History Education Research Journal

Kühberger, C., Bramann, C., Weiß, Z. and Meurers, D. (2019) 'Task complexity in history textbooks: A multidisciplinary case study on triangulation in history education research'. *History Education Research Journal*, 16 (1): 139–57. DOI https://doi.org/10.18546/HERJ.16.1.12

Task complexity in history textbooks: A multidisciplinary case study on triangulation in history education research

Christoph Kühberger* and Christoph Bramann – *University of Salzburg, Austria*
Zarah Weiß and Detmar Meurers – *University of Tübingen, Germany*

Abstract

The purpose of history education in Austria has changed over at least the last decade. While the focus used to be to give students a master narrative of the national past based on positivist knowledge, the current objective of history education is to foster historical thinking processes that enable students to form transferable skills in the self-reflected handling and creation of history. A key factor in fostering historical thinking is the appropriation of learning tasks. This case study measures the complexity of learning tasks in Austrian history textbooks as one important aspect of their quality. It makes use of three different approaches to complexity to triangulate the notion: general task complexity (GTC), general linguistic complexity (GLC), and domain-specific task complexity (DTC). The question is which findings can be offered by the specific strengths and limitations of the different methodological approaches to give new insights into the study of task complexity in the domain of history education research. By pursuing multidisciplinary approaches in a triangulating way, the case study opens up new prospects for this field. Besides offering new insights on measuring the complexity of learning tasks, the study illustrates the need for further research in this field – not only related to the development of analytical frameworks, but also regarding the notion of complexity in the context of historical learning itself.

Keywords: computational linguistics; historical thinking; task complexity; textbook research; triangulation

Context of research: Historical thinking and learning tasks

> Recognition in the public culture of the powerful role of collective memory, awareness of rapid demographic changes, and the ubiquity of conflicts over recognition, reparation, and commemoration of historical injustice: all of these pose new opportunities and new demands on history education in schools. (Ercikan and Seixas, 2015: xi)

In the context of the opportunities and demands identified by Ercikan and Seixas (ibid.), a paradigm shift has taken place in Austrian school curricula since 2008, from focusing on historical content to fostering transferable domain-specific thinking skills (Körber *et al.*, 2007).

A new discourse about different types of tasks started in the domain of history education (Heuer, 2011: 447; Köster *et al.*, 2016). As well as tasks to evaluate and assess

historical thinking in tests (Kühberger, 2014; Ercikan and Seixas, 2015), learning tasks became a central key for the development of learning opportunities (Waldis *et al.*, 2012). In the paradigm of focusing on processes of historical thinking, learning tasks are therefore 'crucial variables ... and take into account the individual skill level of the learners. They should be differentiated, … appropriately challenging, and thus sufficiently complex, meaningful, authentic, demanding and adapted to the learning group' (Leisen, 2010: 62).

In the context of historical learning, textbook tasks are significant, because in German-speaking countries history textbooks are still considered to be a core medium of history lessons. Modern history textbooks are intertextual and multimodal representations, designed for working and learning with historical material (historical sources and narratives about the past). They create a multitude of interpretations and constructions about the past (including guided narrations in the author's text). Tasks, often in the form of questions, are essential components of textbooks that serve as a learning medium. They also may serve as key elements to evaluate the implementation of domain-specific learning modes in history textbooks and to identify how history textbooks initialize and foster historical thinking processes (Bramann, 2018: 190). An essential aspect in the construction of suitable learning tasks is the inherent complexity of the given structure.

However, there is hardly any empirical research on task complexity in the field of history education. Research on historical learning tasks so far focuses on normative moments (for example, Heuer, 2011; Thünemann, 2013) or – in the field of textbook research – on specific details such as cognitive performance levels (for example, Bernhard, 2016; Bramann, 2018). In addition, the concrete meaning of the term 'complexity' has not been defined so far. However, complexity is used in a linguistic context – even in history education research on suitable learning tasks (Heuer, 2011: 449). Therefore, a domain-specific definition of complexity is still pending.

Because of the limited connecting factors in the scientific discourse on the 'historical complexity' of learning tasks in general (Von Borries *et al.*, 2005: 78), and on new triangulated approaches to complexity, this case study opens up new prospects for the field of history education research.

Methodological approaches

Triangulation as an approach to task complexity

Since complexity in history education research is very rarely empirically researched, triangulation attempts have been used to capture the common subject (learning tasks from history textbooks), using various methodological processes. Since complexity is a multidimensional concept, our framework is based on three different concepts of complexity that contrast different theoretical approaches: general task complexity (GTC), general linguistic complexity (GLC), and domain-specific task complexity (DTC) (see Figure 1). The main questions are how the different approaches to the complexity of tasks are useful for the analytical examination of learning tasks in the domain of history education research, and whether the multidisciplinary research approach provides findings for future task research that a single, domain-specific approach would not discover.

This 'purpose of method integration … can serve … for the production of a more coherent and complete picture of the investigated domain than mono-method research can yield' (Kelle, 2006: 293). In this context, the chosen research design can

be termed an approach of triangulation to determine the potential and limitations of the single method (Flick, 2003: 315). Hence, the question is whether a domain-specific analysis of learning tasks shows more differentiated results of the same data than other approaches that focus on very general aspects of task complexity (Gürtler and Huber, 2012: 42–3). In this way, the discussion about triangulation moves towards a concentrated view on the phenomenon of learning tasks in history textbooks. For this, 68 tasks from an Austrian history textbook for the eighth grade (age 13) on National Socialism and the Holocaust (Bachlechner *et al.*, 2012: 26–50) were consensually coded and analysed by two experts in history education (GTC/DTC) in an *investigator triangulation*. The same tasks were then analysed by two computational linguists (GLC).

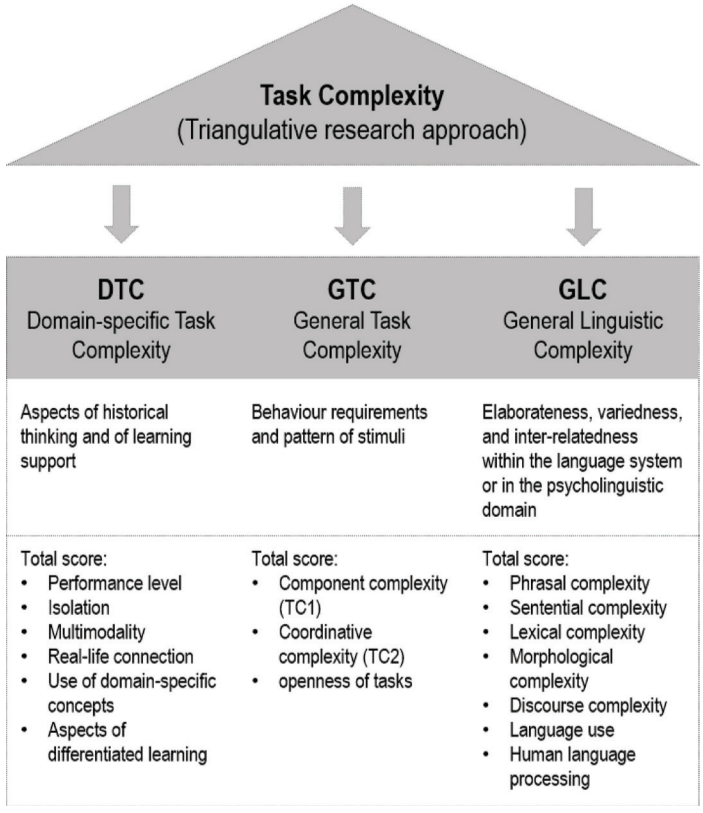


Figure 1: Multidisciplinary approach to task complexity

Complexities of learning tasks

According to the well-established German dictionary *Duden*, *Komplexität* ('complexity') is used in everyday language to denote a multilevel situation. In a more scientific view, it refers to ontological questions. In this regard, complexity can be defined as an entity in which many interdependent features exist in a snippet of reality:

> The complexity of snippets of reality increases with more available features and their increasing dependency on each other. Hence, the degree of complexity arises from the extent to which different pieces of reality and their connections are considered, in order to capture a situation in a respective snippet of reality and to plan actions. (Dörner, 1989: 60–1 [our translation])

Following this definition, the complexity of a task depends on: (1) the number of its features (variables); (2) the dependencies and necessary links between its characteristics (connectedness); and (3) the single steps (acts) required to solve the learning task. Connectedness could occur in domain-specific learning tasks, where students are required to use different materials and critical interpretations in order to solve the task.

While this approach shows scepticism about an empirical acquisition of complexity (ibid.: 61–2) – at least because complexities today are researched in more multidisciplinary ways, to predict economic or ecological developments (for example, biodiversity in the rainforest or economic markets) – empirical educational research includes attempts to measure complexity through an empirical investigation of variables and links (Robinson, 2001). Objective complexity can only be approximated in this way, since its application ultimately depends on experience and practice (Dörner, 1989: 61–2; Elen and Clark, 2006). Relating to tasks in history textbooks, it has been pointed out that tasks are more complex when various interdependent aspects have to be taken into account in order to solve them (Kühberger, 2014: 24).

For the linguistic analysis of the tasks, we employ the operationalization of linguistic complexity from second-language acquisition (SLA) research. It is part of the triad of complexity, accuracy, and fluency (CAF) that characterizes language performance (Housen *et al.*, 2012). Linguistic complexity captures the elaborateness, variedness, and inter-relatedness of the language used, regarding its lexicon and morphology, grammatical structures, meaning, and function at the sentence level, and the distribution of cohesive language devices. Complementing this SLA view of complexity with a psycholinguistic perspective, we can add notions of complexity related to language use and human sentence processing costs.

The different approaches to complexity (see Figure 1) will be presented in the form of mathematical terms. These not only allow for a direct comparison of results, but also for the formation of a hermeneutical cluster, based on various degrees of complexity.

Analysis and results

Domain-specific task complexity (DTC)

To examine the domain-specific complexity of learning tasks in history textbooks and to compare it with other types of complexity calculations, total sum scores will be calculated. Various individual analyses, which are applied to learning tasks, provide different indicators and add up to one score per task. The individual total scores provide information on the respective complexity of the tasks according to the underlying theoretical constructs. The selected variables are designed based on theoretical discussions about task complexity in history education research. They consider general aspects as part of the current discourse on tasks in educational science, as well as domain-specific aspects that address insights on historical thinking.

With respect to the general and the domain-specific discourse on learning tasks, we rely on a tripartite operationalization of the *performance level* required for solving a task (Kühberger, 2011: 6–7), which is loosely based on Bloom's taxonomy (Bloom *et al.*, 1956) and its revised version by Loron W. Anderson and colleagues (Anderson *et al.*, 2001), which was criticized for showing too little sharpness for a valid coding (Bohl *et al.*, 2012: 17).

While the 'original' taxonomy is based on six major categories of the cognitive process – remember, understand, apply, analyse, evaluate, and create – this taxonomy

has been reduced to three levels and transformed for the discipline of history education (see Kühberger, 2011: 6–7) (see Figure 2.1).

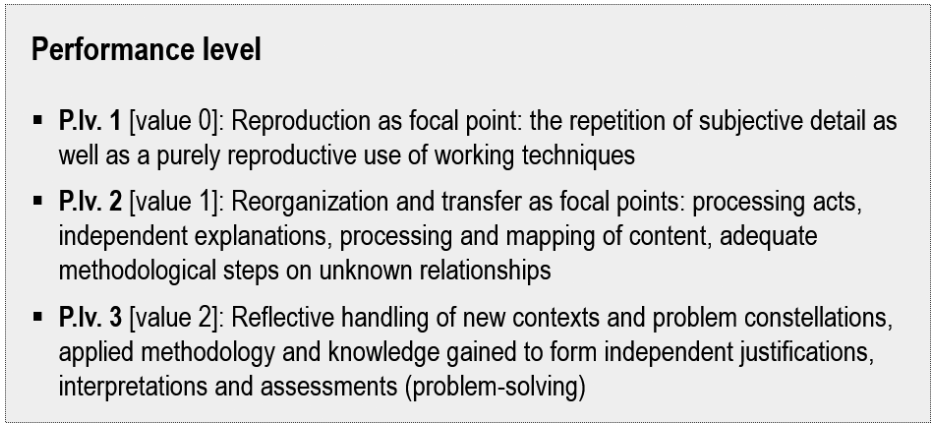


Figure 2.1: Categorical approaches to DTC: Performance level

Since earlier investigations in Austria have already been conducted along this taxonomy (see chapters in Bramann *et al.*, 2018), it seems useful to apply this division again. However, the coding was not derived directly from 'operators' in several national curricula, but rather inferred from the intended (thinking) tasks (see also Bramann, 2018: 193). The analysis shows that less than 5 per cent of the tasks focus on the important level of reflection, independent reasoning, and evaluation.

In this context, approaches to *aspects of differentiated learning* were also evaluated. In the textbook being examined, the following moments were coded: references to materials to use, indications of the degree of difficulty, assistance through method pages, and instructions for the assignment to solve the problem (Bohl *et al.*, 2012: 39). The results do not evaluate the quality of the textbook but count the forms that influence complexity (see Figure 2.2).

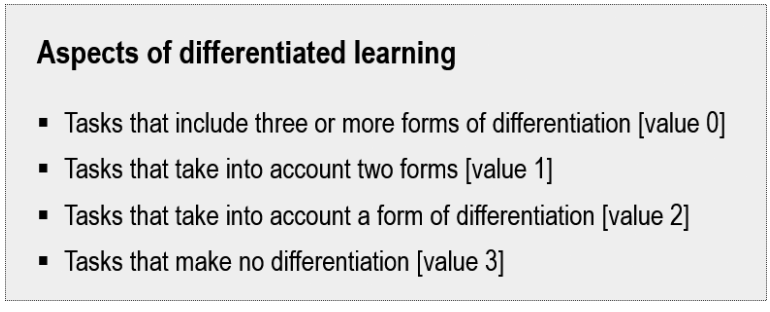


Figure 2.2: Categorical approaches to DTC: Aspects of differentiated learning

The complexity of a learning task is also determined by the interrelations of characteristics to be taken into account in the process. This influences the complexity as to whether learners are faced with tasks that are part of their daily world (*real-life connection*) or not. If a challenge relates to students' real life, this reduces task complexity. In accordance with the already existing discussion in this area, a real-life connection is here understood as a relation between domain-specific knowledge and experience and students' real life (Maier *et al.*, 2010: 89) (see Figure 2.3).

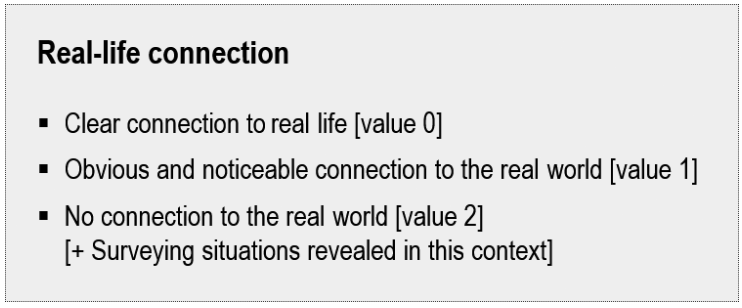


Figure 2.3: Categorical approaches to DTC: Real-life connection

At various points in history education, the argument is made that historical thinking is also expressed by using *domain-specific concepts* (Kühberger, 2012). Tasks should challenge thinkers to open up the mental operations involved in the task, relating them to specific questions, approaches, and concepts (Kühberger, 2011: 8). By assuming that complexity reveals itself in various dependencies, it is also increased by domain-specific concepts (see Figure 2.4).

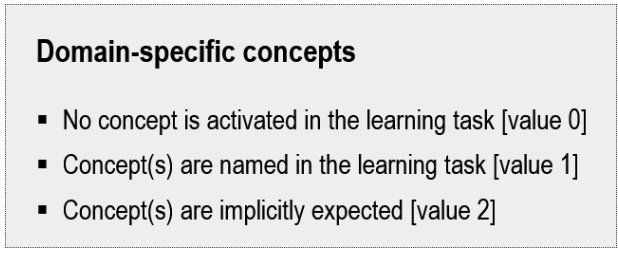


Figure 2.4: Categorical approaches to DTC: Domain-specific concepts

For the complexity of tasks, the *integration of tasks in task sets* is also important. It is expected that an isolated task has lesser complexity, as there may not be any sequence errors. Different tasks that deal with different problems, but build on the same content, are regarded as isolated tasks. Only one task from the history textbook was not isolated, because it built on an answer from another task (see Figure 2.5).

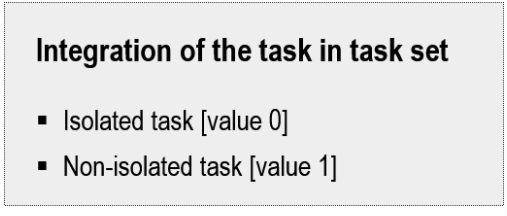


Figure 2.5: Categorical approaches to DTC: Integration of the task in task set

The aspect of *multimodal complexity* of learning tasks in history textbooks considers that modern textbooks present multimodal design opportunities for domain-specific learning. They are not only structured by the author's texts but also by a wide range of additional modules (such as pictures, graphs, and other narratives of historians), which, together with visual elements (such as layout and free space), represent a historical

narrative (Kühberger, 2016: 70–1; Bramann, 2017: 70–1). Learning tasks are part of this multimodality and interact with different elements of a textbook's (double) page.

Multimodal complexity of learning tasks, as used here, does not focus on the narrative interactions of all features, but tries to reveal the interwoven configuration between different kinds of modules created by the (reconstructed) aim of the tasks. To clarify this aspect of complexity in history textbooks, the quantity of implicit and explicit references between the tasks and other elements of the textbooks was counted. In this context, modules are understood as closed areas in the textbook (see Figure 2.6). Furthermore, the different types or genres (author text, historical source, narrative of a historian) controlled by the learning tasks were coded. The results show that the majority of all tasks rely on more than three different elements and, while some tasks do not use the textbook at all, others demand the integration of all the elements presented on a double page of a textbook.

Multimodality / Multimodal complexity (MMC)

- Number of element (and genres) must be taken into account (DTC = 0)
- Number of elements to be taken into account for completing a task
- Number of different genres to be taken into account for completing a task

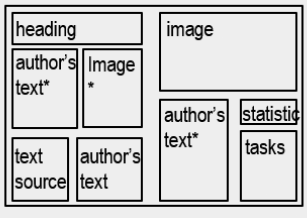
Mathematical term DTC_{MMC}

$$DTC_{MMC} = \sum_{k=1}^{k=n} s_k \sum_{l=0}^{l=p} u_l$$

k = considered module/s
n = sum of the included modules
l = different genres
p = sum of the different genres

Example: ***Name three reasons for the outbreak of the First World War!***

$$DTC_{MMC} = \sum_{k=0}^{k=3} 3 \sum_{l=0}^{l=2} 2 = (1+1+1) \cdot (1+1) = 6$$



* used elements for solving a task on a structured visualisation of elements on a textbook's double page

Figure 2.6: Categorical approaches to DTC: Multimodality / Multimodal complexity (MMC)

The results show that the majority of tasks with a total score of 11 or less can be clearly marked as less complex (see Figure 3). Only a few tasks reach a higher DTC.

Following the theoretical construct, it is especially the multimodal complexity that heavily influences the result. In doing so, it becomes evident that the sum score of task complexity without regard to multimodality is between 2 and 6, which represents a relatively narrow range (see Figure 4).

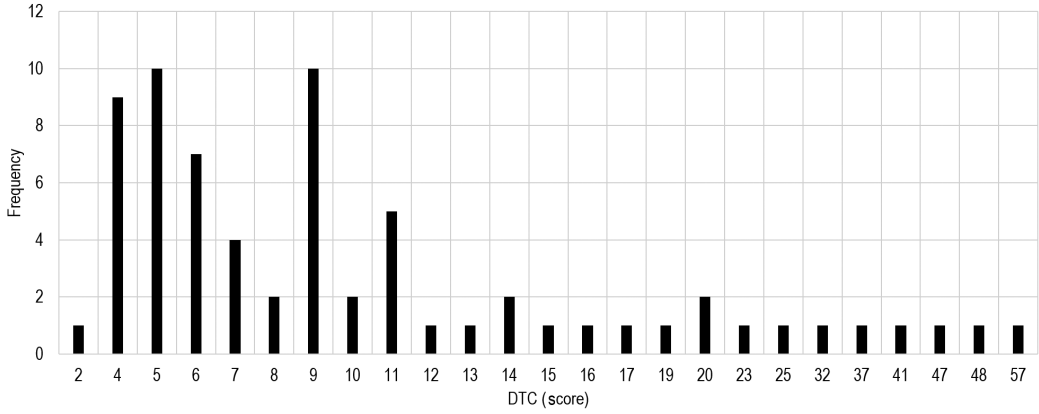


Figure 3: Domain-specific task complexity (DTC) per sum score/frequency

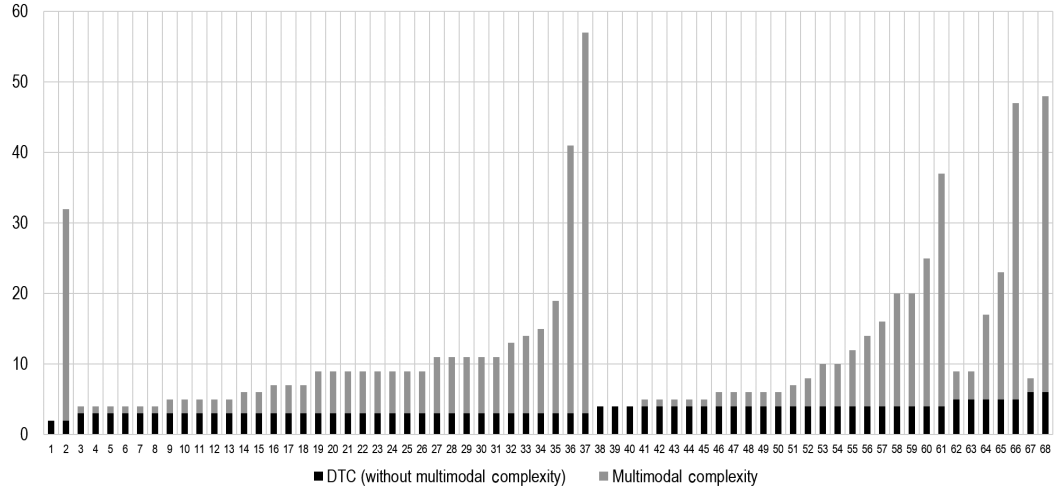


Figure 4: Domain-specific task complexity (per task)

General task complexity (GTC)

To empirically explore the complexity of tasks in general, it makes sense to consider tasks as entities that, in a first approximation, are considered independently by individual processing modes subjects requiring critical thinking, in order to make stimuli and characteristics describable and to apply them as component parts. For this, 'behaviour requirements' and 'pattern of stimuli' are analysed here (Wood, 1986: 62). As the tasks require actions and processing variables that form a limit in terms of knowledge, skills, and resources that subjects demanding critical thinking require to solve a task, these actions are important task components that can be described as *component complexity*:

> The component complexity of a task is a direct function of the number of distinct acts that need to be executed in the performance of the task

> and the number of distinct information cues that must be processed in the performance of those acts. … As the number of acts increases the knowledge and skill requirements for a task also increase, simply because there are more activities and events that an individual needs to be aware of and able to perform. (ibid.: 66)

Therefore, with an increase of information variables, the number of actions to be determined also increases. It has been stressed that complexity is reduced when an overlap of requirements (component redundancy) is repeatedly inserted in the same acts or by placing redundant information (ibid.). Furthermore, it is important to note whether there are tasks within tasks (subtasks). Subtasks are understood as parts of a task that are a component part of the task. This will account only for explicit moments. The calculation shown in this paper of general task complexity refers to component complexity (GTC_1). It focuses on the acts at the level of the subtasks of a task (see Figure 5.1).

Component complexity (GTC_1)

Mathematical term GTC_1:

$$GTC_1 = \sum_{j=0}^{j=p} \sum_{i=1}^{i=n} W_{ij}$$

i = distinct act in subtasks
n = number of distinct acts in subtasks
j = subtask
p = number of subtasks
W_{ij} = number of information cues to be processed in the performance = p × n

Example: ***Name three reasons for the outbreak of the First World War!***

$$GTC_1 = \sum_{j=0}^{j=0} \sum_{i=1}^{i=0} 0$$

i = distinct act in subtasks
n = number of distinct acts in subtasks (= 0)
j = subtask (= 0)
p = number of subtasks (= 0)
W_{ij} = number of information cues to be processed in the performance (= p × n = 0)

Figure 5.1: Categorical approaches to GTC: Component complexity (GTC_1)

Coordinative complexity (GTC_2)

Mathematical term GTC_2:

$$GTC_2 = \sum_{i=1}^{i=n} r_i$$

n = number of treatments (acts)
i = treatment (act)
r_i = number of previous relations between r_{th} and all other treatments (acts) in the task

Example: ***Name three reasons for the outbreak of the First World War!***

$$GTC_2 = \sum_{i=1}^{i=1} 0$$

i = 1
r = 0 [no relation]

Example: ***Describe, analyse and interpret the royal painting!***

Relation 1 Relation 2

$$GTC_2 = \sum_{i=1}^{i=3} 2$$

n = 3
$r_1 \ldots r_2 \ldots r_3 = 2$ [Relation 1 between r_1+r_2; Relation 2 between r_2+r_3]

Figure 5.2: Categorical approaches to GTC: Coordinative complexity (GTC_2)

Additionally, a *coordinative complexity* (GTC_2) has been identified. It focuses on the form of the relationship between information items, actions, and the resulting products, as well as their sequencing (see Figure 5.2):

> The more complex the timing, frequency, intensity, and location requirements, the greater the knowledge and skill an individual must have to be able to perform the task. The appropriate index for coordinative complexity of a task will depend upon the specific aspects of the relationship between task inputs that are being considered. … As the number of precedence relationships for the coordination of acts will also increase because individuals who perform the task will have to learn and perform longer sequences of acts. (ibid.: 69)

Coordinative complexity stresses that there is a difference in complexity when the acts that are set in a specific order are subject to variation, or when the sequence of actions is combined. In this sense, variations as special relations are coded and transformed into a term (Oeser and O'Brien, 1967: 91–2).

Lastly *dynamic complexity* can be made out to be a quality of complexity (Schoeneberg, 2014: 14), which also refers to the performance of the task (see Figure 5.3). This focuses on changes that may occur during the processing of tasks due to causal chain or means–purpose hierarchies. For tasks that have a dynamic complexity, the a priori identified conditions change (that is, *acts* and *cues*), and thus the relationship to the anticipated product in the course of processing: 'Changes in either the set of required acts and information cues or the relationships between inputs and products can create shifts in the knowledge or skills required for a task' (Wood, 1986: 71). In connection with tasks of historical learning, such constellations can be determined in tasks, which include an independent research and/or a production of a historical narration where the methodical way, the historical sources and the domain-specific concepts are not pre-structured. Such tasks are also called 'open tasks' or 'opened tasks' (Kühberger, 2014: 39–49). As the moments of dynamic complexity are only observable on the basis of the concrete solution-based actions of thinking subjects, in the present analysis learning tasks from history textbooks are only highlighted if such dynamic factors are to be expected during processing due to different structures of openness (ibid.).

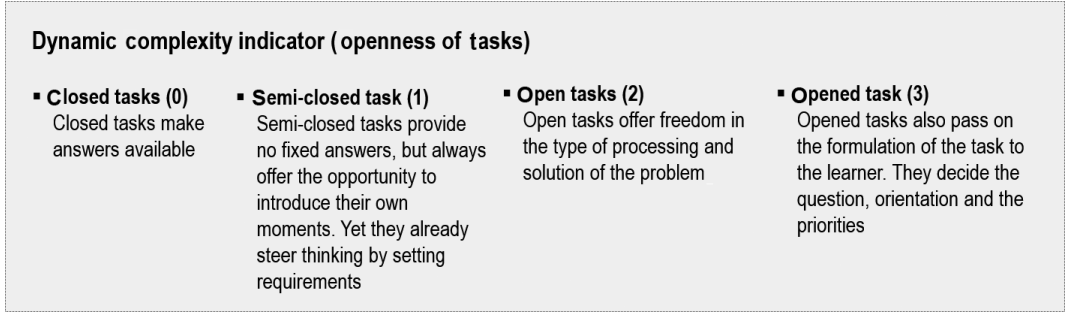


Figure 5.3: Categorical approaches to GTC: Dynamic complexity indicator

In order to make all learning tasks of a textbook comparable, a sum (total score) is formed from GTC_1, GTC_2, and the dynamic complexity indicator. In comparing the total scores of the tasks from the coded history textbook, nothing unexpected shows up. Along the theoretical construction, all tasks that display the various acts, subtasks, and relations, and have openness, appear more complex in the ranking of the total score. The following moments are noteworthy (see Figure 6).

Example	Generalization	GTC
Why are wars fought? Underline in T1!	Tasks that reproduce some aspects from the textbook are less complex.	1
Process M5 per the "Read the pictures" method on page 132!	Tasks that work with methodological requirements are always exceptionally complex. Their degree of complexity is often endless.	47

Figure 6: General task complexity (total score)

Thus, for the textbook examined, a *general task complexity* unfolds, which makes it clear that the tasks are in a less complex area, while there are certainly tasks that, as mentioned above, all work with methodical support and demonstrate a high degree of complexity in the textbook (see Figure 7).

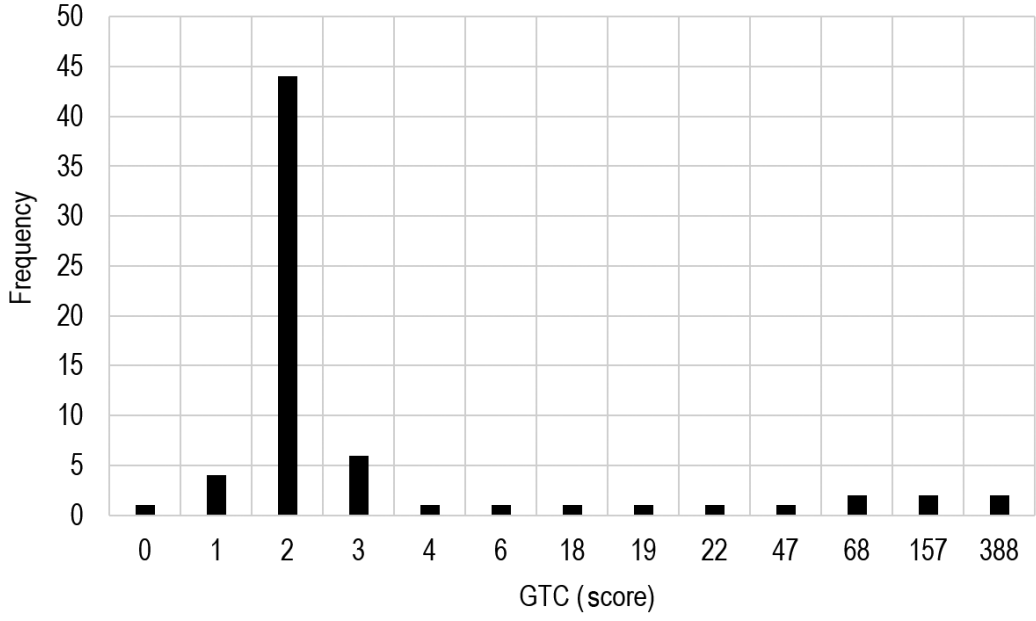


Figure 7: General task complexity (GTC) per sum score/frequency

General linguistic complexity (GLC)

As a third perspective on complexity, we include general linguistic complexity (GLC) in our evaluation. We follow the well-established second-language acquisition (SLA) tradition of analysing language performance by assessing the multidimensional construct of linguistic complexity in terms of syntactic, lexical, and discursive elaborateness, variation, and inter-relatedness, as well as of language use and human language processing. Similar measures have been used in previous research to assess the adaptation of reading demands in German geography schoolbooks to different school types and grade levels (Berendes *et al.*, 2018). The general linguistic complexity of student answers, which relates to research on tasks in foreign-language learning (Alexopoulou *et al.*, 2017) provides first insights into the relationship between task complexity and general linguistic complexity. We extract complexity measures for German based on SLA research and psycholinguistics using the system by Weiß and Meurers (2018; in press), excluding measures of cohesion and grammatical variation that are not meaningful for the short textbook tasks. To aggregate the remaining 215

indices to a single GLC score that is readily interpretable in an education context, we trained a machine learning model with these measures on reference texts. While for English the Common Core State Standards provide a reference data set of texts that children at different grade levels should be able to read (CCSSI, 2010), no such externally validated data set is accessible for German. Given that the development of reading and writing abilities of students are linked, we focused on student writings from the Karlsruhe Children's Text (KCT) corpus (Lavalley *et al.*, 2015), one of the few student writing corpora available for German. We selected 1,470 texts written in free writing tasks by students of Grades 3 to 8 (mean ages rounded up of 9, 10, 11, 12, 13, 14 respectively). Each level is represented by 212 to 283 text instances. We grouped grade levels into pairs, resulting in the classification levels 3/4, 5/6, and 7/8.

Given the limited amount of data, types of writing tasks, grade levels/school types, and the gap between passive and active language knowledge, the machine learner trained on this data set clearly only provides a first approximation of the potential spectrum of linguistic complexity. There is a clear need for externally validated reading and writing reference corpora for German, in order to obtain more firmly grounded interpretations of the evidence becoming available through the broad range of linguistic complexity measures. On the KCT data subset, we trained the Simple Logistic Regression algorithm from the WEKA Machine Learning toolkit (Smith and Frank, 2016), which performs cross-validated feature selection using LogitBoost with simple regression functions during parameter estimation (Landwehr *et al.*, 2005). The classification result is aggregated from three separate, binary regression terms, each determining the affiliation to a certain grade level through feature weights (see Figure 8). The example shows two features of syntactic complexity from the term for Grades 7/8 applied to a schoolbook task.

It contains two noun phrases (NP) and one prepositional phrase (PP) as post-nominal modifier. PPs per sentence (1/1) receives a weight of 0.19, postnominal modifiers per NP (1/2) a weight of 2.39. Both features are positively correlated with Grade 7/8, but post-nominal modifiers are more crucial.

Overall, out of the 215 features, the system attributed non-zero weights to 56, 75, and 45 features for Grades 3/4, 5/6, and 7/8 respectively, all including features of syntactic, lexical, and morphological complexity, human cognitive processing, and language use.

Mathematical term GLC_g:

$$GLC_g = \alpha_g + \sum_{i=1}^{i=n} \beta_{gi}\, \varphi_{gi}$$

g = grade level (3/4, 5/6, 7/8)
α = intercept
i = feature index (with n = 215)
β = feature weight
φ = complexity feature

Example: *Explain the Headline of T1!*

$$GLC_{7/8} = -13.3 + 0.19 * \frac{PP}{Sent} + \cdots + 2.39 * \frac{p.Mod.}{NP}$$

PP = prepositonal phrase *zu T1*
NP = noun phrase: *die Überschrift*, *T1*
p.Mod. = postnominal modifier: *zu T1*
Sent = sentence (*Erkläre die Überschrift zu T1!)*

Figure 8: Categorical approaches to general linguistic complexity (GLC)

For *syntactic complexity*, aside from the two measures used for illustration in Figure 8, comparative noun modifiers and complex clausal structures receive high positive weights in GLC 7/8. GLC 5/6 places positive weights on pre-nominal participles and clausal noun modification, whereas post-nominal and comparative noun modifiers receive negative weights. This shows that across grade levels, NPs are of different complexity across grades, and that clausal modification is more relevant at higher grade levels. Grades 5/6 are also associated with more modal verb clusters than any other grade level and with passive constructions. Writings at Grades 3/4 are associated with unmodified noun phrases (NPs), placing high negative weights on the overall frequency of complex NPs. They also exhibit increased use of *to*-infinitives, but they are characterized by a general lack of sentential sophistication.

The highest feature weights are assigned to indices of *morphological complexity* in all GLC models: Grades 3/4 are associated with high negative weights for noun suffixes with Latinate or Greek origin (for example, *-atur*), while they receive high positive weights in the other models. Also, Grade 3/4 writing is assumed to contain nouns derived from a verb, which is atypical for Grade 7/8. This shows increased use of other derived nouns. Grade 5/6 texts are less characterized in terms of derivation, but show a tendency towards non-nominative case nouns.

Lexical complexity plays a less pronounced role in the models, except for GLC 3/4, which punishes word length with relatively high negative weights. Also, Grade 3/4 writings are characterized as less semantically inter-related, while Grade 7/8 writings are characterized in terms of more specific words, that is, words with more hypernyms, and words with more semantic interrelations.

In terms of *human processing cost*, GLC 3/4 assigns negative weights to indices of cognitive processing load. This grade level is thus associated with less cognitively demanding sentences. In contrast, GLC 5/6 and GLC 7/8 place increasingly high positive weights on these measures.

The models also employ measures of *language use:* writings of Grades 3/4 and 5/6 are associated with the use of more frequent words, while writings from Grades 7/8 are expected to use less frequent vocabulary.

Before applying the models to the textbook tasks, we evaluated their performance on the KCT data using ten-fold cross-validation. Their average classification performance is F1 = 75.1 per cent, with misclassifications occurring predominantly between adjacent grades, for example fourth being confused with fifth grade. Compared to the random classification baseline of 33.3 per cent, the result confirms that the models successfully differentiate between children's writing across grade levels, and should be applicable to textbook tasks after being retrained on the full training set.

The results of applying the full KCT-based GLC model on the examined tasks may be seen in the plot in Figure 9. Overall, 60 of 68 tasks were assigned a GLC score of 7/8, two tasks exhibit a GLC of 5/6, and six tasks obtained GLC scores of 3/4. Tasks with GLC scores of 3/4 and 5/6 clearly rate below the GLC that was observed in the writings of the German peers to the textbook's target audience. They should thus be easily comprehensible for the target audience of eighth-grade students. The 60 tasks that received GLC scores of 7/8 are at or above the level of the target audience. Since our model does not include scores above grades 7/8, the results very likely show a ceiling effect. Some tasks might receive higher ratings by a model trained on texts from higher grades, and might thus prove to exceed the competence level of the target audience. Unfortunately, there is currently no such data set available to follow up on that issue.

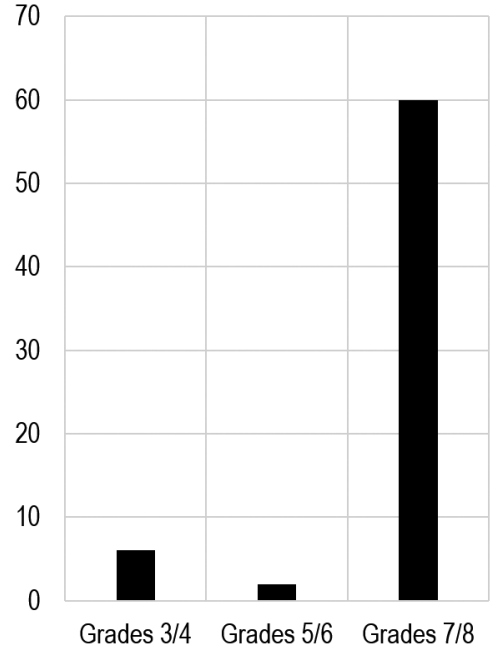


Figure 9: Predicted grade levels (GLC)

Despite these limitations, our operationalization of GLC successfully differentiates between tasks based on a diverse set of features of linguistic complexity. Figure 10 illustrates these for all three GLC levels.

Example	Generalization	GLC
Why are wars fought? Underline in T1!	Tasks with simple, short sentences, short, frequent words, and no nominal modifications	3/4
Work with M4! What is the left poster supposed to make afraid of?	Tasks with some noun modification and passive voice, but predominantly simple sentences and commonly used vocabulary	5/6
Write down, why only few people offered resistance! (T1)	Tasks with clausal subordination, less frequent and semantically more connected vocabulary	7/8

Figure 10: General linguistic complexity (Total score)

Comparison of results and conclusion

Without going into detail on individual results, it has been shown that the different models for representing complexity have led to very different results depending on the theoretical focus and on its representation as numeric data, as the following task illustrates:

> *Process M1 per the 'Analyse posters' method on page 9!*
> [Original tasks: *Bearbeite M1 nach der Methode 'Plakate analysieren' auf Seite 9!*] (Bachlechner *et al.*, 2012: 35)

This task has a low total score in the DTC (= 4). However, the GTC demonstrates an exceptionally high level, endlessly pointing in an infinite direction (GTC = 385). In terms of GLC, the task also scores high (GLC = 7/8) due to the frequency of prepositional phrases and the complexity of the noun phrase.

Following the theoretical aspects of the triangulated research approach, this example illustrates that a triangulation of coded tasks by different theoretical constructs is necessary to avoid the risk of following only one model of complexity. At the same time, it cannot be a question of simply favouring one of the models, because they offer information on very different aspects of complexity. Combining the complexity scores obtained from all three models would be possible, because they encode different moments. This would, however, not reveal more specific insights into the tasks, due to the levelling effect of the GTC.

If one ultimately compares the results of the GTC with the DTC, this at least demonstrates a trend for the evaluated textbook, namely that there is a tendency according to which less-complex tasks clearly dominate (see Figure 3 and Figure 7). Given these results, it can be assumed that the authors of history textbooks are not aware that a task can grow in complexity that is too high (if not even infinite) due to the many acts in connection with materials from the textbook, which may be restricted only by an intervention on the part of the teacher or that may be borne from an implicit *commitment* between students and teachers, whose conditions are known from a different context and/or have been learned through repeated application.

When comparing the GLC of the tasks with their GTC and DTC, another interesting pattern emerges: while tasks show high GLC (7/8) irrespective of their GTC or DTC, medium (5/6) and low (3/4) GLC is only observed for tasks with GTCs ≤ 7 and DTCs ≤ 9 (see Figure 11 and Figure 12). Thus, combined low GTC and DTC seems to align with reduced GLC. However, since this observation is only supported by eight tasks, and lower scores for GTC and DTC are more common than higher tasks, this remains a tentative first impression until more data are collected.

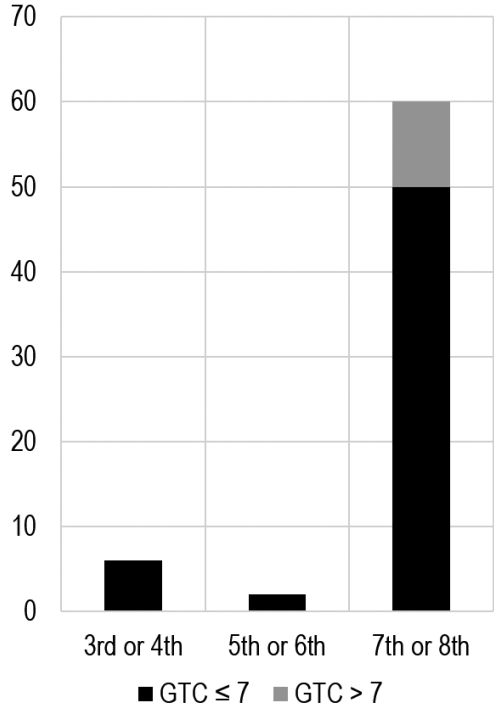


Figure 11: GTC related to GLC

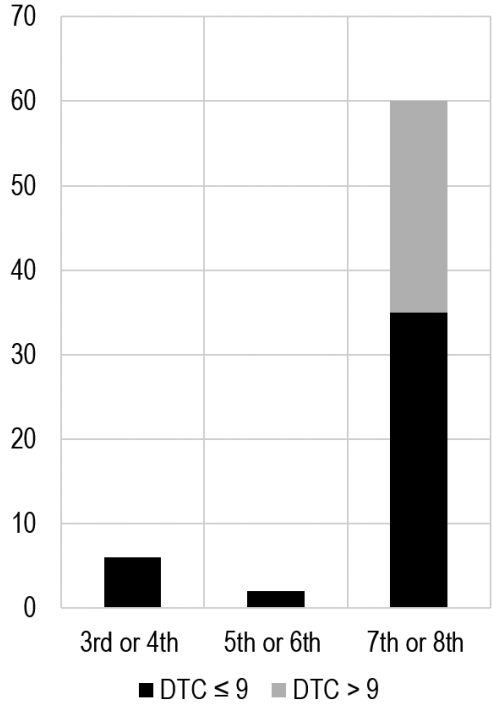


Figure 12: DTC related to GLC

In order to investigate the link between GLC and GTC/DTC further, we calculated the Pearson correlation of individual complexity features. (We excluded outliers deviating more than 2 standard deviations from the mean and performed log transformation if necessary.) A highly significant ($\alpha \leq 0.001$), moderate positive correlation ($r = 0.51$) was found between the Log-GTC and infinitives per word, and highly significant, weak positive correlations for lexical types and lemmas present in the KCT corpus ($r = 0.44$), VP length in words ($r = 0.4$), noun dependents per noun with dependents ($r = 0.38$), and several type and lemma frequencies from the dlexDB data base ($r = 0.35$ to $r = 0.34$). Furthermore, a significant ($\alpha \leq 0.05$), weak negative correlation was found for nominative case markings per noun ($r = –0.33$). For the logged DTC, we found significant, weak positive correlations for noun modifiers per NP ($r = 0.36$) and pre-nominal modifiers per NP ($r = 0.33$), as well as for several dlexDB and KCT age-of-active-use-based frequency measures ($r = 0.32$ to $r = 0.31$). These measures are identical or belong to the same type of GLC features that receive high feature weights in the GLC models. Hence, the findings support the initial impression of a link between the linguistic complexity of a task and its GTC or DTC, especially in terms of language use and phrasal complexity.

At this point in the study, or in the trial phase of coding, it especially makes sense to code more textbooks and to compare the results to be able to derive structural insights for each school textbook from them. Because the chosen research approach consists of measuring the general, linguistic, and domain-specific aspects of learning tasks, and shows how to analyse learning tasks in a goal-oriented manner, the differentiated outcomes can help textbook authors to develop specific tasks in the future that take into account different levels of student abilities. Triangulation as a recent approach in the field of history education is an area that remains central for the comparison of results for differentiated insight into history textbooks and the tasks presented there.

Notes on the contributors

Christoph Kühberger is Professor for History and Civic Education at the University of Salzburg (Austria). He was Professor for European Cultural History at the University of Hildesheim (Germany) and Professor for History and Civic Education at the Pedagogical University of Salzburg Stefan Zweig (Austria). His current research interests include history education and civic education, ethnography, historical culture, new cultural history and ethics of historical sciences.

Christoph Bramann is a scientific assistant at the Department of History at the University of Bochum (Germany) and a lecturer in history and civic education at the University of Salzburg (Austria). He studied history and German studies at the University of Frankfurt (Germany) and was a scientific assistant at the University of Salzburg. His main research fields are history education, task research, textbook research and empirical research.

Zarah Weiß is a scientific assistant at the Department of General and Computational Linguistics at the University of Tübingen (Germany). She studied German studies, general linguistics, and computational linguistics at the universities of Frankfurt (Germany) and Tübingen (Germany). Her research currently focuses on linguistic and statistical aspects of language and data modelling, in particular with regard to language complexity, first- and second-language proficiency and development, task effects and text readability.

Detmar Meurers is Professor of Computational Linguistics at the University of Tübingen (Germany). He was previously an associate professor at the Ohio State University (USA), a professor II at the University of Tromsø (Norway), and a professeur invité de première classe at Université Paris Diderot (France). His research explores computational linguistic methods in education and foreign language teaching and learning, including a focus on adaptive and interactive materials.

References

Alexopoulou, T., Michel, M., Murakami, A. and Meurers, D. (2017) 'Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques'. *Language Learning*, 67 (S1), 180–208.

Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J. and Wittrock M.C. (2001) *A Taxonomy for Learning, Teaching, and Assessing: A revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman.

Bachlechner, M., Benedik, C., Graf, F., Niedertscheider, F. and Senfter, M. (2012) *Bausteine 4: Geschichte, Sozialkunde, Politische Bildung*. Wien: öbv.

Berendes, K., Vajjala, S., Meurers, D., Bryant, D., Wagner, W., Chinkina, M. and Trautwein, U. (2018) 'Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track?'. *Journal of Educational Psychology*, 110 (4), 518–43.

Bernhard, R. (2016) 'Lernaufgaben zur Förderung historischer Denkprozesse: Normative Richtlinien für Geschichtsschulbücher und deren Implementierung in Österreich'. In Matthes, E. and Schütze, S. (eds) *Schulbücher auf dem Prüfstand – Textbooks under Scrutiny*. Bad Heilbrunn: Klinkhardt, 243–53.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H. and Krathwohl, D.R. (1956) *Taxonomy of Educational Objectives*. New York: Longman.

Bohl, T., Kleinknecht, M., Batzel, A. and Richey, P. (2012) *Aufgabenkultur in der Schule: Eine vergleichende Analyse von Aufgaben und Lehrerhandeln im Hauptschul-, Realschul- und Gymnasialunterricht*. Baltmannsweiler: Schneider.

Bramann, C. (2017) 'Arbeiten mit dem Geschichtsschulbuch? Zur paradoxen Stellung eines Leitmediums in Unterricht und Lehrkräfteausbildung'. *ph-script: Beiträge aus Wissenschaft und Lehre*, 12, 69–76. Online. https://tinyurl.com/y9dejv47 (accessed 27 December 2018).

Bramann, C. (2018) 'Historisch Denken lernen mit Schulbuchaufgaben? Medienspezifische Analyse von Arbeitsaufträgen in österreichischen Geschichtsschulbüchern'. In Bramann, C., Kühberger, C. and Bernhard, R. (eds) *Historisch Denken lernen mit Schulbüchern*. Frankfurt am Main: Wochenschau Verlag, 181–214.

CCSSI (Common Core State Standards Initiative) (2010) *Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects*. Washington, DC: Common Core State Standards Initiative. Online. www.corestandards.org/assets/CCSSI_ELAStandards.pdf (accessed 10 January 2018).

Dörner, D. (1989) *Die Logik des Mißlingens: Strategisches Denken in komplexen Situationen*. Reinbek bei Hamburg: Rowohlt.

Elen, J. and Clark, R.E. (eds) (2006) *Handling Complexity in Learning Environments: Theory and research*. Amsterdam: Elsevier.

Ercikan, K. and Seixas, P. (eds) (2015) *New Directions in Assessing Historical Thinking*. New York: Routledge.

Flick, U. (2003) 'Triangulation in der qualitativen Forschung'. In Flick, U., Von Kardorff, E. and Steinke, I. (eds) *Qualitative Forschung: Ein Handbuch*. Reinbek bei Hamburg: Rowohlt, 309–18.

Gürtler, L. and Huber, G.L. (2012) 'Triangulation: Vergleiche und Schlussfolgerungen auf der Ebene der Datenanalyse'. In Gläser-Zikuda, M., Seidel, T., Rohlfs, C., Gröschner, A. and Ziegelbauer, S. (eds) *Mixed Methods in der empirischen Bildungsforschung*. Münster: Waxmann, 37–50.

Heuer, C. (2011) ‚Gütekriterien für kompetenzorientierte Lernaufgaben im Fach Geschichte'. *Geschichte in Wissenschaft und Unterricht*, 62 (7–8), 443–58.

Housen, A., Kuiken, F. and Vedder, I. (2012) *Dimensions of L2 Performance and Proficiency: Complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins.

Kelle, U. (2006) 'Combining qualitative and quantitative methods in research practice: Purposes and advantages'. *Qualitative Research in Psychology*, 3 (4), 293–311.

Körber, A., Schreiber, W. and Schöner, A. (eds) (2007) *Kompetenzen historischen Denkens: Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik*. Neuried: Ars Una.

Köster, M., Bernhardt, M. and Thünemann, H. (2016) 'Aufgaben im Geschichtsunterricht'. *Geschichte Lernen*, 29 (174), 2–11.

Kühberger, C. (2011) 'Aufgabenarchitektur für den kompetenzorientierten Geschichtsunterricht: Geschichtsdidaktische Verortungen von Prüfungsaufgaben vor dem Hintergrund der österreichischen Reife- und Diplomprüfungsreform'. *Historische Sozialkunde*, 41 (1), 3–13.

Kühberger, C. (ed.) (2012) *Historisches Wissen: Geschichtsdidaktische Erkundung zu Art, Tiefe und Umfang für das historische Lernen*. Schwalbach am Taunus: Wochenschau Verlag.

Kühberger, C. (2014) *Leistungsfeststellung im Geschichtsunterricht: Diagnose – Bewertung – Beurteilung*. Schwalbach am Taunus: Wochenschau Verlag.

Kühberger, C. (2016) 'Intertextual and multi-modal construction of history via textbooks and its reception'. In Lehmann, K., Werner, M. and Zabold, S. (eds) *Historisches denken jetzt und in Zukunft*. Münster: Lit, 67–81.

Landwehr, N., Hall, M. and Frank, E. (2005) 'Logistic model trees'. *Machine Learning*, 59 (1–2), 161–205.

Lavalley, R., Berkling, K. and Stüker, S. (2015) 'Preparing children's writing database for automated processing'. Paper presented at the Language Teaching, Learning and Technology (LTLT) Workshop, Leipzig, 4 September 2015.

Leisen, J. (2010) 'Lernaufgaben als Lernumgebung zur Steuerung von Lernprozessen'. In Kiper, H., Meints, W., Peters, S., Schlump, S. and Schmit, S. (eds) *Lernaufgaben und Lernmaterialien im kompetenzorientierten Unterricht*. Stuttgart: Kohlhammer, 60–7.

Maier, U., Kleinknecht, M., Metz, K. and Bohl, T. (2010) 'Ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben'. *Beiträge zur Lehrerbildung*, 28 (1), 84–96.

Oeser, O.A. and O'Brien, G. (1967) 'A mathematical model for structural role theory, III'. *Human Relations*, 20 (1), 83–97.

Robinson, P. (2001) 'Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA'. In Robinson, P. (ed.) *Cognition and Second Language Instruction*. Cambridge: CUP, 287–318.

Schoeneberg, K.-P. (2014) 'Komplexität – Einführung in die Komplexitätsforschung und Herausforderungen für die Praxis'. In Schoeneberg, K.-P. (ed.) *Komplexitätsmanagement in Unternehmen: Herausforderungen im Umgang mit Dynamik, Unsicherheit und Komplexität meistern*. Wiesbaden: Springer, 13–27.

Smith, T.C. and Frank, E. (2016) 'Introducing machine learning concepts with WEKA'. In Mathé, E. and Davis, S. (eds) *Statistical Genomics: Methods and protocols*. New York: Springer, 353–78.

Thünemann, H. (2013) 'Historische Lernaufgaben: Theoretische Überlegungen, empirische Befunde und forschungspragmatische Perspektiven'. *Zeitschrift für Geschichtsdidaktik*, 12 (1), 141–55.

Von Borries, B., Fischer, C., Leutner-Ramme, S. and Meyer-Hamme, J. (2005) *Schulbuchverständnis, Richtlinienbenutzung und Reflexionsprozesse im Geschichtsunterricht: Eine qualitativ-quantitative Schüler- und Lehrerbefragung im Deutschsprachigen Bildungswesen 2002*. Bayerische Studien zur Geschichtsdidaktik, Vol. 9. Neuried: Ars Una.

Waldis, M., Hodel, J. and Fink, N. (2012) 'Lernaufgaben im Geschichtsunterricht und ihr Potential zur Förderung historischer Kompetenzen'. *Zeitschrift für Didaktik der Gesellschaftswissenschaften*, 3 (1), 142–57.

Weiß, Z.L. and Meurers, D. (2018) 'Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation'. In Bender, E.M., Derczynski, L. and Isabelle, P. (eds) *Proceedings of the 27th International Conference on Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 303–17.

Weiß, Z.L. and Meurers, D. (in press) 'Broad linguistic modeling is beneficial for German L2 proficiency assessment'. In Abel, A., Glaznieks, A., Lyding, V. and Nicolas, L. (eds) *Widening the Scope of Learner Corpus Research: Selected Papers from the 4th Learner Corpus Research Conference 2017*. Louvain-La-Neuve: Presses Universitaires de Louvain.

Wood, R.E. (1986) 'Task complexity: Definition of the construct'. *Organizational Behavior and Human Decision Processes*, 37 (1), 60–82.