second version it was 23. As Table 1 shows, the criteria were divided into four categories, namely 1) *Technical answering skills* which includes criteria that relate to general aspects of writing a text that meets the requirements of an essay, 2) *Mastering historical content-knowledge* with criteria dealing with knowledge and concepts of history, 3) *Use of the source material* with criteria relating to historical literacy practices, and 4) *Reflection and argumentation* that include criteria dealing with how students present their ideas. (The order and names of the categories are the ones used in V2.) The four categories are the result of content

analysis of teachers' group interviews in the first phase of the development process. In the analysis, the criteria were identified and placed into four thematic theory-driven categories. Categories cover aims related to the historical content-knowledge (2), disciplinary literacy (3, 4; Goldman, et al., 2016), and general intermediate skills (1; Shanahan and Shanahan, 2008).

Table 1 presents the number of criteria in each category in the two versions. In version 2, the distribution of criteria into the four categories was relatively even, whereas in version 1 content-knowledge and answering skills dominated. Full versions of the criteria (V1 and V2) are included in Appendix 1 in which the criteria have been numbered, and the first digit refers to the version of the criteria (1 or 2).

TABLE 1

Distribution of criteria into four categories in versions 1 and 2.

Version 1		Version 2	
Name of the category	N of criteria	Name of the category	N of criteria
Mastering disciplinary content-knowledge	10	Technical answering skills	7
Use of the source material	6	Mastering historical content-knowledge	5
Own reflection	3	Use of the source material	6
Answering skills	8	Reflection and argumentation	5
Total	27	Total	23

The number of criteria was reduced from version 1 to version 2 since some criteria were omitted and some combined. The omitted criteria include the following three: 1) *Issues are connected to the present* (the category Reflection and argumentation), 2) *The source material is presented in the essay by providing key source information (author, date, publication platform)* and 3) *The text is fluent*, both from the category Technical answering skills. The first of these criteria was considered irrelevant to source-based essays in general albeit it was relevant to the assignment of essays that teachers scored in phase 1. As one of the teachers, Tomi, said: "Sometimes the connection to the present is a bit forced. I don't think it's that important." This idea of connection to the present was also omitted from another criteria (see Appendix 1 criterion 1.24). The second one was omitted as it was considered to overlap with another criterion related to critical examination of the source material (Appendix 1, criterion 1.15) in which similar information about the materials is taken into account.

The third criterion about text fluency was omitted as no one was able to define it in a way that would be concrete and accessible to students. As teacher called Tenho described: "[It is] what separates the nineteen points from the twenty points, from the best answer. That cannot be defined, in my opinion." Tenho's description seems to refer to a holistic approach in assessment (Arter and MicTighe, 2001), that is, to assessment that results in a general impression of a student's performance. As our aim was to

develop criteria that would communicate to students what is expected of them (Reddy and Andrade, 2010), this was not considered a useful criterion. Moreover, two criteria from version 1 were combined so that the latter sentence provides a concrete example of the former: *The language of the essay is polished: the choice of words and expressions in the essay are appropriate for a (factual) text on history*. The combined criterion was considered to better communicate to students what the polished language means and thus be more accessible (Zhao, 2022).

The changes between versions 1 and 2 included also moving several criteria from one category to another. These changes and discussion around them show how sometimes it is quite difficult to distinguish the *way* things are expressed from the content. This reflects how deeply intertwined the disciplinary literacy and knowledge-construction practices are (Moje, 2015; Goldman, et al., 2016). In group discussions, one of the teachers, Marko, referred to this by saying that "Sometimes with a badly written essay you get fewer points than you might deserve because it's hard to tell what you mean. With good writing skills you might be able to fool the teacher." Criteria that were moved from one category to another included, for example, *The sequence of presenting issues is logical and coherent* which was transferred from the content-related category into the category on answering skills (1.8 -> 2.5). Moreover, some criteria were moved from the category Mastering historical content-knowledge into the Reflection and argumentation category as they concerned primarily students' own deliberations. This was the case, for example, with criterion *Precise and plausible judgements of the significance of historical events are presented* (1.6 -> 2.20). Finally, the criterion about referring to the source material in appropriate manner was moved from Use of the source material category into Technical answering skills (1.16 -> 2.6). Suvi, one of the teachers, argued: "Referring to a work by its name and the author's surname is a kind of skill that once you have acquired, it just keeps repeating itself, so it felt more like a kind of answering skill". Suvi's reflection showed how she found it a general skill relevant for many subjects rather than being a disciplinary one, reflecting the distinction between content-area literacy and disciplinary literacy (Shanahan and Shanahan, 2012; Derlinger Kane, Ortlieb and Cheek, 2024).

Teachers' understanding of the role of general literacy skills and their relations with disciplinary ones were also raised in discussions about the order of categories and their headings. In version 1 answering skills was the last category as researchers drafting the criteria gave priority to discipline-specific categories. However, the teachers had a different idea. In the group interview, Rami suggested the following: "should we have the answering skills first - - I thought that they are kind of general things from which we then go to more specific ones". Maria agreed and explained: "I also go through the most general instructions that no matter what kind of essay it is, how you write it. So yes, I think it would be good to have them at the beginning and only then go to the history subject." Teachers views thus reflected the Shanahan and Shanahan (2008) model in which generic literacy practices are the foundation for more specialized disciplinary practices. The heading of the category was also changed by adding the word "technical" to it. Teachers found that this would clarify the general nature of the skills in question. As Tenho explained: "Isn't this considered a technical point? Then I

understand this as a category of its own but the answering skills may slightly refer to a broader skill which includes the argumentation." These discussions make visible how history teachers distinguish generic literacy practices from disciplinary ones.

Another topic abundantly discussed was the category on students' own reflection (as it was called in V1). Teachers found the idea that students would be expected to participate in knowledge construction and providing their own interpretations (as the curriculum requires) quite unimaginable as they thought it is too demanding for secondary students. As Rami pondered: "In history assignments it is quite difficult to require students' interpretations or conclusions. Some might be able to do it and I would just about be able to do it, but it feels risky to say to students that they should come up with explanations [to historical events]." Teachers' beliefs about the role of students' reflections have been reported in detail elsewhere (Sulkunen et al., 2022) showing that teachers actively look for other interpretations for students' own reflections: instead of knowledge construction and presenting own interpretations, teachers suggested it could mean compiling and considering others' interpretations or considering the significance of (others') interpretations and sources. Related to these discussions, the heading of the category was changed from Own reflection to Reflection and argumentation. Teachers felt argumentation described the category more aptly. As Suvi said: "That the argumentation is coherent, as if the construction of the argument is perhaps at the center of it [this category]?" Teachers also found the word 'argumentation' familiar to students from other school subjects. This shows how the teachers' views conflicted with the research and theory on disciplinary literacy suggesting students should be engaged in knowledge construction (Moje, 2015; Goldman, et al., 2016). In the context of the content-based tradition of teaching history in Finland (Rantala, et al., 2020), the teachers felt this requirement too demanding. As teachers' role was to be co-designers (Aksela, 2019) in this process, their views were integrated to the criteria development.

Additionally, some phrasing was modified to become more accessible for students. Usually this meant changing individual words to more familiar and transparent ones. For example, the word 'discipline' was changed into 'history' in the name of one of the categories for better accessibility (Zhao, 2022).

In all, several changes were made to the criteria from first version to the second one. Some of these stemmed from transforming the criteria drafted based on one assignment to criteria that would be useful for assessing any source-based history essay. Some reflected the difficulties in distinguishing mastering content-knowledge, use of sources and ways knowledge is presented, and some concerned fundamental reflections on the role of procedural knowledge in secondary history teaching.

Teachers' experiences of the development process and the criteria

In the content analysis of the two group discussions, teachers expressed their views on both the co-design process and the professional development resulting in changes in their assessment practices. Teachers' views about the co-design process were clustered into the following theme categories: 1) consistency of assessment, 2) power in/of assessment, 3) sharing assessment practices and 4) changes in assessment practices. The

first two categories were mostly related to teachers' concerns over challenges in assessment. The latter two made visible the benefits of the co-design project for teachers' professional development.

The consistency of assessment was a concern for teachers already at the beginning of the co-design project (phase 1). Teachers noticed that they did not always arrive at the same scores and noted differences between their individual scoring scales in phase 1. This resulted in teachers becoming aware that each of them had their own implicit evaluation criteria and scoring rubrics which were influenced by the Matriculation Examination and general course practices. Tomi explained this as follows (group discussion, phase 1): "I've used the Matriculation Examination scoring rubrics in these twenty-point tasks as a help. I used it here. So, from fourteen to twenty points is commendable. Then seven to thirteen is good through satisfactory. Below seven is then weak answers." Tenho, for example, described his typical assessment as follows (phase 1): "I have indeed evaluated them a little bit through course evaluation. So, you have to get the thirty percent to pass." Also, Suvi confirms in phase 1 that each teacher may have their own assessment rubrics: "[Each has] their own line when the courses are evaluated." Moreover, the consistency of assessment was a concern when teachers assess bundles of students' assignments: "how much is missed in the authentic assessment situations when there are ninety or so texts" (Vappu, phase 1). Consistency issues raise concerns about fairness and equity in assessment.

The theme of power in assessment is a minor issue in the group discussion data but an important one. This is illustrated by the following quote from Tenho: "it is a bit frightening how much power we have in it, and no one really ever checks our assessments" (phase 1). It is noteworthy that while teachers became aware of many issues related to assessment, power aspect was discussed only little.

Group discussions offered teachers opportunities to share and discuss their assessment practices during the co-design. In group discussions teachers pointed out that they are not in the habit of discussing evaluations among themselves, and these discussions proved to be a valued learning experience. Rami, for example, said already in phase 1 that "we very rarely discuss how we have evaluated - - you do learn from this." Also, Suvi confirmed in phase 1: "On this basis [having different criteria] [we] could perhaps more sometimes discuss the fact that what kind of things people evaluate and what kind of points they give." The discussions during co-design project offered teachers reflecting experiences that are needed for professional development (Coburn and Penuel, 2016; Mansvelder-Longayroux, Beijaard and Verloop, 2007). According to the teachers, the discussions and reflections raised teachers' awareness of challenges in assessment mentioned above. Yet, it is beyond the scope of this study to determine how lasting the impact was. The co-design experience made Suvi and her colleagues to realize that joint discussions on criteria and scoring rubrics would be useful. This and other above-described understandings of the challenges in assessment opened new avenues for teachers for professional development.

The opportunities of professional development were realized in teachers' assessment practices. This was particularly the case with Vappu who piloted the criteria with students. After piloting the criteria (phase 3), Vappu highlighted the usefulness of the

piloting experience. She described how "the assessment was different from before, the criteria made me think more about concrete things, before I thought more about the essay as a whole". The pilot therefore provided Vappu with an opportunity to shift the focus of assessment from holistic to criterion-based assessment (Arter and McTighe, 2001). However, she did have room for holistic deliberation in terms of the overall grade of the assignment: “When I thought about the scores for the different areas [categories], I thought about the whole and when I decided how do I convert the scores into grades -- then I thought about the whole.” She also said that she emphasized the category Use of the source material since “interpretation of sources was emphasized in the course”. From the two alternative scoring rubrics, Vappu preferred the more detailed one as “it was more informative for the students” and “easier to use”. Thus, Vappu’s estimation was that the more detailed scoring rubrics would communicate better the strengths and development points of students’ essays, possibly allowing the teacher to justify his/her overall assessment to the student in more detail (Arter and McTighe, 2001).

Next, we turn to students’ experiences on the criteria during the pilot in which Vappu evaluated their essays using the criteria.

Students’ views on accessibility of the evaluation criteria

To fulfill the function of communicating to students what is expected of their performance, criteria need to be accessible to students, i.e. students need to comprehend their meaning in relation to the assignment (Zhao, 2022), in this case writing a source-based history essay. To investigate students’ views on the accessibility of the evaluation criteria, we gathered student feedback using an online questionnaire with both multiple choice and constructed-response questions. Students (n = 32) responded to the questionnaire after the pilot in which their history teacher assessed students’ source-based essays using the criteria. In the questionnaire, each of the criteria from version 2 (see appendix 1) was presented to students with the question “how well does each of the evaluation criteria below help you understand what is expected of you in the essay?”. Students responded to the question using a scale from 1 to 4 (1 = not at all or poorly, 2 = somewhat, 3 = well, 4 = very well).

Overall, students found the criteria quite accessible as with most of them the mean was 3 or higher. The mean was below 3 only with four criteria (see Table 2), namely the criteria about linking examples to a wider issue, about presenting appropriate knowledge on history, about presenting own thinking avoiding too many absolutes, and black-and-white views as well as connecting source materials to historical knowledge. These criteria also had the lowest percentage of students answering that the criteria help them understand what is expected “well” or “very well”. Moreover, there were three criteria with the mean response being 3,5 or higher (Table 2): criteria about essay matching the assignment, referring appropriately to the source material and the requirement to use the source materials.

TABLE 2.

The criteria with the lowest and highest means for accessibility (scale 1–4).

Criterion	Mean	% of S's answering "well" or "very well"
2.16 The (data) examples are linked to a wider phenomenon/issue.	2,4	41
2.8 Appropriate knowledge on history is presented in the essay.	2,7	62
2.23 One's own thinking is presented avoiding too many absolutes, black-and-white views and generalizations.	2,8	65
2.14 The source materials are connected to historical knowledge.	2,8	63
. . .		
2.6 The references to the source material are appropriate (e.g. by the name of the work, by the author's surname).	3,5	97
2.13 Source materials are used.	3,5	91
2.1 Essay matches the assignment: all the questions in the assignment are addressed and answered.	3,6	94

It is worth noting that while there is relatively little variation within or between categories of criteria, none of the criteria with the lowest mean was from the category Technical answering skills, and two of the criteria with highest mean came from this category (2.1, 2.6). The accessibility of the criteria related to technical answering skills was supported by the percentage distribution of students' responses. In this category, the percentage of students answering that the criteria help them understand what is expected "well" or "very well" varied from 81 % to 97 %. In other categories, there was more variation: the percentages varied between 62 and 87 with criteria related to mastering historical content-knowledge, between 41 and 91 with criteria on using the source material and between 65 and 91 on reflection and argumentation. As Technical answering skills includes generic content-area literacy practices introduced in many school subjects and their assignments (Shanahan and Shanahan, 2012; Derlinger Kane, Ortlieb and Cheek, 2024), it is likely that they are quite familiar to students.

Indeed, the accessibility of the criteria seems to be related to their familiarity to students. In the question about familiarity, 88 % of the students answered that the criteria were familiar or very familiar. The criteria familiar from other school subjects were found to be most accessible (2.6, 2.1). The connection between familiarity and accessibility was confirmed by the content analysis of the student constructed responses: many students stated that "almost all" or "all criteria were very familiar since we hear about them every time we write an essay", and many mentioned particularly criteria related to technical answering skills saying for example "these are the basics for an essay" or "these are the criteria in all subjects". The category of Technical answering

skills clearly represented content-area literacy practices (Derlinger Kane, Ortlieb and Cheek, 2024) in the developed criteria.

Likewise, the criteria found least accessible were mentioned as being unfamiliar. A student commented linking data examples to wider issues (2.16) saying "I do not remember seeing this before", and another one had a similar comment about presenting own thinking without absolutes (2.23): "I had not heard about this before". However, these least accessible criteria were also difficult to understand, and students would have needed more explanation on their content. Students struggled with the concept of historical knowledge and its appropriateness (criteria 2.8 & 2.14): "I did not understand what appropriate knowledge on history means" and "what is history knowledge, how is it defined and what is appropriate knowledge on history?" Moreover, understanding what some of the criteria mean in practice was challenging: "I do not know how to link example to wider issues in a history essay" (2.16) and about presenting own thinking avoiding too many absolutes (2.23), "I don't understand what it means and how to do this in an essay". These results suggest a need to teach the disciplinary literacy practices explicitly (Sulkunen and Saario, 2019).

These results support the earlier findings of studies in which students needed more concrete and contextualized information about the meaning of the criteria to communicate what is expected in the assignment (Zhao, 2022). Also, it is to be noted that with the absence of quality definitions, the verbalization of the criteria was even more scarce than it would be with the definitions (Bearman and Ajjawi, 2018).

Discussion and conclusion

In this article, we have examined the co-design process of evaluation criteria for source-based history essays implemented with a group of history teachers in a mid-sized Finnish upper-secondary school. We studied both the process and the product of the co-design: we examined how this collaborative effort worked from the perspective of teachers and how the process transformed their assessment practices and how accessible the criteria developed were to students. The co-design research aimed at supporting skill-based history teaching, particularly teaching of historical literacy practices, by creating criteria for evaluating a typical history assignment, source-based essay. Assessment criteria provide teachers concrete tools for teaching and assessment, and at the same time make expectations of the assignment clear to students (Reddy and Andrade, 2010).

In the study, we have shown that the co-design approach and group discussions were a valuable opportunity for the teachers to share their assessment practices and reflect on them. This supports the findings in the earlier studies which show that teachers' and researchers' collaboration supports teachers' professional development (Coburn and Penuel, 2016; Mansvelder-Longayroux, Beijaard and Verloop, 2007). More specifically, the mere realization of varying assessment practices and discussion about them during the partnership made challenges and important aspects of assessment visible. Moreover, the criteria that were developed in partnership gave teachers

opportunities to discuss also the relationship between content-knowledge, inquiry practices as well as general and discipline-specific literacy practices. This became evident in discussions about criteria and their categories in phase 2. The pilot of the criteria showed how criteria enable the transfer from holistic to criterion-based assessment. While holistic assessment can be less time-consuming for teachers, it may not provide detailed feedback for students which is the benefit of the criterion-based assessment (Jönsson, et al., 2021). While the co-design process has been in many ways valuable for the teachers participating in it, for developing the criteria the number of teachers and the scale of the pilot study is quite small. With larger sample of teachers, some criteria not yet included might emerge even if the ones developed in the study were validated by theory and previous research. Nevertheless, future studies could include more extensive piloting and further development of the criteria.

The accessibility of the evaluation criteria for students is important for transparency of the assessment and for communicating the expectations of the assignment (Zhao, 2022; Reddy and Andrade, 2010). Our findings showed that criteria were on average quite accessible for students, but this was associated with the familiarity of the criteria. The technical answering skills that deal with following the instructions and general literacy skills of writing an essay regardless of the discipline, that is, content-area literacy practices (Derlinger Kane, Ortlieb and Cheek, 2024), were among the most accessible ones being familiar from several subjects over school years. However, some of the discipline-specific criteria, even ones dealing with traditional category of content-knowledge, caused confusion and difficulties in interpreting them. Explaining the new criteria once in the context of explaining the assignment clearly was not enough but they need to be explained to students more thoroughly. More extensive, perhaps repeated, discussion about the criteria between teacher and students as well as concrete examples to illustrate the criteria in an essay might be needed (Zhao, 2022). Furthermore, there certainly is room for development in the wordings of some of the criteria. In all, piloting the criteria and results of the student questionnaire are extremely valuable in a development process like this.

While we succeeded in co-designing criteria for evaluating source-based essay that cover essential aspects of an assignment requiring historical inquiry – that is content-related (Breakstone, 2014), epistemological (Pitt and Quinlan, 2021) and disciplinary language and literacy related aspects (Pessoa, Mitchell and Miller, 2017) – this was only a first step in a demanding process. The process and the criteria have their limitations, too. One of them is that the assessment rubrics include only evaluation criteria and scoring rubrics. With quality definitions the rubrics would communicate the expectations to students even more clearly and would perhaps be more useful and clearer to teachers, too. After all, also teachers who were not involved in our co-design research may use them. Moreover, in the pilot, the criteria were introduced and discussed with students only once and quite briefly. And, most importantly, we have not so far examined long-term use and effects of the criteria. The prior studies show that using analytic and task-specific criteria improve the reliability of the assessment (Jonsson and Svingby, 2007) and task-specific criteria are more likely to improve

students' performance than general ones (Panadero and Jonsson, 2013) but studying the use and effects of our newly developed criteria is left for future studies.

While criterion-based assessment is becoming more common in Finnish schools (Atjonen, et al., 2019), achieving absolute transparency and unambiguity in assessment is challenging and the belief that criterion-based assessment would by itself lead to the elimination of subjectivity in evaluation, is unrealistic. At best, however, the use of criteria will systematize the assessment and allow clearer justification of grades. Criteria also allow the teacher to give concrete feedback to students, who can also use them as a tool for self- and peer-assessment as well as guidance during the assignments.

References

Aksela, M. (2019). Towards student-centred solutions and pedagogical innovations in science education through co-design approach within design-based research. *LUMAT: International Journal on Math, Science and Technology Education*, 7(3), 113–139. https://doi.org/10.31129/LUMAT.7.3.421

Arter, J. & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. London: Sage.

Atjonen, P., Laivamaa, H., Levonen, A., Orell, S., Saari, M., Sulonen, K. & Immonen, J. (2019). *"Että tietää missä on menossa" Oppimisen ja osaamisen arviointi perusopetuksessa ja lukiokoulutuksessa.* [Assessment of learning and performance in basic education and upper-secondary education.] Helsinki: Finnish Education Evaluation Centre.

Bearman, M. & Ajjawi, R. (2018). "From "Seeing Through" to "Seeing With": Assessment Criteria and the Myths of Transparency." *Frontiers in Education*, 3(96). https://doi.org/10.3389/feduc.2018.00096

Breakstone, J. (2014). Try, Try, Try Again: The Process of Designing New History Assessments. *Theory & Research in Social Education*, 42(4), pp. 453–485. https://doi.org/10.1080/00933104.2014.965860

Brookhart, S. M. & Chen, F., (2015). The quality and effectiveness of descriptive rubrics. *Educational Reviews*, 67(3), pp. 343–368.

Brozo, W., Moorman, G., Meyer, C. & Stewart, T. (2013). Content Area Reading and Disciplinary Literacy. A case for the radical center. *Journal of Adolescent & Adult Literacy*, 56 (5), pp. 353–357. https://doi.org/10.1002/JA AL.153

Coburn, C. E. & Penuel, W. R. (2016). Research–practice partnerships: Outcomes, dynamics, and open questions. *Educational Researcher*, 45(1), pp. 48–54.

Delinger Kane, B., Ortlieb, E. & Cheek, E. H. (2024). Tracing the History of Disciplinary Literacies. In E. Ortlieb, E., B. Delinger Kane & E. H. Cheek (Eds.), *Disciplinary Literacies: Unpacking Research, Theory, and Practice*. Guilford Publications, pp. 1–15.

van Drie, J. & van Boxtel, C. (2008). Historical reasoning: Towards a Framework for analysing students' reasoning about the past. *Educational Psychological Review*, 20, pp. 87–110.

Edelson, D. C. (2002). Design research: What we learn when we engage in design. *The Journal of the Learning Sciences*, 11, 105–121

The Finnish National Agency for Education (2014). *National core curriculum for basic education 2014*. Helsinki: Finnish National Agency for Education.

The Finnish National Agency for Education (2019). *Lukion opetussuunnitelman perusteet 2019*. [National curriculum for upper secondary school 2019.] Helsinki: Finnish National Agency for Education.

The Finnish National Agency for Education (2026). *General Upper Secondary Education*. Helsinki: Finnish National Agency for Education. Available at: <https://www.oph.fi/en/education-system/general-upper-secondary-education> [accessed February 16, 2026].

The Finnish National Board on Research Integrity (2023). *The Finnish code of conduct for research integrity and procedures for handling alleged violations of research integrity in Finland*. Helsinki: Finnish National Board on Research Integrity 4/2023.

Gee, J. P. (2014). *How to do discourse analysis*. Abingdon: Routledge.

Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M., Greenleaf, C., Lee, C. D., Shanahan, C. & Project READI (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist*, 51(2), pp. 219–246.

Halliday, M. A. K. (1973). *Explorations in the functions of language*. London: Arnold

Hardy, A. & Iwatani, E. (2021). *Rubrics for examining historical thinking skills in high school world history activities and student work: Construct validity evidence from the literature*. Washington, D.C.: Digital Promise.

Henrick, E. C., Cobb, P., Penuel, W. R., Jackson, K. & Clark, T. (2017). *Assessing research-practice partnerships: Five dimensions of effectiveness*. William T. Grant Foundation. Available at: <https://wtgrantfoundation.org/new-report-assessing-research-practice-partnerships-five-dimensions-effectiveness> [Accessed June 24 2024].

Humphrey, S., Martin, J. R., Dreyfus, S. & Mahboob, A. (2010). The 3x3: Setting up a linguistic toolbox for teaching and assessing academic writing. In A. Mahboob & N. Knight (eds.) *Applicable linguistics: Texts, contexts and meanings*. London: Continuum, pp. 185–199.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), pp. 130–144.

Juuti, K., Lavonen, J., Salonen, V., Salmela-Aro, K., Schneider, B. & Krajcik, J. (2021). A Teacher–Researcher Partnership for Professional Learning: Co-Designing Project-Based Learning Units to Increase Student Engagement in Science Classes. *Journal of Science Teacher Education*, 32(6), pp. 625–641.

Jönsson, A., Balan, A. & Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers´ grading. *Assessment in Education: Principles, Policy & Practice*, 28(3), pp. 212-227.

Laine, M., Bamberg, J. & Jokinen, P. (2007). Tapaustutkimuksen käytäntö ja teoria. In M. Laine, J. Bamberg, & P. Jokinen (eds.) *Tapaustutkimuksen taito* [Practice of case studies]. Helsinki: Gaudeamus, pp. 9–38.

Lok, B., McNaught, C. & Young, K. (2016). Criterion-referenced and norm-referenced assessments: compatibility and complementarity. *Assessment & Evaluation in Higher Education*, 41(3), pp. 450–465.

Löfström, J. & Kaarninen, M. (2013). Historian ja yhteiskuntaopin kokeet ylioppilastutkinnossa 2006–2012: havaintoja ja tutkimustehtäviä. In L. Tainio, K. Juuti & S. Routarinne, S. (eds.) *Ainedidaktinen tutkimus koulutuspoliittisen päätöksenteon perustana* [Subject didiactic research as a basis for educational policy]. Helsinki: The Finnish Association for Subject Didactic Research, pp. 177–198.

Mansvelder-Longayroux, D. D., Beijaard, D. & Verloop, N. (2007). The portfolio as a tool for stimulating reflection by student teachers. *Teaching and Teacher Education*, 23(1), pp. 47–62. https://doi.org/10.1016/j.tate.2006.04.033

Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. In A. Bikner-Ahsbahs, C. Knipping & N. Presmeg (eds.) *Approaches to qualitative research in mathematics education: Examples of methodology and methods*. London: Springer, pp. 365–380.

Moje, E. (2015). Doing and teaching disciplinary literacy with adolescent learners: A social and cultural enterprise. *Harvard Educational Review*, 85 (2), 254–279. https://doi.org/10.17763/0017-8055.85.2.254

Monte-Sano, C. & De La Paz, S. (2012). Using writing tasks to elicit adolescents' historical reasoning. *Journal of Literacy Research*, 44(3), pp. 273–299.

Nokes, J. D. (2013). Building Students' Historical Literacies. Learning to Read and Reason with Historical Texts and Evidence. New York: Routledge.

O'Donovan, B., Price, M. & Rust, C. (2004). "Know What I Mean? Enhancing Student Understanding of Assessment Standards and Criteria." *Teaching in Higher Education*, 9(3), pp. 325–335. https://doi.org/10.1080/1356251042000216642

Paldanius, H., Sulkunen, S., Luukka, M.-R. & Saario, J. (2021) Lukion historian opettajien käsityksiä esseen arviointiin vaikuttavista piirteistä [Upper-secondary school history teachers' beliefs on features affecting assessment of essays]. *Ainedidaktiikka*, 5(1), pp. 3–22. https://doi.org/10.23988/ad.98334

Panadero, E. & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, pp. 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Pessoa, S., Mitchell, T. D. & Miller, R. T. (2017). Emergent arguments: A functional approach to analyzing student challenges with the argument genre. *Journal of Second Language Writing*, 38, pp. 42– 55.

Pitt, E. & Quinlan, K. M. (2021). Signature assessment and feedback practices in the disciplines. *Assessment in Education: Principles, Policy & Practice*, 28(2), pp. 97–100.

Puustinen, M., Paldanius, H. and Luukka, M.-R. (2020). Sisältötiedon toistamista vai aineiston analyysia? Tiedonalakohtaiset tekstitaidot historian ylioppilaskokeen tehtävänannoissa ja pisteytysohjeissa. [Repeating content or analysing sources? Disciplinary literacy in matriculation exam's history tasks and scoring guide.] *Kasvatus & Aika*, 14(2), pp. 9–34. https://doi.org/10.33350/ka.84579

Radinsky, J., Goldman, S. R. & Pellegrino, J. 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 and P. Seixas (eds.) *New Direction in Assessing Historical Thinking*. New York: Routledge, pp. 132–142.

Rantala, J., Puustinen, M., Khawaja, A., Van den Berg, M. & Ouakrim-Soivio, N. (2020). *Näinkö historiaa opitaan?* [Is this the way to learn history?]. Helsinki: Gaudeamus.

Reddy, Y. M. & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35(4), pp. 435–448. https://doi.org/10.1080/02602930902862859

Rust, C., Price, M. & O'Donovan, B. (2003). "Improving Students' Learning by Developing their Understanding of Assessment Criteria and Processes." *Assessment & Evaluation in Higher Education*, 28(2), pp. 147–164. https://doi.org/10.1080/02602930301671

Shanahan T. & Shanahan C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content area literacy. *Harvard Education Review*, 78, 40–59.

Shanahan, T., & Shanahan, C. (2012). What is disciplinary literacy and why does it matter? *Topics in Language Disorders*, 32(1), 7–18. https://doi.org/10.1097/TLD.0b013e318244557a

Sulkunen, S., Luukka, M.-R., Paldanius, H. & Saario, J. (2022). Historian opettajien käsityksiä tiedonalan tekstitaitojen opettamisesta lukiossa. [History teachers' beliefs in teaching disciplinary literacu at upper-secondary school.] *Kasvatus & aika*, 16(4), pp. 43–61. https://doi.org/10.33350/ka.112693

Sulkunen, S., Luukka, M.-R., Saario, J. & Veistämö, T. (2019). Monilukutaito lukion historian opetuksessa. [Multiliteracies in upper secondary school history teaching.] *Ainedidaktiikka*, 3(2), pp. 2–23. https://doi.org/10.23988/ad.76111

Sulkunen, S. & Saario, J. (2019) Tiedonmuodostus ja tekstityö: tapaustutkimus lukion historian opetuksen tekstikäytänteistä. [Knowledge construction and literacy practices in teaching history.] *Finnish Educational Journal Kasvatus*, 50(2), pp. 149–163.

Veijola, A. (2016). Historiatietoisuus, historiallinen ajattelu ja historian tekstitaidot: Uuden opetussuunnitelman moninaiset lähtökohdat. [Historiacl consciousness, historical thinking and historical literacy: many premises of the new curriculum.] *Kasvatus & Aika*, 10(2), pp. 6–18.

Virta, A. (2014). Historia ja yhteiskuntaoppi reaalikokeessa 1921–1969 – koetehtävien kehitys oppiennätysten ja opetusta koskevan keskustelun näkökulmasta. [History and social sciences in matriculation examination 1921–1969 – the evolution of tests in terms of learning records and the debate on teaching.] *Ennen ja Nyt*, 4. Available at: <https://journal.fi/ennenjanyt/article/ view/108600/63603> [Accessed 24 June2024].

Wineburg, S. (1991). On the reading of historical texts: Notes on the breach between school and academy. *American Educational Research Journal*, 28(3), pp. 495–519. https://doi.org/10.3102/00028312028003495

Ylioppilastutkintolautakunta (2023). *Hyvän vastauksen piirteet: FI – historia. Syksy 2023*. [Features of good response: FI – history. Fall 2023.] Available at: <https://tiedostot.ylioppilastutkinto.fi/kokeet/2023-09-13_HI_fi/grading-instructions.html> [Accessed 2 June 2024].

Zhao, H. (2022). Promoting accessibility of assessment criteria: shifting from a product- to a process- and future-oriented approach. *Teaching in Higher Education*, 29(5), pp. 1283–1301. https://doi.org/10.1080/13562517.2022.2129964

APPENDIX 1

Assessment criteria versions 1 and 2

Version 1 was created based on group-interviews in spring 2020, and discussed by teachers and researchers in two groups on 14. and 16.9.2020. Version 2 was created based on these discussions.

Version 1 (V1)	Version 2 (V2)
Mastering disciplinary content-knowledge: 1.1 Relevant knowledge on history is presented in the essay. 1.2 Relevant disciplinary concepts are used and they are defined. 1.3 Events are timed accurately enough. 1.4 The topic is addressed broadly, deeply and/or analytically enough. 1.5 The views presented are based on historical knowledge. 1.6 Precise and plausible judgements of the significance of historical events are presented. 1.7 Different perspectives on the issue at hand are presented. 1.8 The sequence of presenting issues is logical and coherent. 1.9 The issues are linked with each other and form a coherent whole. 1.10 The choice of words and expressions in the essay are appropriate for a (factual) text on history.	**Technical answering skills:** 2.1 Essay matches the assignment: all the questions in the assignment are addressed and answered. 2.2 The essay has a clear structure and paragraph structure: a beginning, a reading section and an ending. 2.3 The essay has a good start: it introduces the key concepts and issues relevant to the assignment. 2.4 The essay has an effective ending: it presents a conclusion. 2.5 The sequence of presenting issues is logical and coherent. 2.6 The references to the source material are appropriate (e.g. by the name of the work, by the author's surname). 2.7 The language of the essay is polished: the choice of words and expressions in the essay are appropriate for a (factual) text on history.
Use of the source material: 1.11 Source materials are used. 1.12 The source materials are connected to historical knowledge. 1.13 Examples of data supporting the answer are given. 1.14 The (data) examples presented are linked to a wider phenomenon. 1.15 The source material is examined critically (author, date of origin, 'type' of material) and this is taken into account when interpreting the material. 1.16 The references to the source material are appropriate (e.g. by the name of the work, by the author's surname).	**Mastering historical content-knowledge:** 2.8 Appropriate knowledge on history is presented in the essay. 2.9 Appropriate disciplinary concepts are used and they are defined. 2.10 The essay outlines temporal cause-and-effect relationships. 2.11 The topic is addressed broadly, deeply and/or analytically enough. 2.12 The issues are linked with each other and form a coherent whole.
Own reflection: 1.17 Issues are connected to the present. 1.18 Own views or conclusions on the issue under discussion are presented.	**Use of the source material:** 2.13 Source materials are used. 2.14 The source materials are connected to historical knowledge. 2.15 Examples of data supporting the answer are given (quotes or expressed in own words).

1.19 One's own thinking is presented avoiding too many absolutes, black-and-white views and generalisations.	2.16 The (data) examples are linked to a wider phenomenon. 2.17 The source material is examined critically (author, date of origin, 'type' of material) and this is taken into account when interpreting the material. 2.18 One's own ideas and those of others are distinguished clearly referring to the material (e.g. "The source material says").
Answering skills: 1.20 All the questions in the assignment are addressed and answered. 1.21 The source material is presented in the essay by providing key source information (author, date, publication platform). 1.22 One's own ideas and those of others are distinguished clearly referring to the material (e.g. "The source material says"). 1.23 The essay has a good start: it introduces the key concepts, issues and source material relevant to the assignment. 1.24 The essay has an effective ending: it connects the issues discussed to the present day or presents a conclusion. 1.25 The text is fluent. 1.26 The essay has a clear structure and paragraph structure: a beginning, a reading section and an ending. 1.27 The language of the essay is polished.	**Reflection and argumentation:** 2.19 Different perspectives on the issue at hand are presented. 2.20 Precise and plausible judgements of the significance of historical events are presented. 2.21 The views presented are based on historical knowledge. 2.22 Own conclusions on the issue under discussion are presented. 2.23 One's own thinking is presented avoiding too many absolutes, black-and-white views and generalisations.