American Journal of Qualitative Research[1]
2026, Vol. 10 No. 1, pp. 37-60
https://doi.org/10.29333/ajqr/17367

AMERICAN JOURNAL OF QUALITATIVE RESEARCH
ISSN: 2576-2141

Critical Thinking Performance Assessment in an Undergraduate Physical Chemistry Laboratory Course: Developing a Contextual Critical Thinking Coding Manual

Lilian Danial
Humboldt University of Berlin, Berlin, Germany

Jenna Koenen
Technical University of Munich, Munich, Germany

Rüdiger Tiemann[2]
Humboldt University of Berlin, Berlin, Germany

ABSTRACT
Critical thinking (CT) performance assessment in authentic settings remains challenging. One reason is that the CT focus might shift depending on the context. This article describes the inductive and deductive processes of developing a contextual CT coding manual (CCTCM) and presents an exemplary analysis to demonstrate its application in assessing CT performance in an undergraduate physical chemistry laboratory course. Students' verbal interactions during the course were used as the primary data for assessment. The CCTCM we developed comprises two components: the contextual profile of the physical chemistry laboratory and an operationalized set of CT skills. The tool enabled a contextualized analysis of CT performance. It produced data that reflected ongoing learning, accommodated quantitative analyses of small samples, and could inform curriculum design by identifying commonly applied or underused CT skills. Moreover, it allowed for monitoring individual CT performance and understanding CT skill interconnections. The CCTCM operates independently of teaching methods or students' problem familiarity, thus preserving intuitive behaviors.

KEYWORDS: Critical thinking, domain specific, performance assessment, undergraduate, chemistry education, science education, laboratory course, coding manual, qualitative data analysis, authentic setting, inductive, deductive, higher education

This work focuses on critical thinking (CT) assessment. CT is widely regarded as a key competence in higher education (HE) because it fosters students' college and workplace success (European Commission, Directorate-General for Education, Youth, Sport and Culture, 2019; Voogt & Roblin, 2010). Assessment tools are important for curriculum development, teaching, and learning outcomes because they provide necessary feedback on their alignment (La Marca et al., 2000). However, the available assessment tools, such as closed-ended, generic, and structured questionnaires, fail to capture the complexity of CT development and integration in teaching and learning (Lai, 2011).

[2] Corresponding author: a professor for chemistry education research at the Chemistry Department, Humboldt University of Berlin, Berlin, Germany, Brook-Taylor-Str. 2, 12489 Berlin, Germany. Email: ruediger.tiemann@hu-berlin.de

The challenges in CT assessment can be linked to one main argument on CT that it is domain specific (Bailin, 2002; McPeck, 1990; Willingham, 2007). A prominent argument on the domain specificity of CT suggests that CT transferability to other domains is not automatic (Kennedy et al., 1991; Pithers & Soden, 2000; Willingham, 2007). However, the available tools for CT assessment (see Carter et al., 2015; Ennis, 1993; Stupple et al., 2017) are, in general, not context specific and mostly claim to be suitable across contexts by focusing usually on real-life problems. Given the current research on CT assessment and its challenges, we argue that to assess progress in CT, it is essential to first assess CT within the same domain in which it is being practiced.

When addressing CT at the undergraduate level in a physical chemistry laboratory course, how can we expect to measure changes in CT using generic tests? One major issue raised here is validity. Do these tools assess what we aim to measure? Raising these issues points to the current gap. To date, no domain-specific assessment tool exists for undergraduate physical chemistry laboratories that enables the in situ analysis of CT in student group discourse. This emphasizes the need to develop a valid and reliable domain-specific tool to assess the CT applied in practice.

This article addresses this gap by presenting the deductive and inductive development processes of a contextual CT coding manual (CCTCM) to assess domain-specific CT performance in authentic settings of an undergraduate physical chemistry laboratory course. Herein, we outline the steps used to establish rigor in developing the CCTCM, along with the detailed methodology for evaluating CT skills within the context of this course. In addition, we discuss the outcomes through an exemplary analysis.

The CCTCM is designed for the following:

- Instructors: assessing students' CT performance in laboratory courses while identifying areas for targeted instruction and interventions to strengthen specific CT skills.
- Researchers: providing a robust framework for studying CT in discipline-specific, authentic educational contexts, thereby contributing to the broader field of CT assessment and its pedagogical implications.
- Curricular Designers: identifying gaps in existing curricula to inform the evidence-based development of learning materials to foster CT skills in science.

Research Objective

Our objective was to evaluate and capture the use of CT skills within the context of an undergraduate physical chemistry laboratory course. Specifically, we aimed to determine whether students actively employ CT skills, identify which skills are utilized, and understand the context in which these skills are applied.

In this course, students conduct experiments in pairs following predesigned scripts (in the style of a "cookbook") and discuss issues related to the experimental procedures as they arise. To assess the applied CT skill within the course, we opted for a performance assessment based on a coding manual, which enabled us to systematically analyze the students' verbal interactions during the experiments. The discussions during the experiment were used as the source of the primary data for assessing CT, which aligns with Lai's (2011) and Norris's (1989) recommendations that encourage making CT visible in authentic learning contexts. An authentic setting refers to learning environments that are free from alterations, interruptions, and interventions in the teaching styles, materials, and syllabus, and thus allow for observations of natural student behaviors and interactions (Bagnato, 2007). The authentic setting in this course provides an ideal environment to observe natural student interactions. For this purpose, we developed the CCTCM, which is designed to code and assess CT performance and applied skills within an authentic physical chemistry laboratory course setting. The course does not provide explicit instructions on CT or its transferability across domains.

Our Guiding Questions

RQ1: What are the steps for ensuring validity, reliability, and rigor in developing the CCTCM?
RQ2: How can we assess domain-specific CT performance in an undergraduate physical chemistry laboratory course?

Theoretical Framework

Critical Thinking

CT is among a family of closely related forms of higher-order thinking skills, such as problem-solving, decision-making, and creative thinking. According to Robert Ennis, one of the founders of and contributors to the field of inquiry into CT, CT is a "reasonable reflective thinking focused on deciding what to believe or do" (Ennis, 1985a, p. 45; 2015, p. 32). This strong focus on the outcome or utility of CT was also at the core of Halpern's (2013, 2014) definition, which highlights CT as a tool to facilitate decision-making or problem-solving (Liu et al., 2014).

In the 1990s, the American Philosophical Association (APA) Delphi report (Facione, 1990a), a landmark report, was published. This report, written by forty-six experts, formed a consensus for the definition of CT and its subskills. According to this report, CT is an essential tool of inquiry. It operates in two dimensions: cognitive skills and affective dispositions. A person skilled at a specific construct does not necessarily mean that this person will apply that set of skills to a given situation. Knowing a person's disposition allows for predicting how the person is most likely to act in various circumstances (Facione, 2000). CT dispositions are a person's consistent internal motivation to act or respond to another person, events, or circumstances using CT skills (Facione, 2000). The Delphi report conceptualizes six core CT skills: interpretation, analysis, inference, evaluation, explanation, and self-regulation (Facione, 1990a; Thomas & Lok, 2015). In their operational framework for teaching CT, Thomas and Lok (2015) reviewed 16 definitions of CT skills and found that all of them are based on the skills stated in the APA Delphi report. This suggests that although the APA Delphi report was published more than 35 years ago, its CT definitions and core skills form the basis for the major consecutive publications focused on CT skills to date (Andreucci-Annunziata et al., 2023; Butler, 2024; Heard et al., 2020).

CT Assessment: Available Tools, Limitations, and Criticisms

The available and widely used tools for measuring CT are multiple-choice, generic, and focused mostly on problem-solving and formal reasoning (Stupple et al., 2017). However, as Chowdhury and Shil (2021) argue, understanding complex thinking in real-life contexts requires capturing the "holistic and meaningful characteristics" of learners' interactions and experiences, which is otherwise difficult to achieve through standardized, multiple-choice formats.

The common tools for measuring CT include the Cornell Critical Thinking Test (Ennis, 1985c), the Oxford, Cambridge and RSA Examination Board CT examination (Wells et al., 2006), the Thinking Skills Assessment (Black, 2012; Black et al., 2008), and the California Critical Thinking Skills Test (CCTST) (Facione, 1990b). Although lacking in comprehensiveness, multiple-choice tests can be used for several purposes: diagnosis, feedback, motivation, impact of teaching, and research. The wide use of such tools relies on their logistic simplicity and cost/time efficiency, unlike CT essay tests (Ennis, 1993). Creating valid and reliable multiple-choice tests is difficult, time-consuming, and requires a series of revisions and tryouts. Therefore, using established tools is easier. The criticisms of the use of multiple-choice tests are not focused solely on their limited ability to convey a comprehensive view of the multidimensional and complex nature of CT. This limitation aligns with the critique in interpretive research that such tools reduce social phenomena to oversimplified variables, ignoring the nuanced and emergent nature of thinking as it unfolds in real contexts (Chowdhury & Shil, 2021). Other issues with multiple-choice tests are also mentioned in the literature:

Ennis (1993) highlights that differences in background beliefs and assumptions between test makers and test takers result in justifiably different answers to test questions (Ennis, 1993; Norris & Ennis, 1989). Another issue is validity, namely construct, content, and face validity. Ennis (1993) explained that when relying on tests for high-stakes purposes, the test items are limited to multiple-choice deductive-logic items in which conclusions necessarily follow the premises, and the "right" answer is less of a subject for debate. This limits the comprehensiveness and compromising content validity of the test (Ennis, 1993; Stupple et al., 2017).

Furthermore, CT is a multidimensional concept and is, by definition, a "reasonable" and "reflective" (Ennis, 1985a, 1985b, p. 45; 2015, p. 32) way of thinking. Thus, limiting tests to multiple-choice items can exclude much of this defining aspect of CT. CT assessment should enable the expression of its complexity. Therefore, a mixed-method approach and triangulation of tools may provide greater validity, reliability, and insight into CT development (Carter et al., 2015).

Carter et al. (2015) reviewed the available tools used to measure CT development in undergraduate nursing and midwifery students. They reviewed studies from nine countries using 16 tools (including the four widely used tools: the CCTST, California Critical Thinking Disposition Inventory (Facione & Facione, 1992), Health Sciences Reasoning Test, and Watson–Glaser Critical Thinking Appraisal). Across the four tools, the issues of inconsistency in findings and invalidity were substantial, placing some doubt on the reliability of these tools in different nursing education contexts. The remaining 12 non-standardized tools had other issues, such as costs, reliability over time, lack of information about the development process, and lack of theory-based development (Carter et al., 2015).

Other CT assessment tools are open-ended (e.g., the Halpern Critical Thinking Assessment, which combines multiple-choice and open-ended formats, and the Ennis Weir Critical Thinking essay test; Ku, 2009). Such tools are more comprehensive and allow for more tailored assessments. However, they are labor intensive in terms of post-analysis and grading. A performance assessment has greater face validity. It assesses situations in realistic and authentic situations (Ennis, 1993), such as performing an experiment in an undergraduate physical chemistry laboratory course. Performance assessments are consistent with the principles of field-based qualitative research, which prioritize immersion in natural settings, evolving research designs, and thick, contextualized representations of behavior (Chowdhury & Shil, 2021; Ferreira & Merchant, 1992). One downside of this assessment lies in its relation to reality. The more realistic the setting is, the less assurance of comprehensiveness (Ennis, 1993). As Ennis (1993) explains, "In real-life situations, people generally reveal only what the situation requires, and most observable situations do not require all aspects of [CT]" (p. 185). While performance assessments may not elicit all aspects of CT in every situation (Ennis, 1993), qualitative approaches, such as case studies, embrace this contextual complexity to capture how CT unfolds in real-life settings through detailed, holistic descriptions grounded in participants' perspectives (Chowdhury & Shil, 2021; Patton, 2002).

In reviewing CT assessment in HE, Liu et al. (2014) argue that even though disagreement still exists regarding whether CT is domain specific, developing tools for major fields can deprive students of CT assessments and trainings in less popular domains. In addition, developing CT assessment tools is complex and requires collaboration between domain experts (Liu et al., 2014). Another issue raised by Ennis (1993) is excessive subjectivity via interpretation in performance assessment. Lastly, performance assessment is the most expensive of all types of assessment, as it requires devoting considerable expert time to assessing data.

Even more recent tools that advocate for performance assessment in CT, such as the International Performance Assessment of Learning in HE (Shavelson et al., 2018), are based on storylines drawn from real-world situations related to business, public policy, civics, medicine, and family. They often involve moral and ethical considerations (Braun et al., 2020), and are related to a general domain, assuming that addressing such challenges requires the application of multiple CT skills suited to such tasks. CT skills are elicited by the questions presented in the tool (Lai, 2011), forming a "stimulus" by "forcing" explicit CT skills in students' performances (Fischer et al., 2009). Such tools are not independent of the CT teaching approach that precedes assessment or students' familiarity with such topics or tasks (Braun et al., 2020). Moreover, such CT assessment tasks may influence student motivation because they can be more challenging, novel, and interesting (Lai, 2011).

Although CT performance assessments are considered to be more valid representations of the CT construct, they can be susceptible to lower reliability and a lack of generalizability. Lai (2011) points out that when generalizability is compromised, the tools should not be used to compare students from various domains or to track student progress or growth over time. However, it can still be used for low-stakes classroom assessments. Despite the broad agreement on CT relevance, Andreucci-Annunziata et al. (2023) highlight the gap in CT assessment that satisfies the complex and diverse aspects of CT.

CT: Domain Specificity and Transferability

In the context of assessment, Ennis (1993) highlights that subject-specific CT tests assess CT within one subject matter. By contrast, general content–based CT tests (e.g., the generic multiple-choice tests) use content from various areas with which test takers are presumed to be already familiar (e.g., everyday problems). Ennis (1990) argues that CT is domain-specific and thus requires background knowledge and is unlikely to transfer between domains without explicit instruction. Consequently, generic multiple-choice tests to measure CT in a specific domain may be unreliable. Ennis (1989) further notes that transfer becomes more likely with sufficient practice across domains and transfer-focused instruction. Lai (2011) supports this idea, emphasizing that many existing tools are not subject-specific and thus measure CT transfer under the presumption of either familiarity with the tool's subject matter or having explicit instruction on CT transfer. Nevertheless, generic tools may be suitable for evaluating CT transferability, regardless of the original instructional context (Ennis, 1993).

Ennis (1989) further elaborates that "it does not even make sense to speak of [CT] or [CT] instruction outside of a subject-matter area, and the idea *general [CT] ability* is meaningless.… as … 'thinking is always thinking *about* something'" (McPeck, 1981, p. 3, as cited in Ennis, 1989, p. 8). On the other end of the argument on CT transferability stands Paul et al. (1997), who argue that "there is no reason in principle that students cannot take the basic tools of critical thought which they learn in one domain of study and extend it *with appropriate adjustments* [emphasis added] to all the other domains and subjects which they study" (Paul et al., 1997, pp. 11, as cited in Heard et al., 2020). However, its "appropriate adjustments" (which was excluded in the citation of Heard et al., 2020, p. 9) are linked to the idea that there is a need for "instruction that focuses on transfer" (Ennis, 1989, p. 5). The idea suggested here goes beyond the issue of whether CT is subject specific or has the likelihood of transferability between domains; it is a matter of correspondence between the CT teaching approach and assessment.

Furthermore, while real-life CT assessment tools project CT usefulness beyond academia, Butler (2024) argues that most fail to predict students' ability to apply CT in everyday situations. It is quite understandable that CT skills can shift focus depending on the context and that different contexts provoke different aspects of CT, which in turn can affect its practical applications (e.g., Liu et al., 2001). Dominguez et al. (2018) found that in STEM (science, technology, engineering, and mathematics) fields, CT is more closely related to problem-solving and technical analysis, whereas in social sciences and humanities, it leans toward ethical reasoning and social analysis.

While some aspects of CT can be generalized to various domains, Stupple et al. (2017) note that CT is "conceptualized and taught in different ways" and added "within each discipline" (Stupple et al., 2017, p. 4). For instance, the National Assessment of College Student Learning Report (1995), which identifies college graduates' essential writing, speech, listening, and CT skills, was developed using the Delphi method, with faculty, employers, and policymakers. However, across the report, there were CT subskills that each sector found to be more important than other subskills (Jones et al., 1995), highlighting the shift in CT depending on the context.

Another example is the Australian Council for Educational Research (ACER) Critical Thinking Framework. It states that "functionally, the same abstract skill [each of the six core skills of the APA Delphi report] manifests as different skills *within different applications* [emphasis added]" (Heard et al., 2020, p. 12). The words "within different applications" clearly show that CT skills can have different foci or purposes depending on the context. In teaching, the assessment must be true to the learning process and context, highlighting the need to develop a valid assessment tool that considers CT in the same context. Thus, there is no denying or attempt to undermine that even when CT "'takes on the particularities of the discipline in which it resides,' there are nevertheless some identifiable core [CT] abilities that are both general in nature and generally applicable" (Jones, 2015, pp. 169–170, as cited in Heard et al., 2020, p. 11).

In our view, CT is "reasonable reflective thinking focused on deciding what to believe or do" (Ennis, 1985a, p. 45; 2015, p. 32). In addition, "in a specific context issue or problem," thinking is naturally "about something." To measure CT in practice, the performance assessment with the greater face validity allows one to "reveal only what the situation requires … [out of] all aspects of [CT]" (Ennis, 1993, p. 185) because different disciplines tend to emphasize different aspects of CT (Liu et al., 2001).

Other practical challenges face CT assessment in HE. Butler (2024) reports that there is often little incentive to prioritize CT assessment, as few institutions require strong evidence of CT development despite claiming its importance. Moreover, assessment is often considered extra work, handled during accreditation, and not integrated into regular faculty responsibilities (Butler, 2024). It is time-consuming to select appropriate tools with trade-offs between multiple-choice and short-answer formats. Financial costs further complicate implementation. Jabali et al. (2024) found that university grades form the basis of education policies and curriculum decisions, which can, in turn, inhibit CT and creativity, thus promoting rote memorization and technical thinking. They also found that faculty members criticized grade prioritization, citing conventional evaluation methods, temporal limitations, and resource constraints. Despite these challenges, CT assessment remains vital and should inform teaching improvements.

Methods

Before addressing RQ1, we provide descriptions of the physical chemistry laboratory setting and the role of the researcher. We then describe the process of developing the CCTCM and its final structure. For this section, we followed Braun and Clarke's (2024) guidelines to structure a rigorous and transparent reporting of this qualitative research.

Research Settings

This study was conducted in a second-year undergraduate physical chemistry laboratory course that comprised 16 experimental units that reflect standard topics in undergraduate physical chemistry, including kinetics, thermodynamics, electrochemistry, and spectroscopy, which are topics related to the theoretical concepts they previously studied in the theoretical course. Each learning activity in the course laboratory manual (LM) follows a consistent seven-part structure: (1) Aim, (2) Prior Knowledge, (3) Theoretical Background, (4) Measuring Principle/Method, (5) Procedure, (6) Reporting/Analysis of Results, and (7) References.

Before each laboratory session, the students underwent an oral examination for primarily assessment of their mastery of "Prior Knowledge" and "Theoretical Background." The students worked in pairs or small groups during the chemistry laboratory to conduct experiments and make real-time decisions about their methods and data collection. They determined whether the procedures were adequate, the data appeared valid, and the measurements needed repetition, drawing on course content and theoretical justifications to guide each choice. The course operates twice a week in a prefixed experiment setup. On a given day, each student group carries out a different experiment and rotates through the experiments in parallel.

Each experiment requires students to integrate theory with practice, operate specialized laboratory instruments, interpret results from advanced measurement methods, and collaboratively reason through problems. This decision-making process requires them to draw on prior knowledge (both conceptual and procedural), justify their decisions, and collaborate effectively. Accordingly, the overall aims of the course, as expected by the teaching staff, are (a) to familiarize students with experimental methods, tools, processes, calculations, and data analysis that are central to physical chemistry; and (b) to enhance their ability to apply and justify theoretical knowledge in practical decision-making within the laboratory setting.

Role of the Researcher

In this study, the teaching staff, syllabus, instruction, and course grading were fully independent of the research team; thus, the likelihood of direct influence on course content or classroom practices was reduced. The first author played the role of an external, passive, nonparticipant observer throughout the data collection period, attending laboratory sessions primarily to take field notes, record contextual observations, and ensure that all video and audio equipment functioned correctly. During these sessions, to preserve the authenticity of the natural classroom settings, the author did not intervene in any instructional activities, provide feedback to students/teaching staff, or influence the flow of the laboratory work.

Ethical Considerations

This study adhered to rigorous ethical standards throughout the research process. Although formal ethical approvals were not required at the time of data collection in Germany, the study complied with the ethical code of the German Research Foundation (DFG; Project number: 194424575), which was published at the planning and implementation stages of the study.

Participation in the study was entirely voluntary and limited to adult students who provided informed written consent after being briefed on the aims and procedures of the study, and their right to withdraw at any time without any consequences. No disadvantages were associated with nonparticipation. To protect the participants' privacy and confidentiality, all personal data were anonymized, with participant identities concealed and replaced by coded identifiers to facilitate data analysis.

Development Process of the CCTCM

In this section, to address RQ1, we outline the development of the CCTCM, guided by an inductive and deductive thematic analysis approach. In this process, the six-step model established by Fereday and Muir-Cochrane (2006) was followed to ensure methodological rigor throughout the process. Fereday and Muir-Cochrane used the “data-driven inductive approach of Boyatzis (1998) and the deductive a priori template of codes approach outlined by Crabtree and Miller (1999) to reach the second level of interpretive understanding (Schutz, 1967)” (p. 81). Their combined approach is integral to deductive thematic analysis, which involves creating codes that capture the qualitative richness of the phenomenon before the formulation of interpretation themes (Boyatzis, 1998). Simultaneously, it allows themes to emerge directly from the data through inductive coding (Fereday & Muir-Cochrane, 2006).

The development of the CCTCM codes relied on self-obtained data and key frameworks for CT skills (Butterworth & Thwaites, 2013; Ennis, 1985a, 1985b, 2015; Facione, 1990a and 2015; Heard et al., 2020). The Fereday and Muir-Cochrane (2006) model comprises six steps (described in the next subsection). Although the model appears to be a linear process, as presented in the original, the development of the CCTCM was “iterative and reflexive” (p. 83), which means that it included rereading, revising each step, and making sure that the codes were representative and grounded in the data.

Description of the Steps for Developing the CCTCM

Step 1: Developing the Coding Manual. The template for the CCTCM was developed on the basis of key definitions and frameworks that allowed us to (1) adapt and define the core CT skills and their subskills in an operationalized way and (2) use and adapt examples to reflect the application of each subskill in the context of the physical chemistry laboratory course. As presented in Figure 1, the CCTCM is composed of two parts: (1) the contextual profile (CP), which focuses on the topics and subtopics discussed by the students in the physical chemistry laboratory context, and (2) the CT skills and subskills profile (CTSP), which includes the operationalized CT core skills and subskills, and related examples that capture CT elements in practice based on the participants’ verbal interactions.

The CTSP was developed on the basis of the key CT frameworks referenced in the manual. The template for the CTSP was developed on the basis of the APA Delphi report for defining CT (Facione, 1990a) and a more recent version that is rich in explanations for the Delphi report and further examples (Facione, 2015). It also relied on definitions presented by Ennis (1985a, 1985b, 2015) specifically for instruction and assessment in HE. We also made adaptations based on the ACER CT framework (Heard et al., 2020) and other frameworks related to CT and problem-solving relevant to the context of the physical chemistry laboratory (Butterworth & Thwaites, 2013). We drew on established categories and coding schemes from the literature while revising or creating new codes and divisions. We focused on incorporating domain-specific examples, instrumentation, and language to reflect the specific nature of physical chemistry laboratory interactions. For each code, we created a parallel to distinguish between statements and questions raised by the students (see example of codes in the Annex). Using a theoretically driven template, as suggested by the model, provides a clear trail of evidence for the credibility of the CCTCM. Such an approach can contribute to construct, content, and face validity.

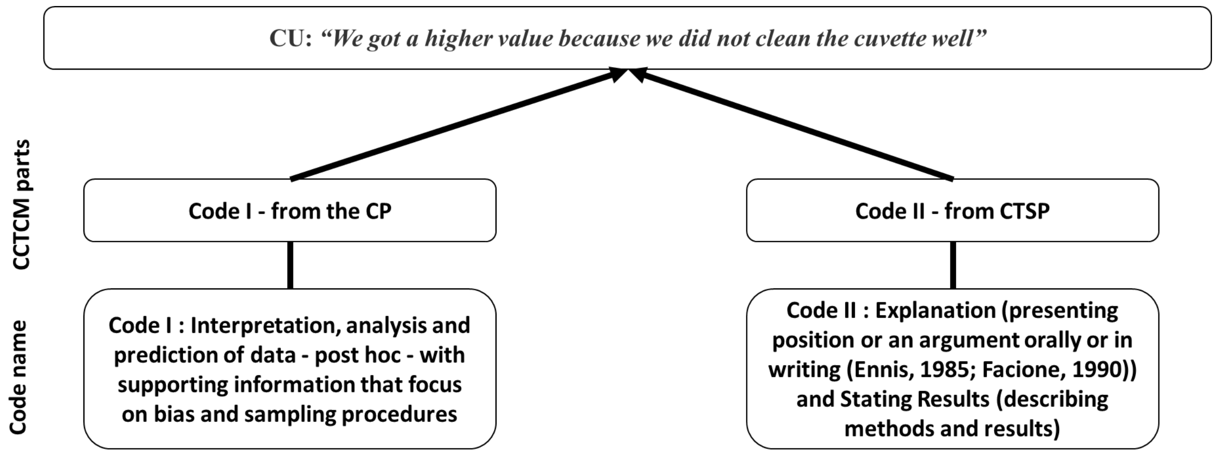
For the CP component, we identified broad thematic categories relevant to typical undergraduate physical chemistry laboratories. We consulted teaching staff, examined laboratory manuals, and reviewed the learning materials of the complete course to compile a set of potential topics and subtopics. This process helped us to create a context-specific coding scheme for the CP. Subsequently, we pilot-tested the initial categories by coding two samples of transcripts of the verbal interactions between the students (which were audio and video recorded, and transcribed) from two experiments by three coders (one author and two external experts in chemistry education) to refine and consolidate the CP codes. Each coder analyzed different segments of the transcripts independently. Subsequently, the author met with each coder to address any discrepancies or emerging codes. Through these discussions, we refined specific categories by, for example, subdividing an initial code labeled "Data Interpretation" into more precise subcodes that distinguish between interpreting obtained data and interpretation while providing explanations on what basis. Other subcodes were added to allow for the unification of the major code categories and structure of the CP. For example, we added a code titled "Generating explanation or expectation from tools – post hoc" to the second main topic that is parallel to all other main categories. For further unification, our final version of the CCTCM includes "others" sections at the end of each main category (e.g., "Theoretical Chemistry – Other") to allow for coding that belongs to the main topic but does not fall under the specific subcode and at the very end of each part of the CCTCM to allow for further analysis. After this final revision for the CP, another cycle of coding using the CP-CCTCM was performed, reaching more than 80% agreement between the coders.

Figure 1
Two Parts of the CCTCM: The CP and CTSP

CCTCM = Contextual Profile (CP) + CT Skills & Subskills Profile (CTSP)

For each coding unit (CU; i.e., statement, question, etc.), the CCTCM assigns two consecutive codes: The first is from the CP specifying the context, and the second is from the CTSP identifying the applied CT skill. See the example for coding in Figure 2. The reasoning behind this two-step coding order is presented in the subsection "Order of the Coding (CP First, Then CTSP)."

Figure 2
An Example of How CU Receives Two Codes From CCTCM: One from CP and the Second from CTSP



Each CCTCM code has a name, label, operational definition, relevant references, and examples in the context of the physical chemistry laboratory. In addition, we added further clarifications for the CP codes on "when to use the code" and "when not to use the code and alternative codes." The CCTCM distinguishes between CUs in question format and those in answer or statement format.

Step 2: Testing the Reliability of the Codes. As stated by Fereday and Muir-Cochrane (2006), the complete sets of codes were confronted with raw data to allow for optimal operationalization and driving examples related to the context of physical chemistry.

Step 3: Summarizing Data and Identifying Initial Themes. To reduce the measurement error that normally arises during coding and lessen the degree of interpretation in defining CUs, especially in transcripts based on performance in authentic settings, we defined CU in the CCTCM.

Using both parts of the CCTCM, CP, and CTSP allowed for the coding of all transcript statements without leaving any uncoded segments behind, thus reducing the number of measurement errors. In addition to the literature-driven codes, we read through, outlined key points, and summarized a sample coding of the raw data. This enabled us to refine our codes and form additional ones. In doing so, we also followed the approach that "a single [statement] was considered as important as those that were repeated [in the transcript]" (Fereday & Muir-Cochrane, 2006, p. 86) and considered all statements in the data when refining the CCTCM.

Step 4: Adapting/Applying the Template of Codes and Additional Coding. As suggested by Fereday and Muir-Cochrane (2006), analysis of the text at this stage was guided, but not limited, by the preliminary codes. During transcript coding, inductive codes were assigned to data segments reflecting new themes, either distinct from the predetermined codes or expanding an existing code from the manual (Fereday & Muir-Cochrane, 2006, p. 88). Applying steps 1–4 by testing the CCTCM on the raw data helped us refine and form a final set of literature- and data-driven codes. The complete coding scheme of the CCTCM (CP and CTSP) was evaluated for robust interrater reliability. For this purpose, we used one full transcript. Three individuals coded the transcript independently (the author and two external coders), and agreement was calculated using the MAXQDA software (VERBI Software, 2023). Following practical guidance from Halpin (2024), we planned intercoder agreement in several steps: We selected Cohen's kappa for dual coding subsets, set a predefined threshold of 0.75 for acceptability, and kept an audit trail to document decisions, revisions, and coder discussions. We achieved an overall agreement of 80.59%, with a Cohen's kappa of 0.81 between the author and the first coder; 87.65%, with a Cohen's kappa of 0.88 between the author and the second coder; and 84.71%, with a Cohen's kappa of 0.85 between the two coders. These figures indicate strong agreement among the coders, demonstrating the reliability of our instrument.

To support a rigorous and systematic qualitative analysis, we used MAXQDA, a CAQDAS platform recognized for enhancing transparency, traceability, and interpretive depth in qualitative coding (Brailas et al., 2023). Following the realist approach described by Madill et al. (2000), we employed multiple independent coders in our qualitative research to establish reliability. In this framework, reliability is tied to "consistency of meaning" (pp. 7–8); that is, obtaining high interrater reliability agreement indicates a shared interpretation of the data and consistent application of the CCTCM codes.

Step 5: Connecting the Codes and Identifying Themes. In this step, we used the validated CCTCM on the data and presented our findings in the Exemplary Analysis: How Does the Coding Process and Outcome Look Like section.

For this study, we did not conduct the sixth step of the Fereday and Muir-Cochrane (2006) model. This final step is concerned with corroborating and legitimating the coded themes. In their work, Fereday and Muir-Cochrane used this step to conduct a cluster analysis of the themes identified in their coding to confirm their findings before interpretation, forming an explanatory framework and drawing conclusions. In this study, we focused on the process of creating the CCTCM and ensuring its quality. Using the CCTCM to fully code collected data from control and intervention groups is the topic of our future publications (Danial et al., 2025a).

Structure of the CCTCM

Figure 3 presents the structure of the CCTCM. The full coding manual is presented in Danial et al. (2025b). Although using both parts of the CCTCM results in a contextualized analysis of the applied CT skills, each part can be used separately for a different purpose in data analysis. Each part of the CCTCM has an introduction that explains how to apply the codes and provides examples.

Figure 3

Structure and Components of the CCTCM

The Contextual Critical Thinking Coding Manual (CCTCM)

Main Introduction – how to use the coding manual and how to assign CU

Part I – CP – in the context of the physical chemistry laboratory course

Main categories of codes:

- 0_CU with incomplete meaning
- 1_Chemical terms and content
- 2_Laboratory equipment, substances and tools
- 3_Computers and digital devices
- 4_Data
- 5_Procedure
- 6_Team work administration and organization
- 7_Final report
- 8_Short answers
- 9_Instruction
- 10_Requesting support and consulting teaching staff
- 11_Others

Secondary introduction – how to proceed from CP to CTSP

Part II – CTSP – CT skills and subskills

1_CT Processing - Main categories of codes:

- 1.1_Interpretation– clarifying meaning
- 1.2_Analysis–examining arguments
- 1.3_Evaluation–assigning value to claims

2_CT Outcomes – Main categories of codes:

- 2.1_Actions and Inference – Deciding on an action and Inference by drawing conclusions
- 2.2_Self/Group -regulation, -examination, and -correction – Employing and reacting to fallacy labels
- 2.3_Explanation – presenting positions or arguments orally or in writing

3_Others

The main categories, CP_1-10 and subcodes (Figure 3), are based on observations in the laboratory, learning materials, and topics addressed in a typical laboratory course. The CTSP comprises six main categories, each corresponding to the core CT constructs and skills outlined in the Delphi report (Facione, 1990a). We further grouped the six categories into two overarching groups, namely "CT processing skills" (CTSP_1.1-1.3) and "CT outcome skills" (CTSP_2.1-2.3; Figure 3), following Ennis's (1985, 2015) dual focus and distinction between the process ("*reasonable reflective thinking*") and the outcomes ("*deciding what to believe or do*"). Our coding manual expanded this scheme to add context-relevant examples related primarily to the context of the physical chemistry laboratory and added for each code of the CTSP a secondary code that distinguishes between the students' statements and questions.

Following Braun and Clarke (2024), we distinguished between our analytic process and steps in creating the CCTCM (how the codes were generated and refined) in , and the final structure of the CCTCM, thus allowing the readers to see how we progressed during the process.

Order of Coding (CP First and Then CTSP)

The CCTCM can be used in its entirety (CP and CTSP), or each component can be applied separately to serve different aims of the assessment. However, when employing the full instrument, we first used the CP codes to categorize whether a given CU pertained to a meaningful topic in the laboratory context. CUs that involve purely procedural remarks (e.g., “Yes,” “No,” “I did this step,” or unrelated social talk) are not coded for CTSP because they do not reflect CT processes. By identifying the relevant context through the CP, we ensured that only meaningful CUs received subsequent CTSP codes, enhancing clarity, consistency, and efficiency in the final analysis.

Exemplary Analysis: How Does the Coding Process and Outcome Look Like

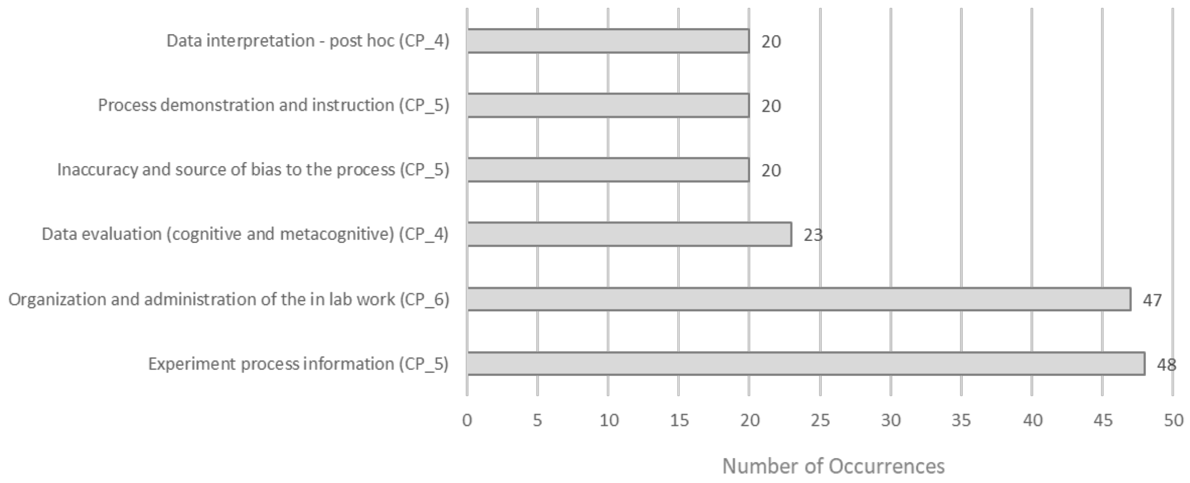
To answer RQ2, the CCTCM was applied to one full transcript of the student groups’ verbal interactions while conducting an undergraduate physical chemistry laboratory experiment. The coding was conducted manually in MAXQDA, where CP and CTSP codes could be assigned to the same segment as specified previously in the methods section. The selected student group was one of the eight groups and comprised two male students. The session lasted approximately 3.5 hours. This exemplary analysis demonstrates the utility of the developed CCTCM for illustrative, not evaluative, purposes. In our other publication, we present a comprehensive dataset in an empirical study that compared control and intervention groups (Danial et al., 2025a).

The current section presents a descriptive analysis of the coding of the full transcript and examples of the most frequent codes from CP and CTSP of the CCTCM, followed by a co-occurrence analysis of CT skills with different contexts. We then examine the overall experiment by focusing on patterns of applied topics and CT skills, as well as trends in instruction and support requests from instructors. Lastly, we present an individual CT performance analysis.

Descriptive Analysis: Most Frequent Codes

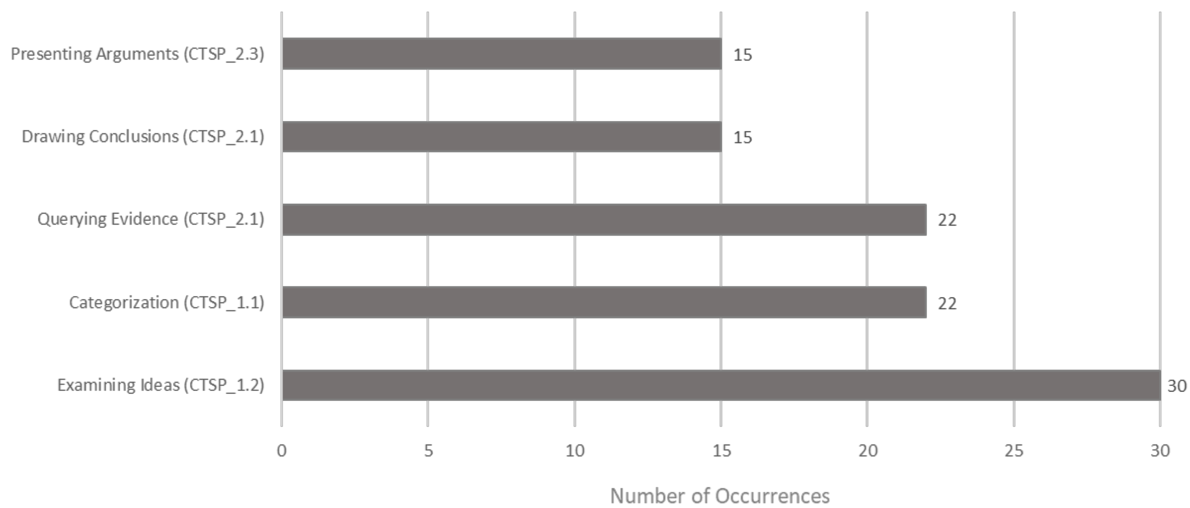
The following two figures present the most frequent codes from the CP (Figure 4) and the CTSP (Figure 5) of the CCTCM.

Figure 4
CP's Most Frequent Codes



The indentation in brackets refers to the number of the main category from the CP to which the specific code belongs. Figure 4 shows that the most frequent verbal interaction within this analysis revolves around the process of carrying out the experiment (CP_5), data evaluation and interpretation (CP_ 4), and work organization and administration (CP_ 6).

Figure 5
CTSP's Most Frequent Codes



By adding the number of occurrences in Figure 5 for the codes belonging to CTSP "CT processing" skills–Categorization (CTSP_1.1) and Examining Ideas (CTSP_1.2), we obtain 52 occurrences, equal to the sum of "CT outcome" skills occurrences–Presenting Arguments (CTSP_2.3), Drawing Conclusions (CTSP_2.1), and Querying Evidence (CTSP_2.1).

Contextualized Coding for CT Skills

Here, using two examples, we demonstrate how a specific CT skill can be applied in different contexts. We used the code co-occurrence model in MAXQDA 2024 (VERBI Software, 2023). Example 1 is for the most frequent "CT processing" skill–Examining Ideas (CTSP_1.2), and Example 2 is for the most frequent "CT outcomes" skill–Querying Evidence (CTSP_2.1), both from Figure 5.

Example 1: CTSP Skill—Examining Ideas in Context

Figure 6 demonstrates the contexts of the CT skill of examining ideas. This co-occurred with 12 contexts from the CP. Eight of the 30 occurrences happened within the context of "inaccuracy and source of bias to the process" (CP_5). Examples of CUs are as follows:

> I'm afraid of adding unintentionally a drop of water in there. That's why I poured it away. (K8_A1_352)
>
> That's why you should hold the cuvettes only at the top, so no fingerprints can influence the results. (K8_A1_782)

Figure 6

Number of Co-occurrences of the CT Skill "Examining Ideas" (N = 30) With Contexts of the Physical Chemistry Laboratory

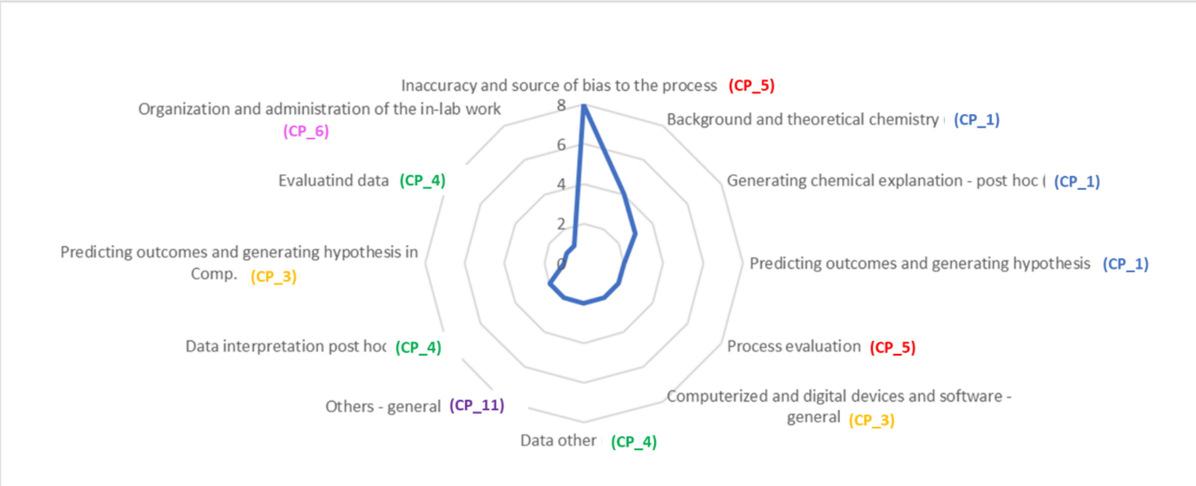


Figure 6 shows that examining ideas was applied in three major contexts from the CP. Of the 30 occurrences, 10 (33%) occurred in two subcodes related to the context of CP_5 "procedure" (red) concerned with the process of carrying out an experiment, nine (30%) occurred in three subcodes related to the context CP_1 "Chemical terms and context" (blue), and five (16%) occurred in three subcodes related to CP_4 "Data" (green).

Example 2: CTSP Skill—Querying Evidence in Context

Figure 7 demonstrates the contexts of the CT skill, Querying Evidence. This co-occurred with nine contexts from the CP. Eight of 22 occurrences occurred within the context of "Process demonstration and instructions" that belongs to CP_5 "Procedure". Examples of CUs are as follows:

> Actually, in the case of iodine, it's important that we don't titrate batchwise but rather little by little and consistently. (K8_A1_182)
>
> Please let me finish my explanation. Now it's 0,0 here. Now we have 30 seconds, and this time, you can make the reference; it always takes some time. About 20 seconds or so. Sometimes, it takes less. So, you are waiting until it reaches 0,0. And then you can take it outside and take this other cuvette inside, with the line mark always to the right side. (K8_A1_929)

Figure 7
Number of Co-occurrences of the CT Skill "Querying Evidence" (N = 22) with Contexts of the Physical Chemistry Laboratory

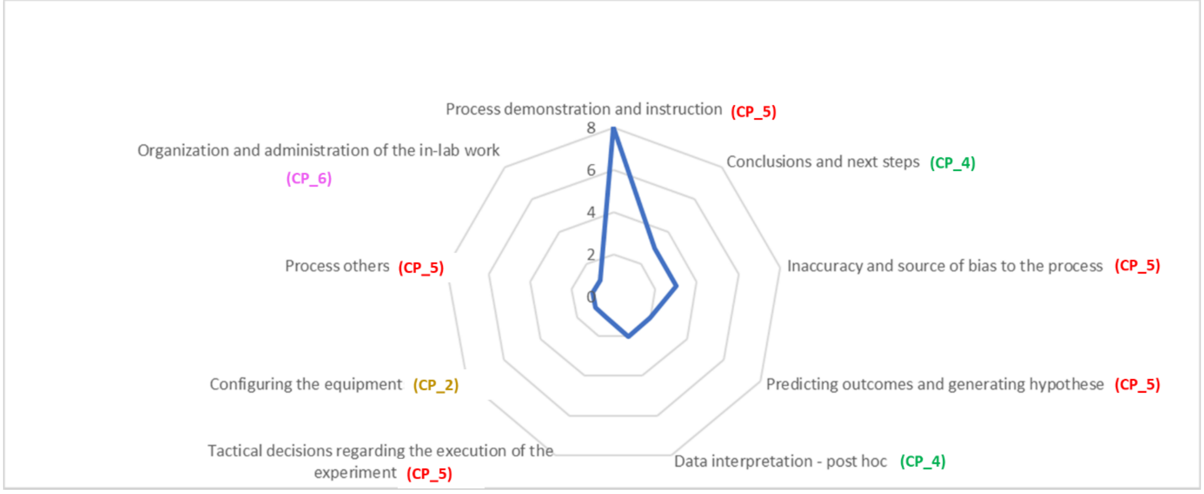


Figure 7 shows that Querying Evidence was applied in two major contexts from the CP. Of the 22 occurrences, 15 (68%) occurred in five subcodes related to the context of the CP_5 "Procedure" (red), and five (22%) occurred in two subcodes related to the context CP_4 "Data" (green).

Examples 1 and 2 (Figures 6 and 7) show that each CT skill was applied differently. For instance, although "Examining Ideas" and "Querying Evidence" co-occurred with most codes related to the CP_5 Procedure, they varied greatly within the subcodes, thus shifting focus. The first focus was mainly concerned with what could have jeopardized the accuracy of the process and what could have contributed to distorting the obtained data owing to the steps they took. The second focus was on the technicalities of doing things. Another notable distinction is that the second major context related to the skill "Examining Ideas", CP_1 Chemical terms and context, did not appear in any of the main or subcodes related to "Querying Evidence." While both CT skills had CP_4 Data as the most common context, they only had one similar subcontext (Data Interpretation),but varied in the remaining three.

This analysis of the co-occurrence of codes is essential for understanding the nature of applying CT skills in the various contexts of the physical chemistry laboratory course. It provides valuable insights into which topics tend to prompt specific CT skills and identifies those less frequently applied skills in this setting. This analysis could further highlight gaps in which certain CT skills are underutilized, providing an opportunity for targeted instructional improvements in the course.

Our analysis demonstrates examples that emphasize specific applied CT skills and examined the corresponding contexts. However, when conducting contextual analysis, in which a single context from the CP is examined on the basis of the co-occurrence with CT skills (Danial et al., 2025a), we observed that different topics also tend to engage different CT skills.

The analysis demonstrates that even from a single transcript, the CCTCM enables an in-depth exploration by answering the following key questions:

In a physical chemistry laboratory course:
a. What CT skills are applied?
b. What experimental contexts are discussed?
c. For a specific CT skill, which topics are most frequently addressed?
d. Which CT skills are being applied to a particular experimental topic?

Such analysis can provide detailed formative and summative insights, allowing a deeper understanding of individual cases, comparisons between cases (e.g., pre and post, and examining how effectively CT skills were applied), and a comprehensive overview of a course or specific context.

Flow of Topics and CT Skills: Insights

By utilizing the CCTCM, we examined both the intensity and timing of occurrences for specific topics or CT skills applied in individual cases and compare multiple cases. Figures 8 and 9 illustrate the topics (Figure 8) and CT skills (Figure 9) utilized throughout the day through their occurrences in the transcripts, highlighting variations in their application.

Figure 8
Trends in CP Codes

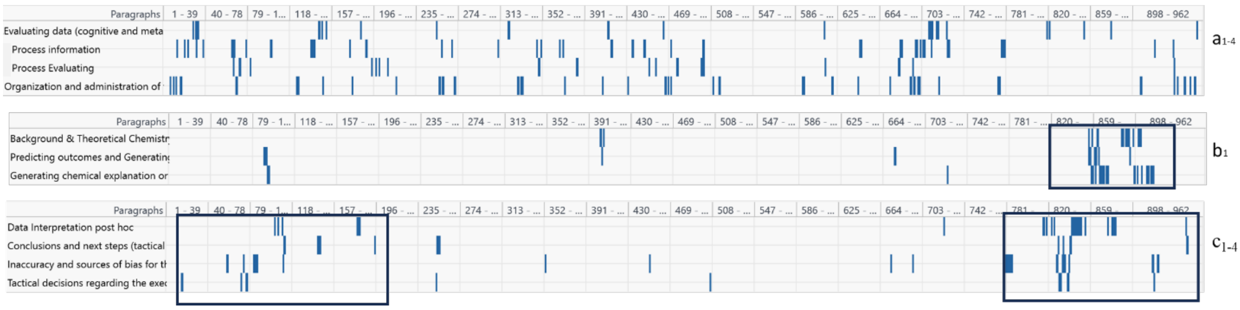


Figure 9
Trends in CTSP Codes

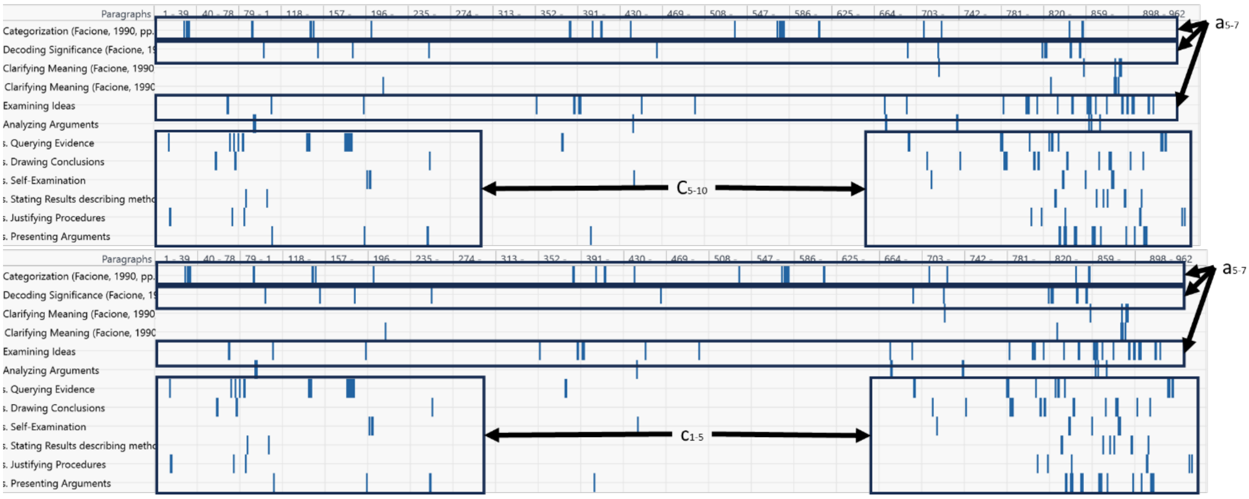
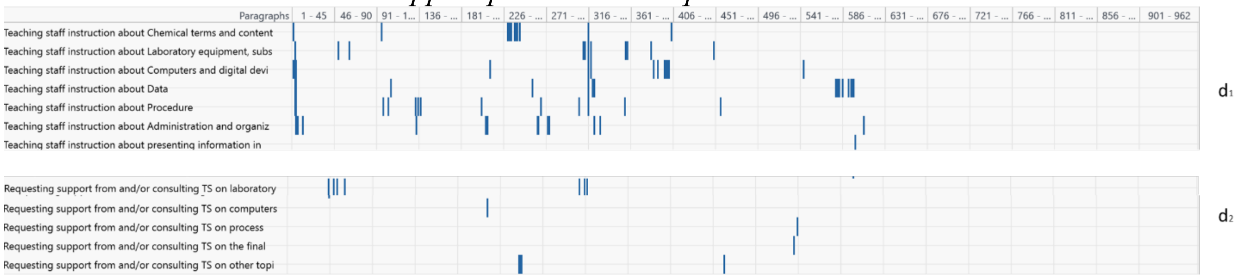


Figure 8 shows that the topics from the CP, such as Data Evaluation, Process Information, Process Evaluation, and Organization of the Work ($a_{1\text{-}4}$), were applied throughout the day in the laboratory. Similarly, the CT skills in Figure 9, which are part of the “CT processing” in CTSP, such as Categorization, Decoding Significance, and Examining Ideas ($a_{5\text{-}7}$), occurred throughout the day. However, we noticed that topics related to chemical terms and content (b_1) were more apparent toward the end of the day, and others, such as Data Interpretation, Tactical Decisions Based on Data, Process Evaluation, and Tactical Decisions Based on Process ($c_{1\text{-}4}$), appeared mostly at the beginning and end of the day. In addition, CT skills that are part of the “CT outcome” in the CTSP, such as Querying Evidence, Drawing Conclusions, Stating Results, Justifying Procedure, and Presenting Arguments ($c_{5\text{-}10}$), appeared mostly at the beginning and end of the day.

Such information can help us determine when CT skills are applied and how we can design learning tasks that effectively promote CT throughout the day. Further information that can help guide this can be supported by the course instructions. Figure 10 shows the focus of the instruction given by the teaching staff (TS) of the course (CP_9) and the incidences when students addressed the TS for support (CP_10).

Figure 10

Trends in TS Instruction and Support per Student Request



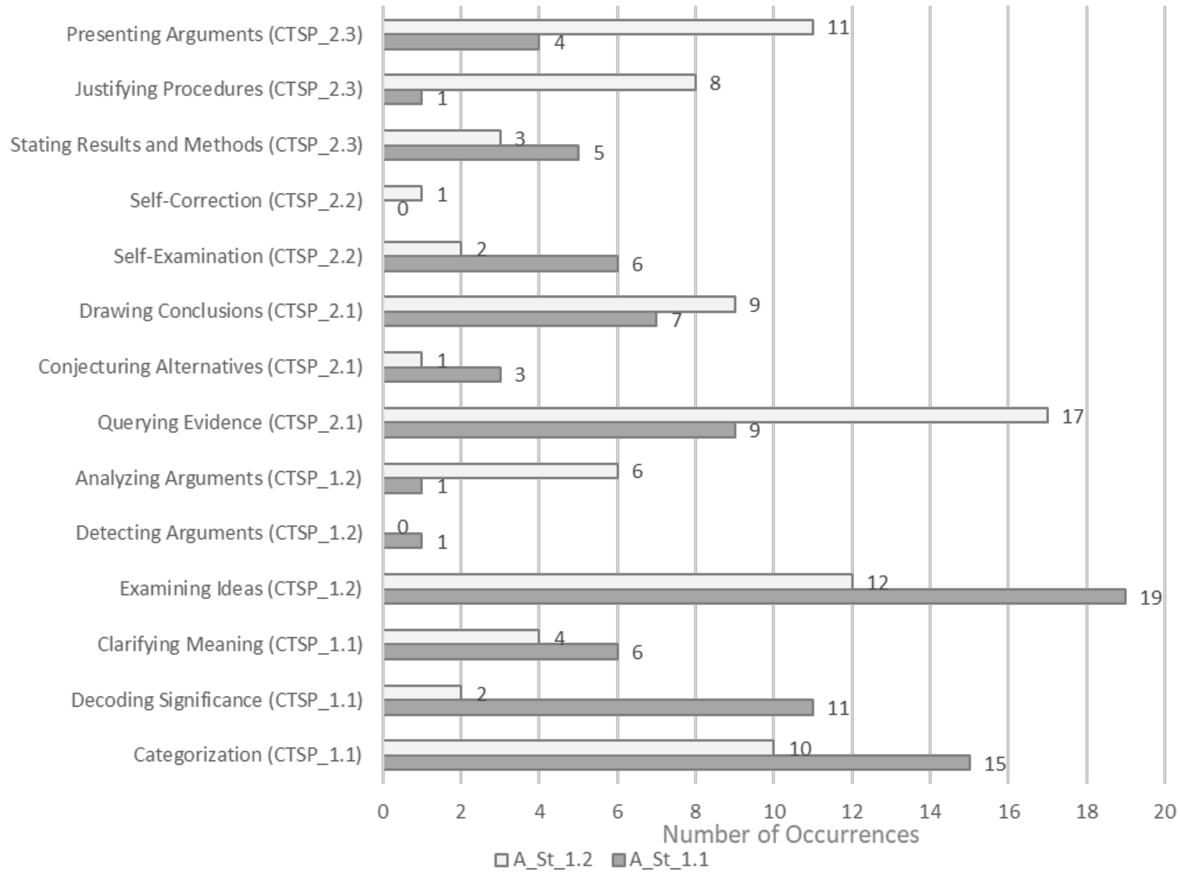
Both instructions initiated by the TS (d_1) or requested by the students (d_2) predominantly occurred during the early and mid parts of the day, as illustrated in Figure 10. Conversely, fewer instructional events were observed later in the day, coinciding with the active application of CT skills, as observed in Figure 9. This pattern suggests that a lack of TS instruction and support during these periods may negatively impact CT skill application and development.

Individual and Relative CT Performance

The CCTCM provides an in-depth analysis of individual CT performance in a collaborative setting (Figure 11).

Figure 11

Participants' Relative CT Performances



In this analysis, Student A1.1 demonstrated CT in 88 instances, while his partner, Student A1.2, did so in 86 instances. Figure 11 highlights individual areas of CT strength. While the number of "CT processing" skills was identical to "CT outcomes," both 87 occurrences in total, individual contributions differed. Student A1.1 accounted for 60% (53/87) of the "CT processing" skills, whereas Student A1.2 contributed 40%. Conversely, for "CT outcome" skills, Student A1.2 performed 60%, while Student A1.1 contributed 40%, showing a complementary dynamic in their CT performance. Some of the most prominent contexts related to CT in which Student A1.1 participated were data evaluation and interpretation (CP_4), providing chemical background and explanation pre and post hoc (CP_1), and process evaluation (CP_5). Inaccuracy in the process (CP_5) and data evaluation and interpretation (CP_4) were most prominent in Student A1.2's performance.

Conclusions and Outlook

We conceptualize CT as a reasoned, reflective thought process that informs judgment and action in a given context (Ennis 2015; Facione, 1990a). In particular, we emphasize the importance of interpreting information, analyzing arguments, and evaluating evidence (referred to in this work as CT processing skills), as well as inferring conclusions, explaining reasoning, and monitoring or correcting one's own thinking (termed in this work as CT outcomes skills). These skills are particularly important in the context of a physical chemistry laboratory, as students must not only process information but also determine how best to act by alternating procedures by adjusting an experimental procedure, dismissing flawed data, and proposing new steps. The APA Delphi report (Facione, 1990a) breaks down CT skills into six core and subskills that form, to this day, the basis for defining CT skills (Andreucci-Annunziata et al., 2023; Butler, 2024; Heard et al., 2020). We group the six core CT skills following Ennis's

(1985, 2015) dual focus and distinction between the process ("*reasonable reflective thinking*") and the outcomes ("*deciding what to believe or do*"). We map interpretation, analysis, and evaluation to "CT processing skills," reflecting how students analyze information, question assumptions, and judge credibility or relevance. Inference, self-regulation, and explanation align with "CT outcomes," capturing how students transition from critical analysis to purposeful concrete action, such as querying evidence and altering methods, articulating coherent justification for their decisions, and revising their understanding. Our view of CT and this two-part structure also reflect Halpern's (2013, 2014) view of CT as a purposeful, context-sensitive, goal-directed set of cognitive strategies.

Different disciplines emphasize distinct aspects of CT (Liu et al., 2001). We argue that assessing applied CT requires doing so within the same domain using performance assessment, which offers the strongest face validity for capturing CT in practice. Although performance assessment is often criticized for "reveal[ing] only what the situation requires" (Ennis, 1993, p. 185), this aligns with our goal in this study of assessing domain-specific CT in a physical chemistry laboratory through authentic students' observable, verbalized reasoning. This approach aligns with the interpretive perspective in qualitative research, which seeks to establish meaning from participants' views and emphasizes understanding phenomena as they naturally occur in specific social contexts (Chowdhury & Shil, 2021).

The CCTCM was developed to enable the performance-based assessment of domain-specific CT in a physical chemistry laboratory context. Its codes are grounded in key CT frameworks (e.g., Butterworth & Thwaites, 2013; Ennis, 1985a, 1985b, 2015; Facione, 1990a, 2015; Heard et al., 2020), addressing concerns around construct validity in existing open-ended tools. Following Fereday and Muir-Cochrane's (2006) model, the development process ensured methodological rigor and minimized subjectivity. The two components of the CCTCM (CP and CTSP) allow for contextualized analysis of how CT manifests in student interactions. In addition to the contextualized CT analysis, the CCTCM generates data that (1) capture real-time verbalized CT performance, (2) are true to the learning process and not just an end-point result, and (3) allow qualitative analysis and quantification, even in the case of a small sample size. The CCTCM and its features offer nuanced insights into the interplay between experimental tasks and cognitive processes, which are often missed by standardized assessments, by highlighting the different contexts that can potentially prompt the application of a specific CT skill and, alternatively, the less common CT skills. Unlike tools such as the Danczak–Overton–Thompson (DOT V3) and Chemistry Critical Thinking Test (Danczak et al., 2020), which uses structured scenarios and assesses CT in the broader context of chemistry, the CCTCM focuses on authentic verbal interactions within the discipline, ensuring richer, individualized, and more contextually valid understanding of CT skills. CT is neither a single monolithic set of skills nor necessarily a comprehensive checklist that every student must master in identical ways. Instead, it can manifest differently depending on the specific goals, contexts, and disciplinary needs of the course (Moore, 2011). Therefore, the CCTCM provides a modular framework that allows instructors or researchers to select or prioritize selected skills rather than mandating the entire code set. This flexibility recognizes that teaching goals vary across courses, disciplines, and even institutions, enabling educators to identify which skills students already apply regularly and which may need further development. The CCTCM can be used as an evidence-based tool to inform content-relevant instructional strategies and curriculum development that ensures that students have authentic opportunities to practice, engage in, and demonstrate CT. The data obtained from the CCTCM and additional measures as part of our broader research allow for further in-depth examination and triangulation of students' CT from what emerges in spoken dialogue and beyond (Danial et al., 2025a). Triangulation, whether within or between methods, can strengthen the validity of inferences drawn from the CCTCM by comparing student reasoning across different forms of data, such as spoken dialogue, written responses, and performance artifacts (Chowdhury & Shil, 2021).

The CCTCM's independence from specific teaching strategies, participant familiarity, and preset problems ensures minimal interference with natural thinking processes and preserves the authenticity of the observed CT. Lastly, its flexible design offers conceptual guidance for assessing CT in other STEM domains by adapting its CP to suit different disciplinary settings. However, its current empirical validation is limited to the context of undergraduate physical chemistry laboratories. Nonetheless, this adaptability has strong potential to inform future research on CT and related cognitive processes across science education contexts.

Limitations

Despite the strength of using the CCTCM to capture CT in an authentic physical chemistry laboratory context, several limitations warrant consideration. First, relying on verbalized discourse captures only a subset of the students' cognitive processes. Although transcription protocols instruct students to note information on observed behaviors (e.g., laughing, head nodding, and moving away from the experiment station), students may engage in internal reasoning that remains unrecorded in transcripts. Consequently, while code frequencies can be useful for highlighting patterns, they do not necessarily indicate mastery or exclusive use of a particular skill. They may also be influenced by group dynamics, personal traits, or signal persistent difficulties within the situation.

Second, the naturalistic performance assessment design of this study provides strong face validity but inevitably "reveals only what the situation requires" (Ennis, 1993, p. 185). This approach does not account for tasks that might, for instance, elicit a distinct set of CT skills. One countermeasure is to employ a pre-post design (with a control group and an elicited CT skills experiment group) or incorporate validated tools to help control variables in participant familiarity, instructional strategies, personal traits, and group dynamics.

Third, the CCTCM's focus on real-time laboratory interactions captures in-the-moment decision-making rather than reflective or retrospective forms of CT. While this offers valuable insights into students' immediate thought processes and reasoning for the next steps, it also entails triangulation via reflective journals, interviews, or other tools to deepen and broaden the analysis. Future research should examine how integrating multiple data sources can enrich our understanding of students' engagement in CT, whether verbally or nonverbally, immediately or retrospectively, and across diverse learning environments.

Fourth, this study was focused on the development and preliminary application of the CCTCM. As such, its scope is intentionally limited to an illustrative analysis rather than a full-scale empirical validation. The findings from the coded session are meant to demonstrate the tool's structure, usability, and alignment with real laboratory discourse, not to offer statistically generalizable claims. More extensive empirical application is required to establish reliability and validity across groups and settings. These follow-up studies are ongoing and discussed in a separate paper.

Lastly, despite conducting intercoder calibration to ensure consistency in applying the coding manual, some degree of subjectivity in interpreting and assigning codes might have persisted. Qualitative coding inherently involves interpretive judgment, especially when analyzing nuanced verbal interactions. While we established a shared understanding of the coding categories through iterative discussions and test rounds, subtle differences in how the coders interpreted context or intention might have influenced the coding decisions. This potential subjectivity does not invalidate the findings but highlights the importance of continued coder training, transparent documentation of coding procedures, and, where possible, triangulation with additional data sources to further strengthen the reliability of the analysis. Recent developments in qualitative coding methodology, such as metaphor coding and categorical theming (Saldaña, 2021), as described in a recent review (Lungu, 2022), provide different approaches for richer qualitative interpretation beyond traditional first-cycle schemes.

These techniques may inform future adaptations of the CCTCM to capture more nuanced reasoning patterns in student discourse.

Acknowledgment

We are especially thankful for the funding by the German Research Foundation (DFG GSC 1013 SALSA) for the School of Analytical Sciences Adlershof (SALSA), where this research was conducted. We express our gratitude to all participating students and faculty members, especially Dr. Wolfgang Christen and the course's teaching staff, who contributed to the data collection process. Their support and collaboration were invaluable to the success of this project.

References

Andreucci-Annunziata, P., Riedemann, A., Cortés, S., Mellado, A., del Río, M. T., & Vega-Muñoz, A. (2023). Conceptualizations and instructional strategies on critical thinking in higher education: A systematic review of systematic reviews. *Frontiers in Education, 8*, 1-18. https://doi.org/10.3389/feduc.2023.1141686

Bagnato, S. J. (2007). *Authentic assessment for early childhood intervention: Best practices*. Guilford Press.

Bailin, S. (2002). Critical thinking and science education. *Science & Education, 11*(4), 361–375. https://doi.org/10.1023/A:1016042608621

Black, B. (2012). An overview of a program of research to support the assessment of critical thinking. *Thinking Skills and Creativity, 7*(2), 122–133. https://doi.org/10.1016/j.tsc.2012.04.003

Black, B., Chislett, J., Thompson, A., Thwaites, G., & Thwaites, J. (2008). *Critical thinking: A definition and taxonomy for Cambridge Assessment*. Cambridge Assessment. https://doi.org/10.17863/CAM.100474

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage.

Brailas, A., Tragou, E., & Papachristopoulos, K. (2023). Introduction to qualitative data analysis and coding with QualCoder. *American Journal of Qualitative Research*, *7*(3), 19–31. https://doi.org/10.29333/ajqr/13230

Braun, H. I., Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Borowiec, K. (2020). Performance assessment of critical thinking: Conceptualization, design, and implementation. *Frontiers in Education, 5*, 1-10. https://doi.org/10.3389/feduc.2020.00156

Braun, V., & Clarke, V. (2024). Reporting guidelines for qualitative research: A values-based approach. *Qualitative Research in Psychology, 22*(2), 399–438. https://doi.org/10.1080/14780887.2024.2382244

Butler, H. A. (2024). Predicting everyday critical thinking: A review of critical thinking assessments. *Journal of Intelligence, 12*(2), 1-12. https://doi.org/10.3390/jintelligence12020016

Butterworth, J., & Thwaites, G. (2013). *Thinking skills: Critical thinking and problem solving* (2nd ed.). Cambridge University Press.

Carter, A. G., Creedy, D. K., & Sidebotham, M. (2015). Evaluation of tools used to measure critical thinking development in nursing and midwifery undergraduate students: A systematic review. *Nurse Education Today, 35*(7), 864–874. https://doi.org/10.1016/j.nedt.2015.02.023

Chowdhury, A., & Shil, N. C. (2021). Thinking 'qualitative' through a case study: Homework for a researcher. American *Journal of Qualitative Research*, *5*(2), 190–210. https://doi.org/10.29333/ajqr/11280

Crabtree, B., & Miller, W. (1999). A template approach to text analysis: Developing and using codebooks. In B. Crabtree & W. Miller (Eds.), *Doing qualitative research* (pp. 163–177). Sage.

Danczak, S. M., Thompson, C. D., & Overton, T. L. (2020). Development and validation of an instrument to measure undergraduate chemistry students' critical thinking skills. *Chemistry Education Research and Practice, 21*(1), 62–78. https://doi.org/10.1039/C8RP00130H

Danial, L., Koenen, J., & Tiemann, R. (2025a). *Prompting and assessing critical thinking in undergraduate physical chemistry labs: Effects of instructional design and performance in-situ assessment*. ChemRxiv. doi: 10.26434/chemrxiv-2025-ms8cp

Danial, L., Koenen, J., & Tiemann, R. (2025b). *The Contextual Critical Thinking Coding Manual (CCTCM) - The complete coding manual.* Zenodo. https://doi.org/10.5281/zenodo.16755624.

Dominguez, C. E., Dumitru, D., Bigu, D., Elen, J., Jiang, L., Railiene, A., Penkauskiene, D., Papathanasiou, I. V., Tsaras, K., Fradelos, E. C., & Ahern, A. (2018). *A European collection of the critical thinking skills and dispositions needed in different professional fields for the 21st century*. UTAD.

Ennis, R. H. (1985a). A logical basis for measuring critical thinking skills. *Educational Leadership, 43*(2), 44–48.

Ennis, R. H. (1985b). Critical thinking and the curriculum. *National Forum, 65*(1), 28–31.

Ennis, R. H. (1985c). Tests that could be called critical thinking tests. In A. L. Costa (Ed.), *Developing minds: A resource book for teaching thinking* (pp. 303–304). Association for Supervision and Curriculum Development.

Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher, 18*(3), 4–10. https://doi.org/10.3102/0013189X018003004

Ennis, R. H. (1990). The extent to which critical thinking is subject-specific: Further clarification. *Educational Researcher, 19*(4), 13–16. https://doi.org/10.3102/0013189X019004013

Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice, 32*(3), 179–186. https://doi.org/10.1080/00405849309543594

Ennis, R. H. (2015). Critical thinking: A streamlined conception. In M. Davies & R. Barnett (Eds.), *The Palgrave handbook of critical thinking in higher education* (pp. 31–47). Palgrave Macmillan US. https://doi.org/10.1057/9781137378057_2

European Commission, Directorate-General for Education, Youth, Sport and Culture. (2019). *Key competences for lifelong learning*. Publications Office of the European Union. https://doi.org/10.2766/569540

Facione, P. A. (1990a). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction* (The Delphi Report). California Academic Press.

Facione, P. A. (1990b). *The California Critical Thinking Skills Test (CCTST): Forms A and B; and the CCTST test manual*. California Academic Press.

Facione, P. A. (2000). The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill. *Informal Logic, 20*(1), 61–84. https://doi.org/10.22329/il.v20i1.2254

Facione, P. A. (2015). *Critical thinking: What it is and why it counts* (2015 update). Insight Assessment. https://www.insightassessment.com

Facione, P. A., & Facione, N. C. (1992). *The California Critical Thinking Disposition Inventory (CCTDI)*. California Academic Press.

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 80–92. https://doi.org/10.1177/160940690600500107

Ferreira, L. D., & Merchant, K.A. (1992). Field research in management accounting and control: A review and evaluation. *Accounting, Auditing and Accountability Journal*, *5*(4), 3-34. https://doi.org/10.1108/09513579210019503

Fischer, S. C., Spiker, V. A., & Riedel, S. L. (2009). *Critical thinking training for army officers: A model of critical thinking* (Vol. 2, Technical Report). U.S. Army Research Institute.

Halpern, D. F. (2013). *Thought and knowledge: An introduction to critical thinking* (5th ed.). Psychology Press. https://doi.org/10.4324/9781315885278

Halpern, D. F. (2014). *Critical thinking across the curriculum: A brief edition of thought and knowledge*. Routledge. https://doi.org/10.4324/9781315805719

Halpin, S. N. (2024). Inter-coder agreement in qualitative coding: Considerations for its use. *American Journal of Qualitative Research, 8*(3), 23–43. https://doi.org/10.29333/ajqr/14887

Heard, J., Scoular, C., Duckworth, D., Ramalingam, D., & Teo, I. (2020). *Critical thinking: Skill development framework*. Australian Council for Educational Research. https://research.acer.edu.au/ar_misc/41

Jabali, O., Hamamra, B., & Ayyoub, A. (2024). Critical thinking, assessment, and educational policy in Palestinian universities. *International Journal for Educational Integrity, 20*(1), 1-23. https://doi.org/10.1007/s40979-024-00160-9

Jones, E. A., Hoffman, S., Moore, L. M., Ratcliff, G., Tibbetts, S., & Click, B. A. (1995). *National assessment of college student learning: Identifying college graduates' essential skills in writing, speech and listening, and critical thinking. Final project report*. U.S. Government Printing Office. https://files.eric.ed.gov/fulltext/ED383255.pdf

Kennedy, M., Fisher, M. B., & Ennis, R. H. (1991). Critical thinking: Literature review and needed research. In L. Idol & B. F. Jones (Eds.), *Educational values and cognitive instruction: Implications for reform* (pp. 11–40). Lawrence Erlbaum Associates.

Ku, K. Y. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity, 4*(1), 70–76. https://doi.org/10.1016/j.tsc.2009.02.001

Lai, E. R. (2011). Critical thinking: A literature review. *Pearson's Research Reports*, *6*(1), 40–41.

La Marca, P. M., Redfield, D., Winter, P. C., & Despriet, L. (2000). *State standards and state assessment systems: A guide to alignment. Series on standards and assessments*. Council of Chief State School Officers.

Liu, G., Long, S., & Simpson, M. E. (2001). Self-perceived gains in critical thinking and communication skills: Are there disciplinary differences? *Research in Higher Education, 40*, 43–60. https://doi.org/10.1023/A:1018722327398

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series, 2014*(1), 1–23. https://doi.org/10.1002/ets2.12009

Lungu, M. (2022). The coding manual for qualitative researchers. *American Journal of Qualitative Research*, *6*(1), 232–237. https://doi.org/10.29333/ajqr/12085

Madill, A., Jordan, A., & Shirley, C. (2000). Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies. *British Journal of Psychology, 91*(1), 1–20. https://doi.org/10.1348/000712600161646

McPeck, J. (1981). Critical thinking and education. New York: St. Martin's.

McPeck, J. E. (1990). Critical thinking and subject specificity: A reply to Ennis. *Educational Researcher, 19*(4), 10–12. https://doi.org/10.3102/0013189X019004010

Moore, T. (2011). Critical thinking: Seven definitions in search of a concept. *Studies in Higher Education, 38*(4), 506–522. https://doi.org/10.1080/03075079.2011.586995

Norris, S. P. (1989). Can we test validly for critical thinking? *Educational Researcher, 18*(9), 21–26. https://doi.org/10.3102/0013189X018009021

Norris, S. P., & Ennis, R. H. (1989). *Evaluating critical thinking*. Midwest Publications.

Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Sage.

Paul, R. W., Elder, L., & Bartell, T. (1997). *California teacher preparation for instruction in critical thinking: Research findings and policy recommendations*. Foundation for Critical Thinking.

Pithers, R. T., & Soden, R. (2000). Critical thinking in education: A review. *Educational Research, 42*(3), 237–249. https://doi.org/10.1080/001318800440579

Saldaña, J. (2021). *The coding manual for qualitative researchers*. Sage.

Schutz, A. (1967). The phenomenology of the social world (G. Walsh & F. Lehnert, Trans.). Evanston, IL: North Western University Press.

Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Marino, J. P. (2018). International performance assessment of learning in higher education (iPAL): Research and development. In O. Zlatkin-Troitschanskaia, M. Toepell, & R. J. Shavelson (Eds.), *Assessment of learning outcomes in higher education: Cross-national comparisons and perspectives* (pp. 193–214). Springer. https://doi.org/10.1007/978-3-319-74338-7_10

Stupple, E. J. N., Maratos, F. A., Elander, J., Hunt, T. E., Cheung, K. Y. F., & Aubeeluck, A. V. (2017). Development of the Critical Thinking Toolkit (CriTT): A measure of student attitudes and beliefs about critical thinking. *Thinking Skills and Creativity, 23*, 91–100. https://doi.org/10.1016/j.tsc.2016.11.007

Thomas, K., & Lok, B. (2015). Teaching critical thinking: An operational framework. In M. Davies & R. Barnett (Eds.), *The Palgrave handbook of critical thinking in higher education* (pp. 93–105). Palgrave Macmillan US. https://doi.org/10.1057/9781137378057_6

VERBI Software. (2023). *MAXQDA 2024* [Computer software]. https://www.maxqda.com

Voogt, J., & Roblin, N. P. (2010). *21st century skills: Discussion paper. Kennisnet, 23*(3), 1–40.

Wells, I., Burton, A., & Burton, E. (2006). The OCR AS level examination as a means of assessing the critical thinking skills of undergraduate students: A pilot study. *Psychology Learning and Teaching,* *5*(1), 32–36. https://doi.org/10.2304/plat.2005.5.1.32

Willingham, D. T. (2007). Critical thinking: Why is it so hard to teach? *American Educator, 31*(2), 8–19.

Notes on Contributors

Lilian Danial is a high school science teacher, science education researcher, and PhD fellow at Humboldt University in Berlin. She holds degrees in Bachelor of Science in Molecular Biochemistry and Bachelor and Master of Science in Science Education from the Technion–Israel Institute of Technology. Her research investigates how science education can be made more effective, engaging, and socially relevant. In her current work, she focuses on fostering and assessing critical thinking in undergraduate physical chemistry laboratories. Her research interests are science education, education for sustainability, and chemistry education at schools and universities.

Prof. Dr. Jenna Koenen is an associate professor of Chemistry Education at the Technical University of Munich. Her research focuses on scientific problem-solving and critical

thinking in secondary and higher education, digital learning environments, and simulation-based instruction. She also develops and evaluates innovative teaching and learning materials to enhance science education. Her work bridges empirical research and practical applications, with the aim of improving learning outcomes through evidence-based strategies.

Prof. Dr. Rüdiger Tiemann is a full professor of chemistry education research at Humboldt-Universität zu Berlin, Germany. His current research interests include decision making, critical thinking, mental modeling of scientific ways of thinking and working, AI-supported self-regulated learning, and questions of situational cognition in the perception and classification of learning situations or in interaction with personality traits, both in secondary and higher education. In digital settings, various possibilities are being developed and evaluated to combine fundamental empirical research on these topics with applications for educational practice in schools and universities.

ORCID

Lilian Danial, https://orcid.org/0009-0004-4472-0517
Prof. Dr. Jenna Koenen, https://orcid.org/0000-0002-3591-617X
Prof. Dr. Rüdiger Tiemann, https://orcid.org/0000-0002-9398-8710