Bernholt, Sascha; Lossjew, Jannik; Gombert, Sebastian

Analyzing students' conceptual understanding over the course of a teaching unit. Tracking changes in knowledge structures over time

Unterrichtswissenschaft 54 (2026) 3, S. 291-323

Unterrichtswissenschaft (2026) 54:291–323
https://doi.org/10.1007/s42010-026-00244-0

THEMENTEIL

Analyzing students' conceptual understanding over the course of a teaching unit: Tracking changes in knowledge structures over time

Sascha Bernholt · Jannik Lossjew · Sebastian Gombert

Received: 23 January 2025 / Revised: 15 May 2026 / Accepted: 17 May 2026 / Published online: 3 June 2026

Abstract The implementation of adaptive teaching poses an enormous challenge to teachers and requires them to perform a balancing act that is difficult to achieve with traditional teaching methods. Digitalization offers the opportunity to use new support tools to record and analyse students' learning status in real time, and to use this information to adapt teaching and provide direct feedback to learners. The present study aims to illustrate an approach for a digitally supported teaching unit, implemented in a regular chemistry classroom, that offers the possibility to monitor students' progress in real time. A unit on chemical kinetics was developed based on a systematic framework that aims to integrate the affordances of coherent and engaging instruction and reliable assessment. After implementation with $N = 300$ students, responses to a series of different tasks throughout the unit were automatically scored to derive individual knowledge networks that reflect students' ability to enact and connect specific knowledge elements across the unit. Based on both a normative and the individual student networks, specific network parameters were derived to allow comparative analyses. All metrics showed correlations with learner performance at the end of the unit, although different patterns emerged. While the network-level metrics provided insight into students' progress along the unit, the evaluation of the node-level metrics underscored the central role of basic concepts for students' cumulative learning. Overall, the results provide clues to identify potential areas for intervention during the unit, e.g. in the form of task-based feedback or identification of specific difficulties experienced by students.

✉ Sascha Bernholt · Jannik Lossjew
Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik,
Olshausenstr. 62, 24118 Kiel, Germany
E-Mail: bernholt@leibniz-ipn.de

Sebastian Gombert
Leibniz Institute for Research and Information in Education, Rostocker Str. 6, 60323 Frankfurt am Main, Germany

Keywords Individualized instruction · Automated scoring · Formative assessment · Network analysis

1 Introduction

The increasing diversity of the student body with its different individual, cultural, and family backgrounds requires a professional adaptation of teaching to these heterogeneous learning requirements (Prediger and von Aufschnaiter 2017). In order to promote equal opportunities, social participation, and cognitively activating lessons, there is an increasing focus on adaptive lesson design (Dumont 2016). Instead of using standardized methods for all learners, individualized approaches must be created that address individual needs in the student's learning process. However, the implementation of adaptive teaching poses an enormous challenge for teachers and requires them to perform a balancing act that can hardly be achieved with conventional teaching methods.

The overall idea of adapting instruction to students' needs is not new (Skinner 1958); however, this requires the implementation of suitable formative diagnostics as this is an essential prerequisite to being able to respond to the individual needs of learners and also to enabling targeted instructional decisions (Blömeke et al. 2020). Digitalization offers the opportunity to use new support tools to record and analyze the learning status of students in real time and to use this information to adapt teaching and to provide direct feedback to students (Blossfeld et al. 2018; Faber et al. 2017). Although continuous assessment and the automated scoring of individual students' learning progress is feasible in principle (e.g., Nakamura et al. 2016), designing digitally supported learning environments that enable both meaningful science learning and the reliable and valid assessment of individual student learning is challenging (Kubsch et al. 2022).

The present study aimed to address this challenge by analyzing the development of students' conceptual understanding over the course of a digitally supported teaching unit implemented in regular chemistry classrooms. Building upon established frameworks of how students' conceptual understanding develops in chemistry (Emden et al. 2018; F. Yan and Talanquer 2016) and in general (Bransford et al. 2000; Linn et al. 2015), the unit was systematically designed to evaluate, as a proof-of-concept study, the extent to which students' engagement with the materials generates data that allow conclusions to be drawn about their individual learning progress. Based on an analysis of students' artifacts (short and long written responses, drawings, etc.) over the course of the unit, the uptake and integration of knowledge elements was modeled by means of semantic knowledge networks to investigate students' knowledge acquisition. Using a theoretical model of student learning (Linn et al. 2015), specific network metrics were identified to trace students' learning trajectories towards a well-connected knowledge base (Bransford et al. 2000). Finally, implications for classroom practice were identified.

2 Theoretical background

In adaptive teaching, suitable formative diagnostics play a pivotal role as they allow teachers to continuously assess the skills and knowledge of their students and select materials appropriate to their learning level (Dumont 2019). However, a series of valid and reliable science assessments (that may even be scored automatically) does not necessarily result in a good curriculum. Consequently, the specific conditions and affordances of formative diagnostics have to be considered together with the design of meaningful and engaging classroom materials to jointly foster students' knowledge and understanding over time (Kubsch et al. 2022). As part of this study, we developed classroom materials based on theoretical and empirical findings on student learning in order to best support students' cumulative competence development. The students' artifacts when working with these materials then served as the basis for deriving the assessments. Previous research in the area of learning progressions (Section 2.1) and knowledge integration (Section 2.2.) was used as a basis for this approach.

2.1 Learning progressions

Scientific concepts serve as crucial tools for interpreting the world, both in scientific and everyday contexts (Murphy 2004). Consequently, fostering a deep conceptual understanding is a primary objective of science education (American Association for the Advancement of Science & National Science Teachers Association 2007; BIMI 2016; KMK 2004). Learning progressions in science education propose hypotheses about how students' understanding of concepts evolves with instruction (Corcoran et al. 2009). These progressions comprise starting and endpoints, with intermediate developmental stages. Lower anchors denote prior knowledge, while upper anchors reflect the desired competence (Corcoran et al. 2009; Duschl et al. 2007).

In the context of school chemistry, the concept of chemical reaction is one of the four basic concepts used to structure chemistry teaching in lower and upper secondary schools in Germany. The respective educational standards stipulate that students should already learn about chemical reactions at the substance level in Grades 5 and 6 (BIMI 2016). According to these standards, students' progression during their lower secondary education can be described as having three stages, ranging from a phenomenological approach to matter (lower anchor) to a differentiated understanding of the composition of matter and its changes (upper anchor; Emden et al. 2018; Weinrich and Talanquer 2015). In these first years of chemistry lessons, the perspective on chemical changes remains descriptive and reactions are presented as proceeding to completion (van Driel et al. 1998). At the upper secondary level, students are required to draw upon mechanistic and causal patterns of explanations to elucidate the emergent course of reactions (Talanquer 2018; F. Yan and Talanquer 2016). Knowledge of chemical kinetics is an indispensable prerequisite to obtain this level of sophistication (BIMI 2016; KMK 2004).

2.2 Knowledge integration

While the learning progressions approach provides subject-specific guidance on how to sequence and structure content and concepts, which also incorporates empirical findings on students' ideas and difficulties in learning the respective content, the knowledge integration (KI) framework by Linn et al. (2015) provides a theoretical underpinning to design and understand learning environments that support learners in building on what they know and using reasoning strategies to make sense of new information. Linn et al. (2015) suggest that KI is more likely to succeed if educators first elicit students' preexisting ideas (Duit et al. 2013), then introduce new normative ideas, help students distinguish between these and their own preexisting ideas, and ultimately encourage reflection. More specifically, inquiry activities need to elicit students' existing ideas about a phenomenon, for example, whether they are aware of the need to develop predictions. The second process according to Linn et al. (2015) focuses on introducing new normative ideas through learning material, which the learners are expected to integrate into their prior knowledge. This may lead to cognitive conflict, which eventually may promote conceptual change (e.g., Chi and Roscoe 2002) or a shift in the prevalence (and use) of competing conceptions (Potvin et al. 2015), which may lead learners to restructure their semantic network (see M. Schneider and Stern 2010). The third process that is crucial for KI is that students learn to distinguish between their own (often nonnormative) ideas and normative ideas. Here, students should investigate and compare the different ideas in terms of their effectiveness to explain scientific phenomena. Finally, learners should be encouraged to reflect on their set of ideas and assess whether they are consistent and whether any questions are left unanswered. At the end of this process, students should have well-developed knowledge networks organized around the central ideas of a domain. The KI perspective is domain-general as it does not specify which ideas can act as resources and which are nonnormative (Linn et al. 2015). Integrating the perspective of learning progression research thus provides the necessary basis for identifying the core ideas and teaching activities that promote the integration of new ideas into existing knowledge networks and ultimately lead to integrated knowledge (Kubsch et al. 2023).

2.3 Network models of knowledge and knowledge acquisition

The concept of integrated knowledge is informed by the notion of semantic networks, which represent explicit declarative knowledge within long-term memory. Gupta et al. (2010) argue that "conceptual knowledge organization is likely to be network-like" (p. 317). In the context of assessing student knowledge (integration), network models of knowledge and knowledge acquisition might provide a more granular focus on students' learning (compared to, e.g., performance indicators such as the percentage of correct answers), especially when it comes to tracking learning progress over time. Accordingly, the uptake and integration of new knowledge elements, as the constituting components of a student's knowledge network (Koedinger et al. 2012), become the focus of analysis.

The KI framework, as outlined in the previous section, delineates specific implications for the structure of students' conceptual networks. Initially, during the onset of a learning phase, students' networks typically exhibit characteristics such as having (1) a high number of nodes with relatively few connecting edges, which results in the networks having (2) a greater number of isolated nodes, and (3) a lower network density (Cromley et al. 2024). However, with progression or through targeted KI instruction, students with an enhanced understanding of scientific concepts tend to emphasize structure-function connections, thereby integrating elements and relations more cohesively. As a result, more proficient students are expected to develop networks characterized by (1) a large number of nodes interconnected by numerous edges, leading to (2) fewer isolated nodes (indicative of fewer clusters), and (3) increased network density. Additionally, (4) the introduction of important scientific concepts through KI instruction should render these concepts central within students' networks. Supporting evidence from KI researchers indicates that after receiving KI instruction, students' knowledge networks exhibit an increase in the number of both nodes and links, a heightened degree centrality for pivotal scientific concepts, and a greater overall network density compared to preinstruction levels (Cromley et al. 2024; Schwendimann and Linn 2016). More specifically, modeling a system as a network facilitates its characterization through numerical metrics, thereby allowing for comparative analyses of the system's components based on local metrics and analyses between different systems using global metrics (Morrison et al. 2022).

2.3.1 Global network metrics

The total number of nodes represents the most fundamental global metric. As the learning process progresses, students will be required to demonstrate their understanding of new knowledge elements for the first time in subsequent tasks. When students incorporate new knowledge elements into their responses, the network size increases. Prior research indicates that the number of nodes in a knowledge network correlates positively with achievement (Freedman et al. 2024; K. Kim 2025) and is positively affected by instruction (Giabbanelli and Tawfik 2021; Kapuza et al. 2020; M. K. Kim and McCarthy 2021). Examining which knowledge elements are incorporated repeatedly over time also provides information about the stability of knowledge retrieval and, thus, about how firmly the knowledge elements are anchored in the cognitive structure of the learners (Unger et al. 2020).

When new knowledge elements are acquired, these elements must also be connected to existing nodes in the knowledge network. Similarly, existing connections between knowledge elements are strengthened when these knowledge elements need to be brought together to complete a particular task. Knowledge elements that are often activated together form activation patterns that represent conceptual understanding (Derry 1996). Network density is the proportion of observed relations in a network to the maximum number of possible relations. Several studies found positive relations between network density and achievement (Kubsch et al. 2019; Yang et al. 2018), as well as positive effects of instruction on network density (M. K. Kim and McCarthy 2021; Schwendimann 2014).

The connectedness of a network reflects the extent to which each pair of knowledge elements in the network is connected. Accordingly, it is a measure of “weak reachability”, indicating the extent to which the overall structure is either dense (high connectedness) or dispersed in tree-like structures or disjunct subnets (low connectedness). In the literature, connectedness is shown to be positively correlated with achievement (Freedman et al. 2024; Podschuweit and Bernholt 2020; Yang et al. 2018).

2.3.2 Local network metrics

Local network metrics take the location of the node and its connections to other nodes in the network into account; this provides indications of the relative prominence of each node in terms of being involved in many (direct or indirect) connections between nodes. Metrics such as degree centrality (the number of edges connected to a specific node) and betweenness centrality (indicating how frequently a particular node functions as a bridge on the shortest path connecting two other nodes) enable detailed comparisons among network components (Brandes and Erlebach 2005). For example, a concept such as chemical conversion should play a key role in a unit on reaction processes and reaction rates, as it is directly related to the phenomena under consideration and their explanation. Results from previous studies indicated that key nodes had higher degree centrality for higher-achieving students (Kubsch et al. 2019; Schwendimann 2014; Yang et al. 2018) and higher betweenness centrality for experts compared to novices (Wagner and Priemer 2023).

These metrics are well aligned with learning theories (e.g., KI, Section 2.2.; Cromley et al. 2024) and aim to trace students’ learning trajectories towards a well-connected knowledge base over the course of a teaching unit (Bransford et al. 2000). Consequently, analyzing the dynamics of concept application over time could obtain information that could serve as an indicator of learning.

3 Research objectives and hypotheses

Effective formative diagnostics are essential for targeted, individualized instruction. However, integrating meaningful science learning with valid assessment in digital environments remains challenging. This proof-of-concept study addressed this challenge by analyzing students’ conceptual understanding during a digitally supported chemistry unit, evaluating whether students’ responses can be reliably used to track individual learning progress. Specifically, we employed network analysis to examine how students apply and develop their understanding of chemical kinetics as they work through the teaching unit (Lossjew and Bernholt 2025). Almost all teaching and learning materials were implemented in a Moodle course that also served as a digital student workbook during the course. In a sample of $N = 300$ students in the 11th to 13th grade, the knowledge elements employed in students’ answers and artifacts were analyzed over time as indicators of students’ progress in mastering the knowledge and skills targeted during the teaching unit. The analyses employed network models based on these knowledge elements and network metrics to examine differences in students' mastery of the unit's content and to derive entry points for individualized scaffolding or instructional decisions by the teacher.

The design and procedure of the present study were driven by the following research questions:

1. With which accuracy can knowledge elements be automatically scored from student responses and artifacts produced over the course of a teaching unit?

Informed by research on learning progressions in students' understanding of chemical reactions, we identified the core ideas and teaching activities that promote the integration of new ideas into existing knowledge networks with regard to the topic of chemical kinetics. As knowledge elements are considered to be the constituting elements of a student's knowledge base (Bransford et al. 2000) and as using conceptual knowledge productively requires that a new problem or phenomenon is recognized as an instantiation of a particular concept (Goldwater and Schalk 2016), we expected the uptake and integration of knowledge elements to become visible in students' answers. The feasibility of automatically scoring the presence and quality of knowledge elements in students' written responses using transformer-based machine learning (ML) models has been demonstrated (Gombert et al. 2023). However, evidence is currently limited to specific topics in physics and the scoring of students' answers in knowledge tests. Accordingly, we planned to replicate the approach using a different data structure (i.e., a different domain and classroom-based artifacts), but we assumed that the automated scoring of knowledge elements would also be feasible over the course of a teaching unit.

2. To what extent are specific network metrics derived from students' knowledge networks over the course of the unit related to students' performance in a posttest administered after the teaching unit?

On the basis of the KI framework, we expected students to have well-developed knowledge networks organized around central ideas of a domain at the end of the unit that was designed to entail inquiry-based activities and driving questions to structure students' learning (Linn et al. 2015). The degree to which students were able to reach this goal was expected to be reflected in the students' knowledge networks, which were derived from the scored knowledge elements employed in students' answers to tasks across the unit. Regarding network metrics related to performance (Cromley et al. 2024; Morrison et al. 2022; see Section 2.4), substantial relations to students' performance in the posttest were expected, especially for density, degree centrality, and betweenness centrality (Asmussen et al. 2022; Cromley et al. 2024; Podschuweit and Bernholt 2020; Schwendimann and Linn 2016). With regard to the network-level metrics, we expected the network size (i.e., more nodes) to be related to students' posttest scores, provided the network nodes were also densely connected (Cromley et al. 2024; Schwendimann and Linn 2016). Likewise, the connectedness of students' knowledge networks was expected to be positively associated with learning outcomes, given that the KI framework emphasizes the importance of interconnectedness and linking information. With regard to the node-level metrics, we expected students with higher posttest scores to prominently include key concepts in their knowledge network; this would indicate associations between learning outcomes and both higher degree centrality and higher betweenness centrality (Asmussen et al. 2022; Kubsch et al. 2019).

In addressing these research questions, the overall aim was to derive indicators of students' learning progress over the course of the unit that could be used to inform diagnostic decisions and to select adaptive scaffolds.

4 Methods

4.1 Teaching unit

The teaching unit on the topic of chemical kinetics was developed as part of the interdisciplinary project ALICE (Analyzing Learning for Individualized Competence Development in Mathematics and Science Education). Based on the ECD4LPA framework (Kubsch et al. 2022) and taking educational standards into account (BIMI 2016; KMK 2004), as well as prior research on the topic (Bain and Towns 2016; Y.K. Yan and Subramaniam 2016) and on learning progressions in students' understanding of chemical reactions (Emden et al. 2018; F. Yan and Talanquer 2016), together with theoretical considerations regarding student learning (Bransford et al. 2000; Duschl et al. 2007; Linn et al. 2015), the teaching unit focused on project-based learning as the central pedagogy guiding the design process to promote students' science learning (B.L. Schneider et al. 2020). Accordingly, scientific inquiry activities were employed extensively to structure students' interaction with the concepts and knowledge elements considered relevant for the unit. Authentic tasks and different task formats (comprising both constructive response answers, e.g., short and long open-ended written responses, drawings; and closed formats, e.g., multiple

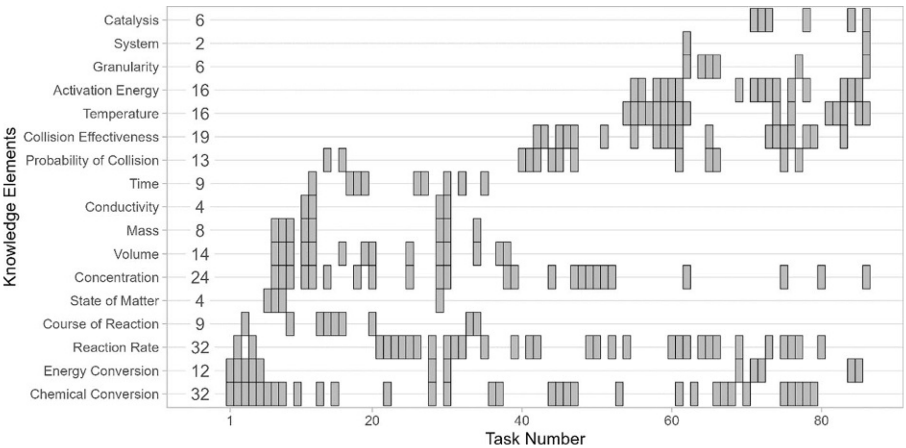

Fig. 1 Distribution of required knowledge elements across tasks. Numbers on the left indicate the sum of occurrences of each knowledge element across the whole unit

choice questions, ranking questions, drag-and-drop questions) required students to interpret and explain real-world phenomena by connecting scientific practices, core disciplinary ideas, and cross-cutting concepts. Moodle was used as the technical platform to provide learning materials and tasks to the students, but it also served as a digital workbook to collect students' answers and further artifacts, for example, drawings. An in-depth description of the unit can be found in (Lossjew and Bernholt 2025).

Extensive scoring guides were drafted to provide evidentiary statements for each task in the unit, along with detailed scoring guides for students' products (see electronic supplement). Overall, students' artifacts from 86 tasks were used for the analyses. Across these tasks, 17 knowledge elements were to be incorporated by students in their responses. The frequency and the total number of occurrences differed between knowledge elements, ranging from two to 32 (Fig. 1), thus requiring students to repeatedly activate and incorporate these knowledge elements across a range of tasks and phenomena. With regard to the storyline, the unit can be divided into two phases: The basic introduction and clarification of the concept of reaction speed in the first phase was followed in the second phase by a closer examination of the various influencing factors (e.g., temperature). The first phase therefore dealt with quantitative reaction dynamics (Tasks 1–35), while the second phase dealt with process control and prediction (Tasks 36–86).

4.2 Sample and data collection

The teaching unit on chemical kinetics was piloted in individual classes to ensure the feasibility of implementation and to identify and, if necessary, reduce problem areas during implementation. Afterwards, the teaching unit was implemented in regular classrooms by regular chemistry teachers over the course of about 10–12 weeks. Fifteen classes in the 11th to 13th grade ($N = 300$ students) and their teachers participated in the study. The teachers participated in a one-day professional development workshop prior to teaching the unit. The workshop focused on the content of the unit and on general strategies for teaching in digitally supported learning environments, especially with Moodle. In addition, the actual implementation of the teaching unit was addressed, and teachers were provided with an overview of the unit. Based on the findings from the piloting, critical phases were considered in detail to highlight phases considered relevant for students' progress during the unit and to illustrate the use of specific task formats (e.g., the use of the geometry applet GeoGebra in class).

After providing informed consent to participate in the study, students completed a pretest designed to assess their prior knowledge and their understanding of chemical reactions prior to the unit, in addition to measures of cognitive abilities, reading skills, and a questionnaire on demographic and person-related variables, such as interest and subject-related self-concept (The results from the pretest and the questionnaire are not presented as part of this manuscript as the posttest is considered the relevant measure to evaluate students' mastery of the content covered in the unit).

During the teaching of the unit, teachers were provided with technical and content-related support upon request. The implementation of the unit was monitored in

Moodle. Visits from the research team were only possible to a small number of classes due to the parallel implementation in many classes.

4.3 Scoring of student artifacts from the teaching unit by human raters

The implementation of the teaching unit in Moodle made it possible to collect students' responses and further artifacts (e.g., drawings) throughout the unit. The initial scoring of students' responses was completed by several (human) raters based on the predeveloped scoring guides. The raters were student assistants with a background in chemistry education. Apart from the information provided in the scoring guides, raters were unaware of the hypotheses or the analytical strategy of the study. To facilitate the scoring procedure, students' answers were exported from Moodle and imported into Inception (Klie et al. 2018). In Inception, scoring was done on a case-by-case basis, that is, the raters scored an individual student's responses to the sequence of tasks in a particular part of the unit to allow for references to previous responses. Knowledge elements were scored with 0 (not present or simple keyword use), 1 (basic integration), or 2 (comprehensive integration), (see electronic supplement). Overall, about 20% of the answers to each task were scored by at least two raters. Interrater reliabilities varied between the individual knowledge elements (Cohen's κ ε [0.55, 0.87]), with an average of $\kappa = 0.72$, indicating a reliable scoring process. Throughout the extensive process, interrater reliability was regularly assessed. Any significant discrepancies were resolved through collaborative discussion, ensuring the quality of the coding.

4.4 Automated scoring

On the basis of the data we scored in Inception, we trained and evaluated an ML model to replicate the respective scoring behavior. For this purpose, we expanded

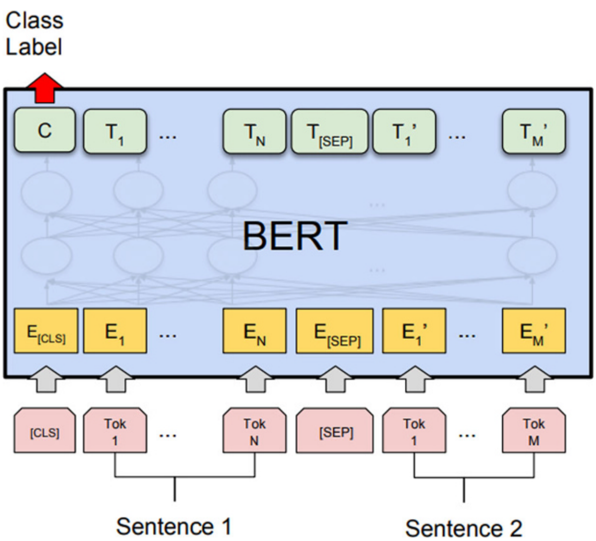

Fig. 2 A schematic illustration of BERT for sentence-pair classification. In our case, Sentence 1 is the name of the knowledge element, while Sentence 2 is the learner's response. Image taken from Devlin et al. (2019), who released it under CC 4.0

on the approach of Gombert et al. (2023), who implemented multiple ML models to solve a very similar problem of detecting knowledge elements within short German answers in the domain of energy didactics. Their strongest model used GBERT (Chan et al. 2020), a German-specific BERT variant, in a multitask learning setup to detect multiple knowledge elements in learners' answers in the form of multiple binary classification tasks. The authors modified the original BERT architecture (Devlin et al. 2019) that GBERT is based on to fit the task. Although their model demonstrated an overall strong performance, their custom modifications to the model architecture made their approach incompatible with a lot of standard tooling that has since become available for transformer language models such as BERT.

For this study, we wanted to focus on approaching the problem with a more standard-oriented approach to guarantee high replicability and allow for easy adoption in the field. Similar to Gombert et al. (2023), we used GBERT to tackle this task. We finetuned *GBERT-base* using the *Huggingface Transformers* framework and its *BertForSequenceClassification* class. To allow the model to deal with the overall range of learning knowledge elements, we prompted it with: 1) the individual label for which a score needed to be assigned, and 2) the corresponding learner's response. Both were separated by the model's separation token. The model then needed to predict a score between 0 and 2 signifying how well-integrated a given knowledge element was, as observed in the learner's response. Thus, on an abstract level, we solved the problem as a sentence-pair classification task, although one of the sentences consisted of just the label. Figure 2 illustrates this setup.

We evaluated this approach in a 3×3 cross-validation setup and we report micro-level F1 scores for all three possible scores. The micro-level F1 score (also known as Micro-Average F1) is a metric used to evaluate classification models in ML, particularly for multiclass or multilabel problems. It is used as an indicator of the overall accuracy of the classification. During each cross-validation step, the model was finetuned for three epochs. The results of this experiment can be found in the Results section.

4.5 Network analysis

Based on the knowledge elements expected in the responses to each task according to the predeveloped scoring guides, a normative network was established. Knowledge elements were considered nodes in the network; edges, that is, connections between nodes/knowledge elements, were established when nodes co-occurred in the same

Fig. 3 Evolution of the normative knowledge network at five time points over the course of the unit. Dots (nodes) represent knowledge elements, dashes (edges) represent connections between knowledge elements

task to reflect that a specific combination of knowledge elements had to be activated to address the demands of the task. Co-occurrence analysis was preferred because its grounding in associative memory models (Barabási et al. 1999; Derry 1996) allows it to capture the implicit, less conscious strength of association between concepts, indicated by their frequency and proximity in a text or discourse. This method reveals how tightly concepts are integrated in a student's memory (Unger et al. 2020), even if they struggle to explicitly articulate the formal relationships. In contrast, coded linkages (Cromley et al. 2024; Kubsch et al. 2019; H.-S. Lee and Liu 2010; Liu et al. 2008) primarily assess the explicit, qualitative, and hierarchical nature of knowledge (Novak 2002). This requires a more formal articulation of meaning that may not be present in the raw data being analyzed, due to students' known lack of commitment to formulating extensive scientific explanations (H.-S. Lee and Liu 2010), as well as their status as novice learners, which might limit the degree to which their knowledge network is organized according to a formal hierarchy, although this is usually expected from experts.

A growing longitudinal network was built up over the sequence of tasks in order to clarify the normative expectation of the knowledge base that should be used to answer the tasks over the course of the unit (Fig. 3).

On the basis of the scoring of knowledge elements in students' responses, individual student networks were established in a comparable way. Knowledge elements scored to be present in a response to a particular task (i.e., scored as 1 or 2; Section 4.3) were taken to establish nodes in the network, and the co-occurrence of multiple knowledge elements in the same response was taken to establish connections between nodes. Over the course of the tasks in the unit, growing longitudinal student networks thus reflected the acquisition and repeated enactment of knowledge elements by each student over the course of the unit.

In the comparison of both types of networks, the normative network can be considered the upper ceiling in the form of an ideal performance, that is, incorporating every normatively expected knowledge element into each task across the whole unit, while the student networks reflect the extent to which individual students demonstrate their capabilities of incorporating the relevant knowledge elements over time. It should be emphasized here that the normative scoring guides were formulated very inclusively in order to cover as many conceivable knowledge elements as possible, from a content perspective. Not all knowledge elements were always necessary and/or of equal importance for a correct answer to a task. Accordingly, "completely correct" solutions based on these normative principles are rather unlikely. Because the scoring of knowledge elements and the networks derived from these scores can be analyzed on a task-by-task basis, they provide an approach that can be used to analyze student learning (as reflected in these networks) on a fine-grained basis throughout the unit.

Based on both the normative and the individual student networks, specific network metrics were derived to allow for comparative analyses of specific network components based on local metrics and analyses between different networks using global metrics. Numerous network metrics can be used for such an analysis (Luke 2015; Morrison et al. 2022); we selected metrics that were shown to correlate with achievement measures in prior studies (Cromley et al. 2024; Daems et al. 2014;

Podschuweit and Bernholt 2020). Relative network size, network density, and connectedness were chosen as global network-level metrics; degree and betweenness centrality were chosen as local node-level metrics for more fine-grained analyses (A detailed description of these metrics is provided in the electronic supplement).

All of these metrics were calculated from the students' knowledge networks along the task sequence. To bridge the gap between the continuous process data (represented by the longitudinal knowledge networks) and the point measure of students' posttest performance, three aggregate measures were calculated for each network metric across the sequence of tasks. These measures capture different facets of the knowledge acquisition process:

First, the *Area Under the Curve* (AUC) was calculated to represent the cumulative complexity and persistence of KI over the course of the unit (Pruessner et al. 2003). While a single-point measure only captures the status quo at a specific time, the AUC integrates the magnitude of a network metric over the entire sequence of tasks. A high AUC indicates that a student consistently maintained a large or highly integrated knowledge network across most tasks.

Second, *Variance* was used as an indicator of cognitive variability (or stability), which often precedes conceptual shifts in learning (Siegler 1994, 2007). It quantifies how much a student's network metrics fluctuated from one task to the next. High variance suggests a "dynamic" or perhaps "unstable" trajectory, where the complexity of the student's answers varied significantly depending on the task demand. High variability can also indicate a transitional phase. For example, when students fluctuate between old and new strategies or elements of knowledge, the variance increases before consolidation occurs (Siegler 2007).

Third, *First-Order Autocorrelation* (ACF1) was computed to assess the "inertia" or temporal stability of the network structures. In the context of dynamic systems, a higher ACF1 indicates a greater dependence on previous states, reflecting the degree of continuity in the student's knowledge construction (D'Mello and Graesser 2012; Wijnants 2014). In a learning context, this can represent a stable, incremental growth.

As the unit was divided into two phases (basic introduction and clarification of the concept of reaction speed; influenceability of reaction speed by various factors, e.g., temperature), the aggregated measures were calculated separately for the first (Tasks 1–35) and second phase (Tasks 36–86).

The network analysis was performed using R (Version 4.4.2; R Core Team 2024) and specific packages (tidyverse 2.0.0, Wickham et al. 2019; network 1.18.2, Butts 2008; ndtv 0.13.4, Bender-deMoll 2024; tsna 0.3.5, Bender-deMoll and Morris 2021). For the different network metrics, linear regression models were used to evaluate the relations between the specific metrics and students' posttest scores.

4.6 Measures

To measure students' competence at the end of the unit, a written test was developed that focused on chemical kinetics but also on students' more general understanding of the concept of chemical reactions. For the test, items were partly taken from prior studies (e.g., Y. K. Yan and Subramaniam 2016) or were newly developed. The

Table 1 Evaluation results of the scoring model

Overall Micro F1	Micro F1 Score 0	Micro F1 Score 1	Micro F1 Score 2
89.85	91.25	19.94	90.55

test comprised mostly open-ended task formats to allow students to express their understanding of particular aspects, but it also comprised closed formats to assess specific knowledge elements.

The open-ended tasks of the test were scored by human raters on the basis of a previously created scoring manual, which contained the knowledge elements required to demonstrate the sample solution. Each knowledge element was scored as present (1) or missing (0). To ensure reliable and objective scoring, approximately 20% of the data material was double-scored. The scoring of the knowledge elements ($\kappa = 0.91$) showed high interrater agreement. The total score for each task was calculated as the sum of all expected knowledge elements. On the basis of these overall scores, the test was scaled using a generalized partial credit model to determine person abilities. The reliability of the test can be considered good (WLE reliability: 0.85; Infit: $0.93 < \text{MNSQ} < 1.04$).

For illustrative purposes, students were categorized into four groups according to their posttest score. To obtain different groups in terms of achievement, students with a posttest score that was at least one standard deviation (SD) below the mean posttest score of the entire sample were classified as "very low" (performance), students with a posttest score within 1 SD below the overall mean were classified as "low", between 1 SD above the overall mean as "good", and above 1 SD above the overall mean as "very good".

5 Results

5.1 Automated scoring of knowledge elements

Scoring students' responses based on *GBERT-base* resulted, overall, in high human–AI interrater agreements and reliabilities, namely, an F1 score of 89.85. Table 1 shows the global results as well as the individual levels of scoring the knowledge elements. F1 scores were high for the overall scoring procedure (overall micro F1), as well as for whether knowledge elements were not present/were only present by keyword use (Score 0) and for the comprehensive integration of knowledge elements (Score 2). However, the Micro F1 score was substantially lower for scoring a basic integration of knowledge elements (Score 1).

5.2 Network-level metrics

5.2.1 Relative size

The relative size of students' networks (i.e., the degree to which students incorporated all required knowledge elements into a given task) varied substantially across

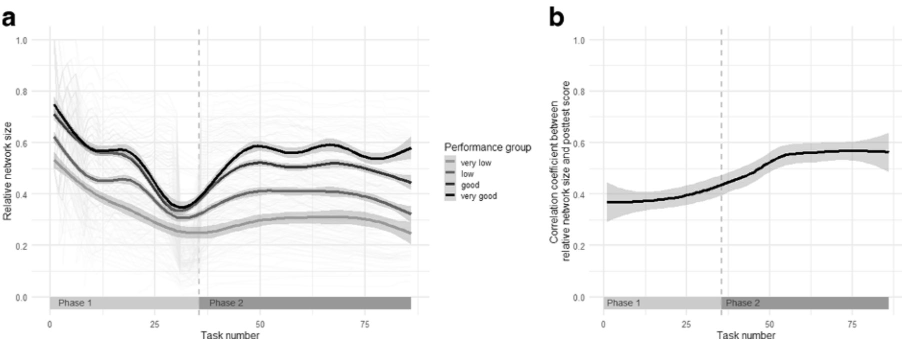

Fig. 4 Distribution of students' relative network size across the whole unit and, for illustrative purposes, divided by the four performance groups (**a**) and correlation coefficients between students' relative network size and posttest score (**b**). The black line indicates the moving average, the gray area its standard error; individual trajectories are displayed in light gray (only in **a**)

the unit (Fig. 4a). For illustrative purposes, the sample was divided into four performance groups. On this basis, a consistently stable ranking emerged, with higher-performing students tending to have larger networks. While the mean relative network size for the first ($M=0.45$, $SD=0.14$, Min = 0.01, Max = 0.79) and second ($M=0.41$, $SD=0.16$, Min = 0.05, Max = 0.81) phase seemed to be comparable, the mean values decreased in the first phase, before appearing to be rather stable in the second phase of the unit (after a short recovery period).

The correlation between the relative size of students' networks and the posttest score was low in the first phase of the unit ($M=0.39$, Min = 0.27, Max = 0.73) and increased substantially in the second phase ($M=0.53$, Min = 0.30, Max = 0.68; Fig. 4b).

The results of a linear regression model used to predict students' posttest scores by the aggregated measures of students' relative network size across the unit (see Section 4.6) revealed that the variance of the relative network size in Phase 1 ($b=9.73$, 95% CI [4.65, 14.81], $t[197]=3.78$, $p<0.001$; $\beta=0.23$, 95% CI [0.11, 0.36]) as well as the AUC ($b=0.04$, 95% CI [0.01, 0.06], $t[197]=2.96$, $p=0.003$; $\beta=0.31$, 95% CI [0.10, 0.52]), the variance ($b=-51.40$, 95% CI [–99.37, –3.43], $t[197]=-2.11$, $p=0.036$; $\beta=-0.12$, 95% CI [–0.23, –7.99e-03]), and the ACF1($b=1.05$, 95% CI [0.40, 1.71], $t[197]=3.17$, $p=0.002$; $\beta=0.24$, 95% CI [0.09, 0.38]) in Phase 2 showed significant effects on students' posttest scores ($F[6, 197]=25.55$, $p<0.001$, $R^2=0.44$, *adj.* $R^2=0.42$). Notably, the variance of the relative network size showed a positive effect in Phase 1 and a negative effect in Phase 2 of the unit.

5.2.2 Density

The density of students' networks (i.e., the proportion of observed connections between knowledge elements to the number of possible connections) seemed to be rather stable across the unit (Fig. 5a). The mean density of students' networks in the first ($M=0.38$, $SD=0.15$, Min = 0.00, Max = 0.78) and second ($M=0.35$, $SD=0.13$, Min = 0.02, Max = 0.69) phase was almost identical. Also, the mean values appeared to be rather stable over the course of the unit.

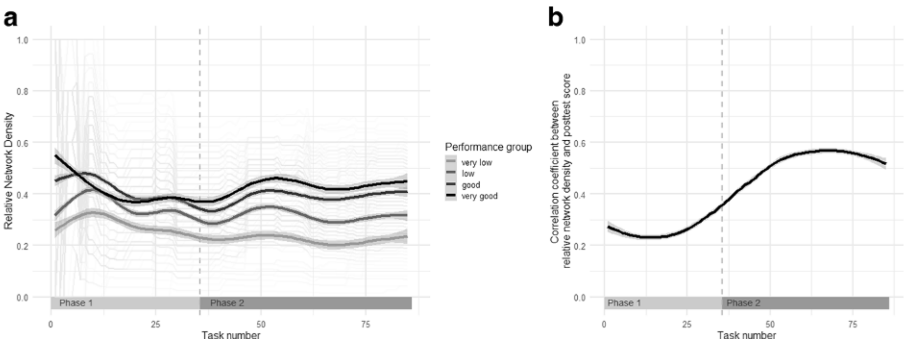

Fig. 5 Distribution of the density of students' networks across the whole unit and, for illustrative purposes, divided by the four performance groups (**a**) and correlation coefficients between the density of students' networks and posttest scores (**b**). The black line indicates the moving average, the gray area its standard error; individual trajectories are displayed in light gray (only in **a**)

The correlation between the density of students' networks and the posttest score was low in the first phase of the unit ($M = 0.26$, Min = 0.16, Max = 0.36) and increased substantially in the second phase ($M = 0.52$, Min = 0.35, Max = 0.57; Fig. 5b).

The results of a linear regression model used to predict students' posttest scores by aggregated measures of student networks' density across the unit (see Section 4.6) revealed that the AUC in both phases (Phase 1: $b = -0.09$, 95% CI [–0.14, –0.05], $t[193] = -4.07$, $p < 0.001$; $\beta = -0.46$, 95% CI [–0.69, –0.24]; Phase 2: $b = 0.15$, 95% CI [0.11, 0.18], $t[193] = 8.85$, $p < 0.001$; $\beta = 0.90$, 95% CI [0.70, 1.10]) and the variance in the first phase ($b = 26.25$, 95% CI [12.97, 39.53], $t[193] = 3.90$, $p < 0.001$; $\beta = 0.25$, 95% CI [0.13, 0.38]) showed significant effects on students' posttest scores ($F[8, 193] = 16.13$, $p < 0.001$, $R^2 = 0.40$, *adj.* $R^2 = 0.38$). Here, the AUC of the network density showed a negative effect in Phase 1 and a positive effect in Phase 2 of the unit.

5.2.3 Connectedness

The connectedness of students' networks (i.e., the degree to which students incorporated all required knowledge elements in a given task) increased from the first ($M = 0.43$, $SD = 0.19$, Min = 0.00, Max = 0.84) to the second phase of the unit ($M = 0.55$, $SD = 0.19$, Min = 0.01, Max = 0.81; Fig. 6a). The patterns across the three higher-performing groups were rather similar, while the group of students with the lowest performance in the posttest showed a rather low level of connectedness throughout the unit.

The correlation between the connectedness of students' networks and the posttest score was low in the first phase of the unit ($M = 0.27$, Min = 0.17, Max = 0.36) and increased substantially in the second phase ($M = 0.48$, Min = 0.36, Max = 0.54; Fig. 6b).

The results of a linear regression model used to predict students' posttest scores by aggregated measures of the connectedness of students' networks across the unit (see Section 4.6) revealed that the variance in Phase 1 ($b = 11.59$, 95% CI [1.67, 21.50], $t[193] = 2.30$, $p = 0.022$; $\beta = 0.17$, 95% CI [0.03, 0.32]) as well as the AUC

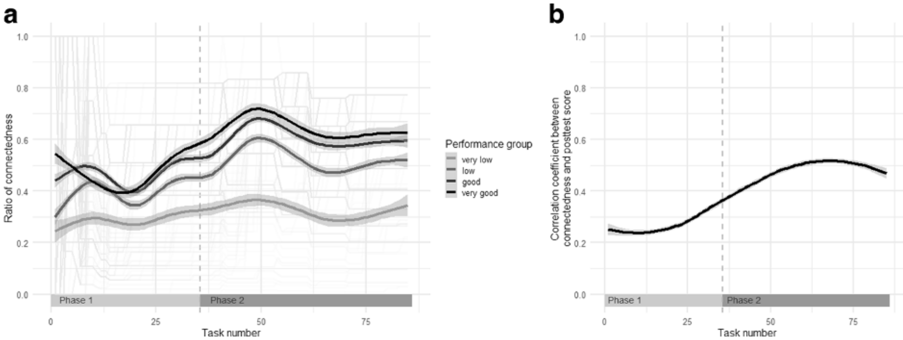

Fig. 6 Distribution of the ratio of connectedness across the whole unit and, for illustrative purposes, divided by the four performance groups (**a**) and correlation coefficients between connectedness and posttest score (**b**). The black line indicates the moving average, the gray area its standard error; individual trajectories are displayed in light gray (only in **a**)

in Phase 2 ($b = 0.06$, 95% CI [0.03, 0.09], $t[193] = 4.64$, $p < 0.001$; $\beta = 0.55$, 95% CI [0.31, 0.78]) showed significant effects on students' posttest scores ($F[8, 193] = 10.96$, $p < 0.001$, $R^2 = 0.31$, *adj.* $R^2 = 0.28$).

5.2.4 Combined analyses

After the aggregated measures of all three network-level metrics were integrated into a combined linear regression model to predict students' posttest scores, the model explained a statistically significant and substantial proportion of variance ($F[18, 183] = 10.73$, $p < 0.001$, $R^2 = 0.51$, *adj.* $R^2 = 0.47$). With regard to the relative network size, both the AUC ($b = 0.05$, 95% CI [1.86e-04, 0.10], $t[183] = 1.98$, $p =$

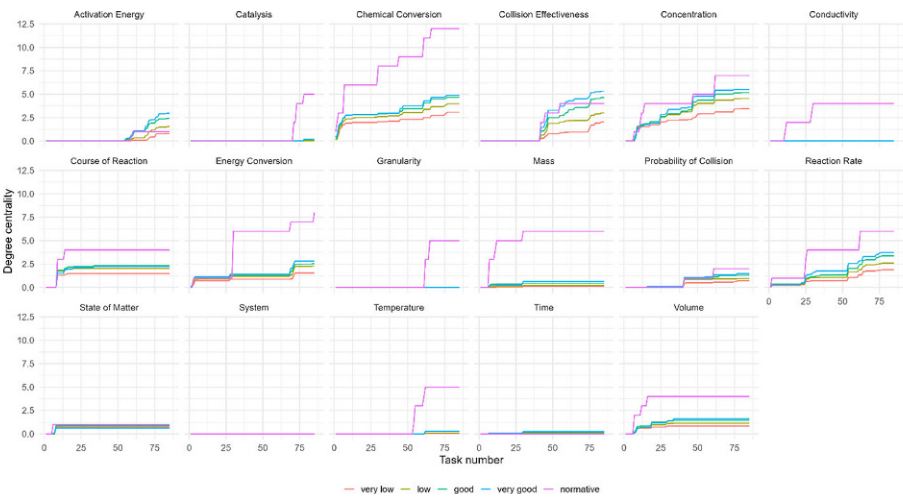

Fig. 7 Distribution of the degree centrality per knowledge element across the whole unit and divided by the normative network (purple) and the four performance groups

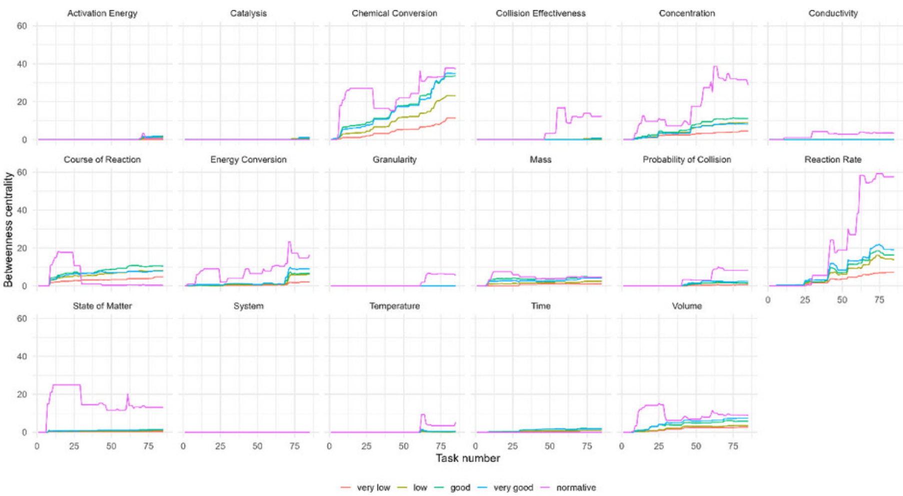

Fig. 8 Distribution of the betweenness centrality per knowledge element across the whole unit and divided by the normative network (purple) and the four performance groups

0.049; β = 0.22, 95% CI [8.62e-04, 0.45]) and the variance in the first phase (b = 9.10, 95% CI [2.78, 15.43], t[183] = 2.84, p = 0.005; β = 0.22, 95% CI [0.07, 0.37]) showed significant effects on students' posttest scores. The effect of the AUC for density was significant in both phases of the unit, but the effect occurred in opposite directions in the two different phases. In the first phase, the AUC for density was negatively related to the posttest score (b = –0.09, 95% CI [–0.18, –1.22e-04], t[183] = –1.98, p = 0.050; β = –0.43, 95% CI [–0.86, –5.93e-04]). In the second phase, the AUC for density had a positive effect on students' posttest scores (b = 0.10, 95% CI [0.03, 0.17], t[183] = 2.77, p = 0.006; β = 0.61, 95% CI [0.18, 1.04]), with a larger effect size than in the first phase.

5.3 Node-level metrics

5.3.1 Degree centrality

Figure 7 illustrates the degree centrality of all knowledge elements over the course of the teaching unit (i.e., the number of edges connected to a specific node). While some knowledge elements showed a low degree centrality throughout the unit (e.g., system and time), some knowledge elements stood out by showing high values for degree centrality in the normative network, for example, chemical conversion, concentration, and energy conversion. In the student networks, the mean values for degree centrality are ordered from lower to higher values in line with the four performance groups for all knowledge elements. For some knowledge elements (e.g., concentration or collision effectiveness), the values for degree centrality in the student networks were also close to the values in the normative network, at least for the high-performing student groups. For other knowledge elements (chemical conversion, energy conversion, or mass), students hardly managed to connect these

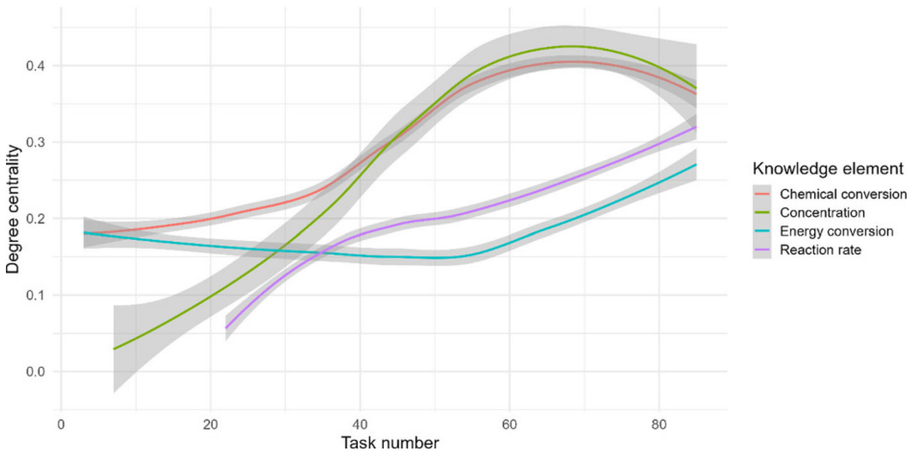

Fig. 9 Moving average of the correlation coefficient between the degree centrality of central knowledge elements and students' posttest scores across the unit. The thick line indicates the moving average, the gray area the standard error

knowledge elements to other knowledge elements in their networks, as normatively expected, indicating that these students struggled to integrate these knowledge elements into their knowledge network.

5.3.2 Betweenness centrality

Across the list of knowledge elements, some stood out by showing high values for betweenness centrality in the normative network, for example, chemical conversion, concentration, and reaction rate (Fig. 8). As in the case for degree centrality, some knowledge elements (e.g., chemical conversion) seemed to actually fulfill their bridging function, also in the students' networks, while other knowledge elements (e.g., concentration or reaction rate) were not integrated into students' networks in a way that reflected such a functional role. Also, students in the higher-performing groups generally showed higher values (i.e., closer to the normative expectation) for betweenness centrality compared to students in the lower-performing groups.

5.3.3 Central knowledge elements

Examining the findings for degree and betweenness centrality, it is clear that four knowledge elements (chemical conversion, concentration, energy conversion, and reaction rate) seemed to have a rather outstanding role. These knowledge elements stood out in that they occurred more frequently, showed higher betweenness, and showed higher degree centrality. In this context, the question arose as to when these knowledge elements become central in such a way that specific network metrics already indicate a substantial relationship between them and students' posttest scores early in the course of the unit. In the comparison of the correlation between the degree centrality of the most central knowledge elements and students' posttest

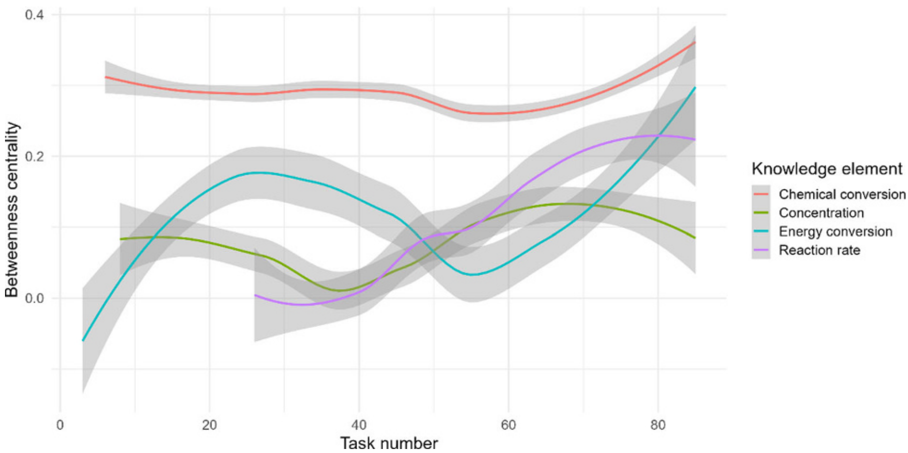

Fig. 10 Moving average of the correlation coefficient between the betweenness centrality of central knowledge elements and students' posttest scores across the unit. The colored lines indicate the moving average for each knowledge element, the gray area the standard error

scores on a task-by-task basis, different patterns emerged for the different concepts (Fig. 9). While all correlation coefficients increased over time, only the correlation coefficients for chemical conversion and concentration rose above a value of $r = 0.4$.

Betweenness centrality indicates a bridge function for specific knowledge elements as these nodes provide short connections between many knowledge elements. The correlation between the betweenness centrality of the most central knowledge elements and students' posttest scores indicated a rather stable pattern across knowledge elements (Fig. 10). Here, the correlation coefficient for chemical conversion remained almost constant across the unit at a (comparably) high value between 0.23 and 0.44, while the coefficients for the other central knowledge elements fluctuated over the course of the unit.

6 Discussion

The present study aimed to identify indicators of students' learning progress in a teaching unit on chemical kinetics that was taught in a digitally supported format. The teaching unit was developed based on a systematic framework that aims to integrate the affordances of coherent and engaging instruction with reliable assessments. Students' responses to a range of different tasks were automatically scored throughout the unit to derive individual knowledge networks that reflected how students enacted and connected specific knowledge elements across the tasks of the unit.

6.1 Automated scoring of knowledge elements

Scoring students' responses both by human raters and automatically, based on *GBERT-base*, resulted, overall, in high interrater agreements and reliabilities, namely, an F1 score of 89.85. Although the feasibility of the automated scoring of knowledge elements in (short) student answers has already been demonstrated in previous studies (Gombert et al. 2023), the present data set entailed the specific challenge that numerous task formats were included across the unit. Nevertheless, the results indicate that it is feasible to score knowledge elements in real time across a full unit, not just in focused lessons (Tissenbaum and Slotta 2019), specific learning environments (Gobert et al. 2013), or task formats (H.-S. Lee et al. 2019).

What becomes apparent from these results is that the model is able to determine well whether a knowledge element is fully contained or fully missing. However, compared to humans, the model struggles to determine whether a knowledge element is only partially contained in a response (as indicated by the low Micro F1 for Score 1). This is an inherent weakness of this approach, which might be explained by the comparably large variance in what a partial containment of a knowledge element might look like and the comparably low number of respective examples in our data set.

However, these observations also offer valuable insights for future applications. Specifically, the description of the intermediate level appears to require greater refinement to ensure clearer distinctions in human scoring (that have a direct influence on the training data set), particularly between the absence of and the comprehensive use of knowledge elements. In this context, the use of specific linguistic examples shows great promise, as they could help identify and classify linguistic features that indicate a basic understanding even when a knowledge element is not explicitly mentioned. Additionally, examining misclassifications could prove crucial, especially to determine whether they reveal a systematic tendency, such as a consistent underestimation or overestimation of learners' performance. This understanding is crucial, as supportive strategies need to differ depending on whether there is a tendency toward under- or overestimation.

6.2 Results of network-level metrics

Based on the scoring of the knowledge elements, normative and student networks were established on a task-by-task basis in order to analyze student learning (as reflected in these networks) on a fine-grained basis throughout the unit. The selection of network- and node-level metrics provided different insights into students' learning over time; interesting insights were also obtained when analyzing the relation of these metrics to students' posttest results.

The extent to which students incorporated new knowledge elements into their existing knowledge network was analyzed using the network-level metric. Over the course of the unit, values for this metric showed substantial fluctuations. In the first tasks, the values were rather high (Freedman et al. 2024; K. Kim 2025). This might be expectable at the beginning of a teaching unit when the focus is on reactivating content from lower secondary classes, as well as on introducing students to the new topic. The values then decreased further to a local minimum at the end of the first phase of the unit (around Task 30). These tasks seemed to challenge all students in the sample, regardless of their performance in the posttest. Many new knowledge elements were introduced here, some of which appeared to present a particular challenge. Such a drop across all students could indicate a critical phase in a lesson, where monitoring student performance and providing support could be central for teachers. While students in the very low-performing group seemed to stay on their level of covering required knowledge elements in the second phase of the unit (from Task 40 onwards), students in the higher-achieving groups seemed to increase their coverage of knowledge elements in this second phase of the unit and, to a greater extent, along the three performance groups (Freedman et al. 2024). On average, however, the ratio did not exceed 60%, suggesting that the better-performing students were better able to meet the normative expectation of the tasks, but their performance was still far from the highest possible performance.

On the basis of the regression models used to predict the posttest scores, it was found that the variance in network size in both phases and the AUC and ACF1 in the second phase showed significant effects. Students whose network size fluctuated more in the first phase achieved a better posttest score. This suggests that the ability or willingness to adapt networks exploratively was beneficial for learning success (Siegler 1994). In the second phase, however, students with a relatively stable network size (low variance) achieved higher posttest scores. This finding was supported by the positive effects of the AUC and ACF1. Here, a larger cumulative number of knowledge elements (i.e., a larger network size over a longer period of time) and high positive ACF1 (i.e., high inertia of the network size over time) were associated with a better posttest score (Freedman et al. 2024; Schwendimann 2014). This suggests that, in the second phase, the established structure was applied consistently to subsequent tasks (D'Mello and Graesser 2012). Students who attained a larger network size earlier might have benefited from this mastery in the posttest (Pruessner et al. 2003).

The analysis of the network density revealed greater consistency in values over the course of the teaching unit (Kubsch et al. 2019; Yang et al. 2018). Nevertheless, there was a higher correlation between density and posttest scores in the second phase of the unit than in the first phase, suggesting significant changes at the individual level. This showed that students whose networks had a higher density, at least in the second phase, performed better in the posttest (M. K. Kim and McCarthy 2021; Schwendimann 2014).

In the analysis of the regression models, opposing effects became apparent in the two phases. The AUC, that is, the cumulative density of the student networks, had a negative effect in the first phase of the unit and a positive effect in the second phase. In addition, the variance of students' network density in the first phase of the unit had a positive effect on the posttest score. Thus, students who had a higher average density (more connections per node) in their knowledge network during the first phase of the teaching unit achieved lower posttest scores (Cromley et al. 2024). A higher density in the first phase of the unit could indicate that some students established connections between knowledge elements without critically evaluating or weighing these connections (H.-S. Lee and Liu 2010). This quickly makes the network "dense" but not necessarily structured (Liu et al. 2008). Weak or irrelevant connections could subsequently make it more difficult to recall and apply the knowledge in the posttest. At this point, the higher-performing students seemed to be more successful in improving their knowledge network through integration (qualitative growth) rather than creating new connections in an unfocused manner (quantitative growth; Liu et al. 2008; Schwendimann 2014).

The variance in density measures how much the density within a student fluctuated over the first phase of the unit. A positive effect means that students whose network density fluctuated more (stronger phases of densification and de-densification) achieved better posttest scores. Although a constantly dense network (a high AUC) seemed to be detrimental in the first phase of the unit, the ability to change density appeared to be beneficial. This suggests that successful learning consisted of actively manipulating the connections in the network: densification during phases of information absorption; de-densification (by deleting or weighting irrelevant edges) during phases of structuring and focusing (Goldwater and Schalk 2016; Koedinger et al. 2012). This higher quality of structure gave these students an advantage in the second phase of the lesson, as they were clearly able to establish key connections to the new knowledge elements (as indicated by the larger network size in these students), significantly expanding their knowledge network in the second phase and substantially increasing its density (Cromley et al. 2024; Linn et al. 2015; Schwendimann 2014).

Similarly, in terms of connectedness, the variance in the first phase and the AUC value in the second phase had a positive impact on the posttest results. Connectedness indicates how connected the knowledge elements of the network are overall (directly or indirectly via paths), and the variance in connectedness measures how much the connectedness within a student fluctuated over time. With regard to the course of the connectedness of students' networks (Fig. 6), the patterns across the three higher-performing groups were rather similar, with phases of decreasing and increasing values in the first and second half of the first phase of the teaching unit, respectively. As in the case of density, this suggests that successful learning consisted of actively manipulating the connections in the network (Siegler 2007). In this first phase, it is likely that it was not the quantity of the connections (density) but rather the quality of the structure or the strategic focus that predicted learning success. This was achieved by the higher-performing students by constantly building up and reviewing connections between knowledge elements, leading to a local maximum in the second phase of the unit (at about Task 50). In the second phase, students with a higher connectedness in their knowledge networks seemed to have an advantage in the posttest, suggesting that the connectedness reflects the quality of the structure of students' networks (Freedman et al. 2024; Podschuweit and Bernholt 2020; Yang et al. 2018). Accordingly, the group of students with the lowest performance in the posttest showed a rather low level of connectedness throughout the unit.

In the combined analysis, which included the different measures of the three network metrics, significant effects on students' posttest scores were found for the AUC of students' network size (positive) and density (negative), as well as for the variance in the network size in the first phase. In addition, the AUC of students' network density in the second phase had a positive effect on their posttest scores. In summary, three out of four effects were related to metrics that indicate the level and quantity (AUC) and, thus, the overall extent to which students enacted knowledge elements while working on the tasks (Pruessner et al. 2003). The fourth measure (variance) relates more to the volatility and adaptability of the use of knowledge elements with which students responded to different requirements during the course of the lesson (Siegler 2007). The different patterns of findings and the sometimes contradictory effects of the same metrics in the two phases of the teaching unit underscore the fact that both are apparently needed to successfully complete a learning unit (D'Mello and Graesser 2012). Consequently, combining different metrics might provide added value when assessing students' learning progress.

6.3 Results of node-level metrics

The results for the degree centrality clearly show that the more frequently occurring knowledge elements were also more strongly connected and thus occupied more prominent positions in the (normative) network (Asmussen et al. 2022; Kubsch et al. 2019; Yang et al. 2018). However, this role of specific knowledge elements was only partially reflected in the student networks. For some knowledge elements (e.g., concentration), it seemed to be easier for students to integrate these elements into their knowledge networks to the required extent and to include them in their answers in combination with various other knowledge elements. For other knowledge elements (chemical conversion, energy conversion), students found this activation and linking much more difficult. These "more difficult" knowledge elements represent more abstract concepts (Asmussen et al. 2022). From a normative point of view, however, it would be much more desirable for these concepts to occupy prominent and highly integrated positions in the students' network, because, as core concepts, they (should) represent structuring elements (Barabási et al. 1999). In a comparison of the knowledge elements, however, these core concepts did not achieve a prominent position in the students' networks. Rather, several knowledge elements showed comparable values of degree centrality, which (with one exception) were all below the values in the normative network. Students in the higher ability group approached the values of the normative network.

When the degree centrality of the four most central knowledge elements was examined in connection with the posttest score, it was observed that the correlation for all four knowledge elements was rather low at the beginning of the unit ($r < 0.2$). Only in the second half of the unit did the correlation coefficients increase significantly, especially for the two concepts of chemical and energy conversion. Here, values of $r \sim 0.4$ were reached in the last third of the unit, while the correlation coefficients for the other two knowledge elements remained below 0.3. This underscores the prominent role of these core concepts but also the fact that the degree centrality becomes a predictive measure of students' posttest performance only at the end of the lesson, that is, as the unit progresses. One potential explanation for this finding might be that, as the unit progresses, an increasing number of knowledge elements are added, necessitating the organization of these elements around the central concepts. Students who demonstrate an ability to cope with this continuous demand will also exhibit superior performance in the posttest. Conversely, degree centrality appears to be rather unsuitable as an indicator of an "early warning system". Nevertheless, it could allow teachers to take more individualized measures before exams are administered in the later (but still ongoing) course of the unit.

The betweenness centrality results should provide insights into the extent to which certain knowledge elements play a bridging role in the network, controlling the co-activation of concepts that are not directly related (Asmussen et al. 2022; Cromley et al. 2024; Kubsch et al. 2019). Chemical conversion played a central role in the normative network (as with degree centrality), as did reaction rate and concentration. Again, this emphasized role of certain knowledge elements was only reflected to a very limited extent in the student networks, with chemical conversion being the element that was reflected most. In terms of correlations with posttest scores, this knowledge element also consistently showed the highest correlation coefficients (between 0.3 and 0.4), while the coefficients for the other central knowledge elements were lower (between 0 and 0.3) and fluctuated significantly over the course of the unit. Correspondingly, the chemical conversion knowledge element, as a core concept, seemed to play a special role in terms of a bridging function. High values of betweenness centrality in the students' networks at an early stage and over the course of the unit were associated with higher values in the posttest. Accordingly, the inclusion of this knowledge element by students should be given constant attention when evaluating task completion. Also, betweenness centrality for the most central concepts of a teaching unit might be a suitable indicator to monitor students' performance even in early phases of learning. Thus, compared to degree centrality, betweenness centrality appears to be more suitable as an "early warning system".

7 Limitations

The limitations concern the database and the selection of the metrics used in this study. The database comprised students' responses to a sequence of tasks from a single teaching unit. The generalizability of the results is limited to the extent that no statement can be made about other topics (in the same or other grades) or other subjects. However, the automated scoring of students' responses has already been successfully applied to various target groups in previous research, and network analysis has also been used extensively to analyze students' knowledge levels. Nevertheless, other patterns of findings may emerge for other subjects or student populations.

The selection of network metrics is necessarily selective due to the sheer number of possible metrics (Morrison et al. 2022). In the analysis presented here, metrics were selected for which correlations with performance tests had been identified in previous studies. However, across the different metrics, very different patterns of results emerged, especially over the course of the unit. It is plausible to assume that additional and different metrics would lead to complementary patterns and, accordingly, different conclusions could be drawn from them. Further studies should therefore continue to systematically investigate which aspects of students' learning progress can be most validly captured by which metric. In addition, it should be noted here that the interrelations between the student networks and the normative network should be evaluated with caution, as the scoring guides were created very inclusively (Section 4.4.). While the expected interrelations between posttest and in-unit performance were found, the size of the correlation coefficients appeared to remain relatively low. The student networks refer to the level of connections between knowledge elements; however, not all inclusively formulated knowledge elements (analytical scoring) need to be included for the successful overall assessment of a task (holistic assessment). Accordingly, it is necessary to find a suitable balance between necessary feedback measures (e.g., for core concepts, see local metrics) and nice-to-have measures derived from these analyses.

In the present study, connections between knowledge elements were established when concepts co-occurred in students' answers to the same task. While there are both theoretical (Barabási et al. 1999; Derry 1996) and empirical (Unger et al. 2020) reasons for this approach, an explicit coding of the connections would provide information about the qualitative nature of these connections and, thus, about the hierarchical nature of students' knowledge structures (Cromley et al. 2024; Kubsch et al. 2019; H.-S. Lee and Liu 2010; Liu et al. 2008). A strength of the present study is that it analyzed the (co-)occurrence of knowledge elements across an extensive period of time and across many instances, tasks, and phenomena; further studies could investigate the development of coded connections over time and examine whether this information provides additional insights into students' learning progress.

The fact that this study was a descriptive study is a further limitation. There were no manipulations between groups of students, and no interventions were implemented over the course of the unit. In this respect, no statement can be made about the effectiveness of the proposed interventions in supporting students' learning development. Further studies are needed to examine whether a data-driven adaptation of the lessons leads to the hypothesized effects regarding the change in network metrics and the learning progress of the students.

8 Conclusions

The present study investigated the feasibility of a real-time derivation of knowledge networks from students' responses in an extensive digitally supported teaching unit on chemical kinetics, as well as the predictive power of the resulting network metrics for learning success. The results have important theoretical and practical conclusions.

8.1 Different findings for the two phases of the learning unit

Analyses at the network and node levels demonstrated that the learning process is dynamic and phase-dependent, with learning success critically relying on a strategic balance between the stability (quality) and the flexibility (quantity) of the knowledge network. The contrasting effects of the same metrics across the first and second phases of the teaching unit underscore the necessity of both stability and adaptability for successful learning. For the first phase, successful learning seems to be characterized by exploration and construction. Higher posttest scores were associated with flexibility and volatility. This was evidenced by the positive variance in both network size and density. Successful learners demonstrated the ability or willingness to adapt networks exploratively and to actively manipulate connections (densification and de-densification). Conversely, a consistently high density (AUC for density) during this early phase was detrimental, suggesting inefficient, unstructured oversaturation or the establishment of weak or irrelevant connections.

For the second phase, successful learning seems to be characterized by consolidation and application. Success in this phase hinged upon stability and consistency. A stable network size (low variance) and a larger cumulative network size (AUC) were predictive of the posttest score. This was supported by the high positive ACF1 of network size, suggesting that the established structure was consistently applied to subsequent tasks.

The contrasting findings for the different phases mirror the structure of the unit, which aimed to first establish the basics, which were then linked and deepened through application tasks. Students' difficulty with the transition from the exploration phase to the consolidation phase may indicate that they have not yet established a large and coherent knowledge network at the end of the first phase. For these students, more targeted repetition might be needed, or the introduction of further new knowledge elements should be delayed so as not to (over)challenge students. This could support learning in the zone of proximal development according to Vygotskij (1981), which—in turn—could promote the experience of competence and motivation to learn, as well as positive performance development (Bureau et al. 2022; Linnenbrink-Garcia et al. 2016).

8.2 The role of core concepts

The node-level analysis highlighted the prominent role of structuring concepts such as chemical conversion and energy conversion. Students found the activation and linking of these more abstract elements particularly difficult. The betweenness centrality of chemical conversion—indicating its bridging role in the network—showed consistently high correlations with the posttest score early in and throughout the unit. In contrast, the degree centrality values (reflecting the integration) of these key concepts became a predictive measure of performance only towards the end of the unit. This difference suggests that the bridging function of core concepts is crucial early on, while their full integration and prominence in the network only become predictive during the final consolidation phase.

However, core concepts rarely seem to be the focus of teaching practice (Bernholt et al. 2020). Accordingly, it may be even more appropriate in the case of these knowledge elements to encourage students to integrate them at an early stage in a unit in corresponding exercises but also to give direct feedback on corresponding individual tasks and to ask students to revise their solutions (Karademir et al. 2024a). In addition, an evaluation at the level of the individual knowledge elements provides very specific information, for example, about which knowledge elements are hardly used by the students in the tasks of a class or about which connections between knowledge elements are especially difficult for the students; teachers could react to this information by employing corresponding exercises and repetition tasks. However, if time resources are limited, the focus should be on the (normatively expected) most central knowledge elements, as they are most likely to contribute to cumulative knowledge building in the long term.

8.3 Implications for teaching and future research

The findings validate the potential of real-time analysis of knowledge networks as a powerful tool for learning process diagnostics. If further studies confirm the different patterns of results, depending on the objectives of the respective teaching unit, across different topics, the findings could be used to help teachers to select specific strategies with which they can support their students in each case. In phases of exploration, for instance, teachers might want to focus on fostering active restructuring (high variance) and preventing excessive, unstructured connectivity (low AUC for density). In phases of consolidation, teachers might better support students by strengthening the consistent application of the established knowledge structure (high ACF1 of network size).

Generally, the successful integration and enactment of the unit's core concepts deserves targeted monitoring. Constant attention should be paid to the bridging function (betweenness centrality) of foundational concepts when evaluating task completion. As these concepts are difficult to integrate, specific interventions to ensure their prominent and highly integrated position (high degree centrality) in the network are desirable.

This study confirms the feasibility of the automated scoring of knowledge elements across an entire, format-diverse teaching unit. Future research should prioritize the refinement of the intermediate scoring level (i.e., "basic integration"), potentially by using specific linguistic examples to ensure clearer distinctions. Furthermore, analyzing systematic tendencies in misclassifications is crucial, as supportive strategies depend on whether performance is consistently underestimated or overestimated.

The findings presented also provide indicators of potential areas for instructional interventions over the course of the unit, for example, in the form of task-related feedback or the identification of student difficulties. Consideration should also be given to preparing the results in a form that can be readily used by teachers (Karademir et al. 2024b). A differentiated presentation of the results at the class level (which knowledge elements are difficult for the whole class?) and at the student level (which individual students show consistent below-average progress and need more support?) could be useful here in order to allocate the teacher's time resources efficiently.

Overall, the results show that incorporating dynamic network metrics contributes significantly to understanding and predicting learning success in complex, digitally supported teaching and learning scenarios; these insights can be used to inform targeted instructional decisions in different learning phases.

Supplementary Information The online version of this article (https://doi.org/10.1007/s42010-026-00244-0) contains supplementary material, which is available to authorized users.

Funding This work was supported by the Leibniz Association, Germany, under Grant K365/2020.

Author Contribution All authors contributed to the study conception and design. Material preparation and data collection were performed by Jannik Lossjew and Sebastian Gombert, data preparation by Jannik Lossjew, and data analysis was performed by Sascha Bernholt. The first draft of the manuscript was written by Sascha Bernholt and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest S. Bernholt, J. Lossjew and S. Gombert declare that they have no competing interests.

References

American Association for the Advancement of Science, & National Science Teachers Association (2007). *Atlas of science literacy: Project 2061*. AAAS.

Asmussen, G., Rodemer, M., Eckhard, J., & Bernholt, S. (2022). From free association to goal-directed problem-solving: Network analysis of students' use of chemical concepts in mechanistic reasoning. In N. Graulich & G. Shultz (Eds.), *Student reasoning in organic chemistry: Research advances and evidence-based instructional practices* (pp. 90–109). Royal Society of Chemistry.

Bain, K., & Towns, M.H. (2016). A review of research on the teaching and learning of chemical kinetics. *Chemistry Education Research and Practice*, *17*(2), 246–262. https://doi.org/10.1039/C5RP00176E.

Barabási, A.-L., Albert, R., & Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A—Statistical Mechanics and Its Applications*, *272*(1–2), 173–187. https://doi.org/10.1016/S0378-4371(99)00291-5.

Bender-deMoll, S. (2024). *ndtv (R package; Version 0.13.4) [Computer software]*. https://CRAN.R-project.org/package=ndtv

Bender-deMoll, S., & Morris, M. (2021). *tsna (R package; Version 0.3.5) [Computer software]*. https://CRAN.R-project.org/package=tsna

Bernholt, S., Höft, L., & Parchmann, I. (2020). Die Entwicklung fachlicher Basiskonzepte im Chemieunterricht – Findet ein kumulativer Aufbau im Kompetenzbereich Fachwissen statt? *Unterrichtswissenschaft*, *48*(1), 35–59. https://doi.org/10.1007/s42010-019-00065-4.

BIMI (2016). *Fachanforderungen Chemie: Allgemein bildende Schulen Sekundarstufe I (Gymnasium) und Sekundarstufe II*. Ministerium für Schule und Berufsbildung des Landes Schleswig-Holstein.

Blömeke, S., Kaiser, G., König, J., & Jentsch, A. (2020). Profiles of mathematics teachers' competence and their relation to instructional quality. *ZDM—Mathematics Education*, *52*(2), 329–342. https://doi.org/10.1007/s11858-020-01128-y.

Blossfeld, H.-P., Bos, W., Daniel, H.-D., Hannover, B., Köller, O., Lenzen, D., McElvany, N., Roßbach, H.-G., Seidel, T., Tippelt, R., & Wößmann, L. (2018). *Digitale Souveränität und Bildung: Gutachten*. Waxmann.

Brandes, U., & Erlebach, T. (Eds.). (2005). *Network analysis: Methodological foundations*. Springer. https://doi.org/10.1007/b106453.

Bransford, J.D., Brown, A.L., & Cocking, R.R. (2000). *How people learn: Brain, mind, experience and school*. National Academy Press.

Bureau, J. S., Howard, J. L., Chong, J. X. Y., & Guay, F. (2022). Pathways to student motivation: A meta-analysis of antecedents of autonomous and controlled motivations. *Review of Educational Research*, *92*(1), 46–72. https://doi.org/10.3102/00346543211042426.
Butts, C. T. (2008). network: A package for managing relational data in R. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v024.i02.
Chan, B., Schweter, S., & Möller, T. (2020). German's next language model. In D. Scott, N. Bel & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6788–6796). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.598.
Chi, M. T. H., & Roscoe, R. D. (2002). The process and challenges of conceptual change. In M. Limón & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 3–27). Kluwer Academic Publishers.
Corcoran, T., Mosher, F., & Rogat, A. D. (2009). *Learning progressions in science: An evidence-based approach to reform*. Consortium for Policy Research in Education. https://doi.org/10.12698/cpre.2009.rr63.
Cromley, J. G., Mirabelli, J. F., & Kunze, A. J. (2024). Three applications of semantic network analysis to individual student think-aloud data. *Contemporary Educational Psychology*, *79*, 102318. https://doi.org/10.1016/j.cedpsych.2024.102318.
Daems, O., Erkens, M., Malzahn, N., & Hoppe, H. U. (2014). Using content analysis and domain ontologies to check learners' understanding of science concepts. *Journal of Computers in Education*, *1*(2–3), 113–131. https://doi.org/10.1007/s40692-014-0013-y.
Derry, S. J. (1996). Cognitive schema theory in the constructivist debate. *Educational Psychologist*, *31*(3–4), 163–174. https://doi.org/10.1080/00461520.1996.9653264.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.
D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001.
Duit, R., Treagust, D. F., & Widodo, A. (2013). Teaching science for conceptual change. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 487–503). Routledge.
Dumont, H. (2016). Die empirische Untersuchung von individueller Förderung als Perspektive für die Unterrichtsqualitätsforschung. In N. McElvany, W. Bos, H.-G. Holtappels, M. M. Gebauer & F. Schwabe (Eds.), *Bedingungen und Effekte guten Unterrichts*. Waxmann-E-Books Didaktik, Schule und Unterricht, (Vol. 1, pp. 107–116). Waxmann.
Dumont, H. (2019). Neuer Schlauch für alten Wein? Eine konzeptuelle Betrachtung von individueller Förderung im Unterricht. *Zeitschrift für Erziehungswissenschaft*, *22*(2), 249–277. https://doi.org/10.1007/s11618-018-0840-0.
Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in Grades K-8*. National Academies Press.
Emden, M., Weber, K., & Sumfleth, E. (2018). Evaluating a learning progression on 'Transformation of Matter' on the lower secondary level. *Chemistry Education Research and Practice*, *19*(4), 1096–1116. https://doi.org/10.1039/C8RP00137E.
Faber, J. M., Luyten, H., & Visscher, A. J. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. *Computers & Education*, *106*, 83–96. https://doi.org/10.1016/j.compedu.2016.12.001.
Freedman, H., Young, N., Schaefer, D., Song, Q., van der Hoek, A., & Tomlinson, B. (2024). Construction and analysis of collaborative educational networks based on student concept maps. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–22. https://doi.org/10.1145/3637313.
Giabbanelli, P. J., & Tawfik, A. A. (2021). How perspectives of a system change based on exposure to positive or negative evidence. *Systems*, *9*(2), 23. https://doi.org/10.3390/systems9020023.
Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521–563. https://doi.org/10.1080/10508406.2013.837391.
Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*(7), 729–757. https://doi.org/10.1037/bul0000043.
Gombert, S., Di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., Grimm, A., Bohm, I., Neumann, K., & Drachsler, H. (2023). Coding energy knowledge in constructed responses with explain-

able NLP models. *Journal of Computer Assisted Learning*, *39*(3), 767–786. https://doi.org/10.1111/jcal.12767.

Gupta, A., Hammer, D., & Redish, E. F. (2010). The case for dynamic models of learners' ontologies in physics. *Journal of the Learning Sciences*, *19*(3), 285–321. https://doi.org/10.1080/10508406.2010.491751.

Kapuza, A., Koponen, I. T., & Tyumeneva, Y. (2020). The network approach to assess the structure of knowledge: Storage, distribution and retrieval as three measures in analysing concept maps. *British Journal of Educational Technology*, *51*(6), 2574–2591. https://doi.org/10.1111/bjet.12938.

Karademir, O., Borgards, L., Di Mitri, D., Strauß, S., Kubsch, M., Brobeil, M., Grimm, A., Gombert, S., Rummel, N., Neumann, K., & Drachsler, H. (2024a). Following the impact chain of the LA cockpit. *Journal of Learning Analytics*, *11*(2), 215–228. https://doi.org/10.18608/jla.2024.8399.

Karademir, O., Di Mitri, D., Schneider, J., Jivet, I., Allmang, J., Gombert, S., Kubsch, M., Neumann, K., & Drachsler, H. (2024b). I don't have time! But keep me in the loop: Co-designing requirements for a learning analytics cockpit with teachers. *Journal of Computer Assisted Learning*, *40*(6), 2681–2699. https://doi.org/10.1111/jcal.12997.

Kim, K. (2025). How the relationship between verbal and nonverbal text signals influences second language expository text comprehension: Knowledge structure analysis. *Educational Technology Research and Development*, *73*(1), 15–34. https://doi.org/10.1007/s11423-024-10405-6.

Kim, M. K., & McCarthy, K. S. (2021). Improving summary writing through formative feedback in a technology-enhanced learning environment. *Journal of Computer Assisted Learning*, *37*(3), 684–704. https://doi.org/10.1111/jcal.12516.

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018). The INCEpTION Platform: Machine-assisted and knowledge-oriented interactive annotation. In D. Zhao (Ed.), *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (pp. 5–9). Association for Computational Linguistics. https://aclanthology.org/C18-2002/.

KMK (2004). *Bildungsstandards im Fach Chemie für den mittleren Schulabschluss*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, *36*(5), 757–798. https://doi.org/10.1111/j.1551-6709.2012.01245.x.

Kubsch, M., Nordine, J., Neumann, K., Fortus, D., & Krajcik, J. (2019). Probing the relation between students' integrated knowledge and knowledge-in-use about energy using network analysis. *EURASIA Journal of Mathematics, Science and Technology Education*, *15*(8), 1–20. https://doi.org/10.29333/ejmste/104404.

Kubsch, M., Czinczel, B., Lossjew, J., Wyrwich, T., Bednorz, D., Bernholt, S., Fiedler, D., Strauß, S., Cress, U., Drachsler, H., Neumann, K., & Rummel, N. (2022). Toward learning progression analytics: Developing learning environments for the automated analysis of learning using evidence centered design. *Frontiers in Education*, *7*(981910), 1–15. https://doi.org/10.3389/feduc.2022.981910.

Kubsch, M., Strauß, S., & Bernholt, S. (2023). Integrating perspectives to promote knowledge integration: How knowledge integration, learning progressions and instructional science can complement each other. In *Proceedings of the 17th International Conference of the Learning Sciences—ICLS 2023* (pp. 1897–1898). International Society of the Learning Sciences. https://doi.org/10.22318/icls2023.643889.

Lee, H.-S., & Liu, O. L. (2010). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education*, *94*(4), 665–688. https://doi.org/10.1002/sce.20382.

Lee, H.-S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, *103*(3), 590–622. https://doi.org/10.1002/sce.21504.

Linn, M. C., Eylon, B.-S., Rafferty, A., & Vitale, J. M. (2015). Designing instruction to improve lifelong inquiry learning. *EURASIA Journal of Mathematics, Science and Technology Education*. https://doi.org/10.12973/eurasia.2015.1317a.

Linnenbrink-Garcia, L., Patall, E. A., & Pekrun, R. (2016). Adaptive motivation and emotion in education. *Policy Insights from the Behavioral and Brain Sciences*, *3*(2), 228–236. https://doi.org/10.1177/2372732216644450.

Liu, O. L., Lee, H.-S., Hofstetter, C., & Linn, M. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment*, *13*(1), 33–55. https://doi.org/10.1080/10627190801968224.

Lossjew, J., & Bernholt, S. (2025). Digital support custom-made or in moderation—Design and evaluation of a digitally supported teaching unit on reaction kinetics and chemical equilibrium. *Chemie Konkret*, *32*(4), 122–131. https://doi.org/10.1002/ckon.202400061.

Luke, D. (2015). *A user's guide to network analysis in R*. Springer. https://doi.org/10.1007/978-3-319-23883-8.

Morrison, D., Bedinger, M., Beevers, L., & McClymont, K. (2022). Exploring the raison d'etre behind metric selection in network analysis: A systematic review. *Applied Network Science*, *7*(1), 50. https://doi.org/10.1007/s41109-022-00476-w.

Murphy, G. L. (2004). *The big book of concepts*. MIT Press.

Nakamura, C. M., Murphy, S. K., Christel, M. G., Stevens, S. M., & Zollman, D. A. (2016). Automated analysis of short responses in an interactive synthetic tutoring system for introductory physics. *Physical Review Physics Education Research*. https://doi.org/10.1103/PhysRevPhysEducRes.12.010122.

Novak, J. D. (2002). Meaningful learning: The essential factor for conceptual change in limited or inappropriate propositional hierarchies leading to empowerment of learners. *Science Education*, *86*(4), 548–571. https://doi.org/10.1002/sce.10032.

Podschuweit, S., & Bernholt, S. (2020). Investigating network coherence to assess students' conceptual understanding of energy. *Education Sciences*, *10*(4), 103. https://doi.org/10.3390/educsci10040103.

Potvin, P., Sauriol, É., & Riopel, M. (2015). Experimental evidence of the superiority of the prevalence model of conceptual change over the classical models and repetition. *Journal of Research in Science Teaching*, *52*(8), 1082–1108. https://doi.org/10.1002/tea.21235.

Prediger, S., & von Aufschnaiter, C. (2017). Umgang mit heterogenen Lernvoraussetzungen aus fachdidaktischer Perspektive: Fachspezifische Anforderungs- und Lernstufungen berücksichtigen. In T. Bohl, J. Budde & M. Rieger-Ladich (Eds.), *Umgang mit Heterogenität in Schule und Unterricht: Grundlagentheoretische Beiträge, empirische Befunde und didaktische Reflexionen*. UTB Schulpädagogik, (Vol. 4755, pp. 291–308). Verlag Julius Klinkhardt.

Pruessner, J. C., Kirschbaum, C., Meinlschmid, G., & Hellhammer, D. H. (2003). Two formulas for computation of the area under the curve represent measures of total hormone concentration versus time-dependent change. *Psychoneuroendocrinology*, *28*(7), 916–931. https://doi.org/10.1016/S0306-4530(02)00108-7.

R Core Team (2024). *R: A language and environment for statistical computing*. https://www.R-project.org/

Schneider, M., & Stern, E. (2010). The developmental relations between conceptual and procedural knowledge: A multimethod approach. *Developmental Psychology*, *46*(1), 178–192. https://doi.org/10.1037/a0016701.

Schneider, B. L., Krajcik, J. S., Lavonen, J., & Salmela-Aro, K. (2020). *Learning science: The value of crafting engagement in science environments*. Yale University Press. https://doi.org/10.12987/9780300252736.

Schwendimann, B. A. (2014). Making sense of knowledge integration maps. In D. Ifenthaler & R. Hanewald (Eds.), *Digital knowledge maps in education* (pp. 17–40). New York: Springer. https://doi.org/10.1007/978-1-4614-3178-7_2.

Schwendimann, B. A., & Linn, M. C. (2016). Comparing two forms of concept map critique activities to facilitate knowledge integration processes in evolution education. *Journal of Research in Science Teaching*, *53*(1), 70–94. https://doi.org/10.1002/tea.21244.

Siegler, R. S. (1994). Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Science*, *3*(1), 1–5. https://doi.org/10.1111/1467-8721.ep10769817.

Siegler, R. S. (2007). Microgenetic analyses of learning. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology*. Wiley. https://doi.org/10.1002/9780470147658.chpsy0211.

Skinner, B. F. (1958). Teaching machines: From the experimental study of learning come devices which arrange optimal conditions for self instruction. *Science*, *128*(3330), 969–977. https://doi.org/10.1126/science.128.3330.969.

Talanquer, V. (2018). Importance of understanding fundamental chemical mechanisms. *Journal of Chemical Education*, *95*(11), 1905–1911. https://doi.org/10.1021/acs.jchemed.8b00508.

Tissenbaum, M., & Slotta, J. (2019). Supporting classroom orchestration with real-time feedback: A role for teacher dashboards and real-time agents. *International Journal of Computer-Supported Collaborative Learning*, *14*(3), 325–351. https://doi.org/10.1007/s11412-019-09306-1.

Unger, L., Vales, C., & Fisher, A. V. (2020). The role of co-occurrence statistics in developing semantic knowledge. *Cognitive Science*, *44*(9), e12894. https://doi.org/10.1111/cogs.12894.

van Driel, J. H., de Vos, W., Verloop, N., & Dekkers, H. (1998). Developing secondary students' conceptions of chemical reactions: The introduction of chemical equilibrium. *International Journal of Science Education*, *20*(4), 379–392. https://doi.org/10.1080/0950069980200401.

Vygotskij, L. S. (Ed.). (1981). *Mind in society: The development of higher psychological processes*. Harvard University Press. Reprint

Wagner, S., & Priemer, B. (2023). Assessing the quality of scientific explanations with networks. *International Journal of Science Education*, *45*(8), 636–660. https://doi.org/10.1080/09500693.2023.2172326.

Weinrich, M., & Talanquer, V. (2015). Mapping students' conceptual modes when thinking about chemical reactions used to make a desired product. *Chemistry Education Research and Practice*, *16*(3), 561–577. https://doi.org/10.1039/C5RP00024F.

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686.

Wijnants, M. L. (2014). A review of theoretical perspectives in cognitive science on the presence of 1/f scaling in coordinated physiological and cognitive processes. *Journal of Nonlinear Dynamics*, *2014*, 1–17. https://doi.org/10.1155/2014/962043.

Yan, F., & Talanquer, V. (2016). Students' ideas about how and why chemical reactions happen: Mapping the conceptual landscape. *International Journal of Science Education*, *37*(18), 3066–3092. https://doi.org/10.1080/09500693.2015.1121414.

Yan, Y. K., & Subramaniam, R. (2016). Diagnostic appraisal of grade 12 students' understanding of reaction kinetics. *Chemistry Education Research and Practice*, *17*(4), 1114–1126. https://doi.org/10.1039/C6RP00168H.

Yang, Z., Zhu, M., Qu, Z., & Zhang, Y. (2018). Research on organization of mathematics knowledge in good mathematical cognitive structure. *EURASIA Journal of Mathematics, Science and Technology Education*. https://doi.org/10.12973/ejmste/78242.