Challenging aspects of critical thinking

A mixed-methods study of students' test results, students' reasoning, and teaching strategies

Vegard Havre Paulsen

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2022

UNIVERSITY OF BERGEN

Challenging aspects of critical thinking

A mixed-methods study of students' test results, students' reasoning, and teaching strategies

Vegard Havre Paulsen

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 21.04.2022

Year: 2022

Title: Challenging aspects of critical thinking

Name: Vegard Havre Paulsen

Print: Skipnes Kommunikasjon / University of Bergen

Scientific environment

This PhD dissertation was written at the University of Bergen, Faculty of Mathematics and Natural Sciences, Department of Physics and Technology. All work was done within the ARGUMENT project, a collaboration between the University of Bergen, Western Norway University of Applied Sciences, and the City of Bergen. In addition, several of the courses I have completed were provided by the Western Norway Graduate School of Educational Research II (WNGER II).

This work was supported by the Research Council of Norway under grant number 275835.

UNIVERSITY OF BERGEN

CITY OF
BERGEN

Western Norway
University of
Applied Sciences

The Research Council
of Norway

ARGUMENT

Acknowledgements

I feel very grateful for having been given the opportunity to do this PhD project over the last few years. This dissertation is the result of countless hours spent reading papers, discussing ideas, thinking, and writing—at any hour of the day (or night). I have genuinely loved it. None of this would be possible without the support of many people, only some of which I will be able to explicitly mention here.

I would like to thank the University of Bergen for the opportunity to pursue my research interests. Furthermore, this research was made possible through the financial support of the Research Council of Norway. Thank you.

I would like to thank my main supervisor, Stein Dankert Kolstø, for providing me with well-timed support and guidance. Your work capacity, knowledge, wisdom, and social aptitude have been of immense value for me during these years, and I find these traits truly inspiring. I would also like to thank my co-supervisor, Erik Knain, for agreeing to help me become a PhD. Furthermore, I would like to thank the participants in our sporadic research-group seminars filled with interesting and enjoyable discussions: Matthias Gregor Stadler, Mildrid Kyte, Vegard Gjerde, and Stein Dankert Kolstø.

I would like to thank Bergen municipality for agreeing to be the owner of the ARGUMENT project. Importantly, I would like to thank the person in charge of managing the project, Janneke Tangen. Your leadership seemed effortless and natural and has been essential for making the project run as smoothly as it did. Many thanks go to the researchers that have been a part of the project, including Stein Dankert Kolstø, Matthias Gregor Stadler, Idar Mestad, Johan Lie, Mette Andresen, Rune Herheim, Suela Kacerja, and Helge Drange. Thank you to the people at Skolelaboratoriet—especially Brage Førland, Olaug Vett Kvam, and Frede Thorsheim—for developing resources and websites as well being part of the development and execution of the project every step of the way. Finally, I would like to thank the schools, teachers, and students that were part of the project. It has been a

true pleasure to meet so many of you in the workshops and in the schools, and to get to observe and collect data in the classrooms.

I would like to thank my good friend, Vegard Gjerde, for notifying me about this PhD position and encouraging me to apply for it. The choice to take our friendship to the next level by becoming colleagues and co-inhabitants of an office space was a good one. Thank you for providing guidance, intellectual stimuli, good discussions, absurdity, and good laughs.

I would also like to thank my loved ones, family (especially my mother for weekly check-up calls), friends, and acquaintances for supporting me and reminding me of the world outside of academia.

Bergen, January 2022

Vegard Havre Paulsen

Abstract

There is a growing focus on critical thinking throughout the education system. Overall, the efforts that have been made to improve students' critical thinking have not yielded the desired results. This could indicate a need for more research concerning which specific aspects of critical thinking that are challenging for students and what barriers students face in this regard. Furthermore, this could also indicate a need for more research on effective strategies for teaching critical thinking, and, more specifically, the details that make these strategies effective in some cases and not in others.

The aim of this dissertation is to provide pragmatic (i.e., useful) knowledge of students' struggles with critical thinking problems. This could hopefully lead to further integration of the insights from critical thinking research into teaching practice. Furthermore, the aim of this dissertation is to investigate whether supposedly effective yet general strategies for teaching critical thinking are effective in the context of lower secondary classrooms.

In this research project, we have used a well-known test of critical thinking skills, together with a modified version of this test, to identify students' challenges concerning their use of critical thinking skills. The modified version of the test includes written justifications to selected multiple-choice items from the test. This could give insights into students' reasoning when facing these items and indicate certain skills and knowledge that should be of particular focus in instruction. Moreover, we measured the effect of the ARGUMENT project—which included supposedly effective strategies for teaching critical thinking—on students' performance on the test.

In Phase 1 of the study in Article I, we quantitatively explored lower secondary students' test results on the unmodified critical thinking test, and tentatively identified challenging items. The items were then qualitatively analyzed and divided into five categories based on their proposed solution strategies (i.e., from the test manual). According to the analyses, three categories of items were particularly challenging for

these students. First, the items that required students to discern between observations and inferences seemed to be the most challenging. Second, many students struggled with the items requiring that test takers recognize a conflict of interest and take that into account when evaluating the credibility of sources and statements. Third, some students also struggled with the items requiring that test takers recognize that certain methods of observation are better than others. In Phase 2 of the study, we administered the modified test which asked for written justifications to selected multiple-choice items from these tentatively identified difficult categories of items. We have not seen any previously published studies that have used this method. The results from the modified test support the hypothesis that these items are challenging, and, importantly, that the challenges relate to the required critical thinking skills.

In Article II, we conducted a thematic analysis of the written justifications from the dataset we collected in Phase 2 of Article I. We identified six overarching general themes of reasoning encompassing 21 sub-themes. In sum, more than a quarter of the responses expressed strong inductive logic yet contained incorrect reasons because the premises used were either based on alternative evidence or were made up by students who introduced elements not originally included in the context of the items. Only a few responses did not express strong inductive logic. Most of these were responses from students who seemed to believe that an inference is just as, or more, believable than an observation. We discuss potential barriers to critical thinking that students seemed to face when working with the test items, and how these barriers relate to skills, dispositions, knowledge, and motivation.

In Article III, we conducted a quasi-experimental study comparing the gain in critical thinking test scores of the lower-secondary students in the ARGUMENT project with a control group. Teachers in the schools within the ARGUMENT project worked with researchers to develop and implement inquiry-based teaching methods with a focus on scientific argumentation and critical thinking in the context of socioscientific issues. The project aspired to implement general strategies for teaching critical thinking which have been found effective in previous research. Students in the three treatment schools and the three non-treatment schools improved

their critical thinking scores significantly from pretest to posttest. However, we did not find a difference in the gain in scores between the two groups. The article discusses potential reasons for this, including the theoretical rationale used in the research and the degree to which the implementation of the project aligns with this rationale. Importantly, we also suggest that the strategies for teaching critical thinking found in the literature could be too general. The article proposes potential avenues that should be explored in future research. In particular, the discussion of the results indicates that there is a need for more detailed insights into the characteristics of the types of authentic inquiry, dialogue, explication of critical thinking principles, and teacher training that are effective in improving critical thinking.

The findings and tentative conclusions from the first two articles could contribute to the literature on critical thinking instruction by providing preliminary insights into which aspects of critical thinking that might be particularly difficult for secondary students, as well as how these students reason when faced with critical thinking problems representing these aspects. With time, especially if future research is able to further validate these conclusions, these insights could indicate which aspects of critical thinking that should be the focus of instruction. Moreover, the insights from Article III might aid instruction and design of other projects with similarities to the ARGUMENT project. Further research on how to explicate critical thinking principles, for example through dialogue and scaffolds for inquiry into authentic issues, could then make use of the insights (from the first two articles) into which aspects of critical thinking that should be the focus of such explication.

List of Publications

Paulsen, V. H., & Kolstø, S. D. (submitted). Aspects of critical thinking that are challenging for students – Conflict of interest, observation, and inference.

Paulsen, V. H., & Kolstø, S. D. (2022). Students' reasoning when faced with test items of challenging aspects of critical thinking. *Thinking Skills and Creativity*, *43*(March 2022), Article 100969. https://doi.org/10.1016/j.tsc.2021.100969

Paulsen, V. H., & Kolstø, S. D. (in review). Large-scale study suggests supposedly effective strategies for teaching critical thinking might be too general.

Author statement

The following author contributions apply to all three enclosed articles.

Vegard Havre Paulsen: Research ideas, methodology, data analysis, data collection, data curation, data presentation, and writing the manuscript.

Stein Dankert Kolstø: Supervision, project administration, funding acquisition, and providing comments on all manuscripts.

In addition, for Article I, Stein Dankert Kolstø took part in the data analysis in a procedure for testing interrater reliability.

Contents

PART TWO | ARTICLES

PART ONE | SYNOPSIS

1. Introduction

> *Everything we know, believe, want, fear, and hope for, our thinking tells us. It follows, then, that the quality of our thinking is the primary determinant of the quality of our lives.*
>
> Richard Paul (1937-2015)

During the years before I started my PhD, I spent some of my time contemplating what type of work and activities I could do that would be the most meaningful and rewarding. This led me to a thought experiment inspired by the writer Alan Watts. He proposes that one asks oneself: How would I spend my time if money was no object? In other words, what would I trade my time for if there were no financial incentives? The ideas resulting from this introspection included several things I was already doing, like feeding my curiosity with a broad selection of knowledge, including about how the mind works and how to improve one's thinking. I also wanted to combine such learning with creating and sharing something that could help others, which would also have the convenient and symbiotic side effect of deepening my own understanding. During the years of working with my PhD, most of my days have consisted of these types of activities.

The PhD-position that I have filled was established within the ARGUMENT project (argument.uib.no/), which was developed in a collaborative effort between researchers, municipal actors, and three lower-secondary schools in Bergen. The project aimed to facilitate teachers' professional development through focusing on the use of inquiry-based teaching methods, and to improve students' scientific argumentation and critical thinking through using real data in the context of socioscientific issues. Students were encouraged to grapple with real data and to gather their own data—both of which promote learning through first-hand experience. In addition, the use of authentic and complex issues required that students considered differing views, weighed evidence, and formulated arguments representing their stance (Kolstø, 2001). Moreover, the project sought to help

teachers encourage students to develop their own tentative ideas through dialogue and the use of contrasting data (Schwartz et al., 2011).

When reading the project description of ARGUMENT, I was intrigued by the potential positive effect this project could have on students' critical thinking, and, in turn, on their quality of life. Importantly, providing teachers and students with the respective resources needed to teach and learn how to critically interpret and evaluate complex information could lead to the development of citizens capable of critical and independent thinking. While a society with few independent and critical thinkers is vulnerable to control by opportunistic and powerful forces, a society that teaches its young to become better thinkers might unleash the full power of free democracy.

The project evolved through one year of trial runs (i.e., fall 2018 and spring 2019), where the project team and teachers planned and implemented five modules, each based on a socioscientific issue. Several researchers (including me) observed and recorded in multiple classrooms, and the modules were improved before the main implementation (i.e., fall 2019 and spring 2020). Together with a few others from the project team, I developed teaching materials and resources that teachers and students could use for two of the modules: *sleep* (argument.uib.no/tema/sovn) and *healthy and sustainable eating* (argument.uib.no/tema/kosthold). I also developed the idea for a scaffold (Bjønness & Kolstø, 2015) to support the students' development of critical thinking. The scaffold consisted of a poster with critical thinking questions related to different critical thinking skills that students and teachers could use to practice those skills. This was inspired by Facione's (2020, p. 8) *Questions to Fire Up Our Critical Thinking Skills* (the first edition was published in 1992) and Facione's (1990b) six core critical thinking skills. One of my first article ideas was to investigate how the use of this poster would affect how critical thinking was handled in classroom discussions and how students' critical thinking was expressed in group discussions. However, I learned an important lesson about the uncertainty related to doing research on people when no teachers in the classrooms we observed implemented the scaffold as part of the modules. Two newer versions of this poster—one for grades eight and nine, and one for grade ten—have now been added to the online resources

for the project (argument.uib.no/kritisk-tenkning) and will hopefully be used in future implementations of the modules. The posters can also be seen in Appendix I. These versions were developed primarily through discussions between Associate Professor Matthias Gregor Stadler, Professor Stein Dankert Kolstø, and me. The rest of the project team also contributed with feedback and design inputs, and Professor Kolstø created an instructional video for how to use these posters.

The combination of the large number of student participants in the project, my quantitative background in physics, and the seeming overlap between critical thinking education and my interests resulted in a desire to somehow investigate the critical thinking abilities of these students. Moreover, because the project seemed to include several strategies and components that previous research has shown to be effective for improving critical thinking ability (for meta-analyses, see Abrami et al., 2015; Abrami et al., 2008), I was interested in measuring the effect of the project on the students' critical thinking. Thus, I arranged for a well-known test of critical thinking skills—The *Cornell Critical Thinking Test* (Ennis & Millman, 2005)—to be translated to Norwegian, and I put it into an online survey solution. The translation was pilot tested and adjusted to fit the population of students in the project. This test has become the basis of all the research in this dissertation. Thus, instead of analyzing the implementation of the project, my research has investigated students' critical thinking as it was expressed when faced with different versions of this test. Furthermore, my research measured the effect of the implementation of the ARGUMENT project on students' critical thinking skills. The ARGUMENT project, then, mainly provided access to students, the quasi-experimental setting, and a context to discuss the results.

The first data collection using this test was done in the fall of 2019. These data were originally intended to just be used as a pretest. However, exploration of the data revealed some potentially interesting aspects of critical thinking among the items that most students answered incorrectly. Therefore, to bring further insight into these aspects, the test was modified to combine written justifications with some of the multiple-choice items. The data collection ended in the spring of 2020 with

administration of the posttest and the modified test. Article I presents a study that focused on identifying the aspects of critical thinking that were challenging for the students, and Article II presents a thematic analysis of students' reasoning related to these aspects. Article III presents a quasi-experimental study investigating the effect of the ARGUMENT project on students' critical thinking skills.

1.1 Motivational background

Critical thinking is viewed as a vital skill for the twenty-first century (Dam & Volman, 2004). Educators, policymakers, governments, and employers recognize critical thinking as essential for academic and professional success, as well as for citizenship, and therefore they want critical thinking to be emphasized in education (Association of American Colleges and Universities, 2011; Casner-Lotto & Barrington, 2006; OECD, 2015; Ventura et al., 2017). Furthermore, critical thinking is considered an important factor in outcomes such as political participation, scientific literacy, digital news literacy, and emotional intelligence (Guyton, 1988; Ku et al., 2019; Li et al., 2021; Rowe et al., 2015). Critical thinking test results have been correlated with fewer negative life outcomes concerning health, finance, and relationships (Butler et al., 2012), and critical thinking might even be a better predictor of such outcomes than intelligence (Butler et al., 2017).

Despite the clear positive outcomes associated with critical thinking and the widespread focus on improving it through education, there is concern that the efforts have failed to improve students' critical thinking (Bouygues, 2018; Case & Wright, 1997; Gunawardena & Wilson, 2021). This highlights the need for improving critical thinking research and instruction. In science education, the critical thinking focus has shifted towards an emphasis on how scientists practice science, resulting in an increased focus on argumentation-based inquiry practices (e.g., NGSS Lead States, 2013). This is also reflected in the renewed core curriculum for primary and secondary education in Norway, which emphasizes critical thinking, inquiry practices, and argumentation as educational goals (Ministry of Education and Research, 2017). The renewed core curriculum has clear parallels with the

ARGUMENT project, which aimed to help teachers develop and implement lessons where students practice scientific inquiry and generate their own arguments using real data in the context of socioscientific, or authentic, issues. Thus, insights from the ARGUMENT project can provide valuable knowledge on how these kinds of changes, when implemented at the school level, affect students' critical thinking.

1.1.1 Positioning the dissertation within the critical thinking literature

There is substantial evidence supporting that critical thinking can be improved through education across all disciplines and age levels, with particularly large effects seen in secondary education (for a meta-analysis, see Abrami et al., 2008). These effects have been found with different types of measures, including standardized tests. Types of interventions where critical thinking principles were made explicit were more effective than interventions where critical thinking was an implicit expectation. Teacher training was also found to be effective.

Another meta-analysis by Abrami et al. (2015) that focused on identifying effective strategies for teaching critical thinking found that inquiry into authentic (or socioscientific) issues, opportunity for dialogue, and one-on-one mentoring had positive effects on both general and content-specific critical thinking skills. In addition, these strategies had positive effects on critical thinking dispositions, and were effective at all levels of education and in all disciplines.

Nevertheless, several studies that seem to implement one or more of the general strategies identified in these two meta-analyses have found little to no effect on critical thinking (e.g., Bixler et al., 2015; Gul & Akcay, 2020; Pellegrino, 2007), indicating a need for more detailed insight into the specifics of what makes these strategies vary in their effectiveness. For example, some principles or components of critical thinking could be more important to explicate than others.

Historically, few studies have reported scores on the measured sub-scales of critical thinking, or any relationship between the sub-scales and the main outcomes, and Liu et al. (2014) suggested that future research should include this. More recent studies have included sub-scales of critical thinking. In one such study, Ku et al. (2019)

found indications that adolescents struggle with evaluating evidence and source credibility in news media. Another recent study showed that students often fail to identify biases in digital news (Nygren & Guath, 2019). Wineburg et al. (2016) analyzed 7804 responses from students ranging from middle school to college and found indications that many students struggle with recognizing conflicts of interest. These students were not able to distinguish an ad from a news story and did not check who was behind a site that presented only one side of a controversial issue. These skills that many students seem to not yet have mastered are potentially important, not least as aspects of digital literacy in today's digital age where online falsehoods have been called a global social problem (Pal & Banerjee, 2019). Furthermore, such problems concerning disinformation and infodemics in today's high-information society highlight an emerging need for fact-checking skills, and training in these skills have been found to correlate with improved job performance and personal development of future graduates (Pérez-Escolar et al., 2021).

The combination of written justifications and multiple-choice items that was utilized in the studies in Articles I and II of this dissertation could bring more detailed insight into barriers that students face concerning critical thinking. Ennis (1996b) reports having tried this, with promising preliminary findings. However, I am not aware of the results being published, nor by Ennis or anyone else. Norris (1990) combined verbal reasoning with a multiple-choice test of critical thinking and found that it did not affect test performance, which he views as an indication that verbal reasoning can be used to validate critical thinking tests.

Concerning sub-scales of critical thinking skills, novice learners tend to struggle with the distinction between observation and inference (Abd-El-Khalick et al., 2002). Moreover, The 2015 PISA-test shows that Norwegian students struggle with using evidence in arguments (Kjærnsli & Jensen, 2016). University lecturers across a range of disciplines seem to agree that the most important critical thinking skills are related to analysis, evaluation, and interpretation (Bellaera et al., 2021). However, there is a need for more insight concerning whether there are particular aspects of critical thinking that are more challenging than others for students. One contribution of such

insights could be to inform which specific principles of critical thinking should be included in effective types of critical thinking interventions (i.e., interventions where critical thinking is an explicit focus, according to Abrami et al., 2008).

Instructors from a range of humanities and social sciences in UK and US universities have reported that they mostly use implicit approaches to teach critical thinking, instead of the more explicit approaches that research has found to be more effective (Bellaera et al., 2021). Concerning Norwegian lower-secondary school teachers, the Teaching and Learning International Survey (TALIS) 2018 showed that these teachers rated themselves lower than teachers in the other Nordic countries on teaching practices and self-efficacy related to teaching critical thinking (Throndsen et al., 2019). This calls further attention to the need for bridging the gap between critical thinking research and teaching practice.

One suggestion for bridging this gap is that teachers should change their practices to align with the research literature on critical thinking (Schulz & Fitzpatrick, 2016). This approach also includes changing the curricula to match the research. Another approach tackles the issue from the opposite side and suggests that researchers should modify their research to take the teachers' perspective into account by positioning their studies in contexts in which learning occurs (Cáceres et al., 2020). Moreover, researchers should pose research questions with relevance for teachers and answer these in ways that make sense to educators. These two approaches imply two different research paradigms and methodologies—quantitative and qualitative research, respectively. Thus, bridging the gap between research and practice invites exploratory and pragmatic designs utilizing whichever methods and perspectives that are best suited to answer the research questions (Johnson & Christensen, 2017). Consequently, there is an increased use of *mixed methods research*, which mixes complementary components from quantitative and qualitative research.

1.2 Aims and research questions

The overarching aim of this dissertation consists of two parts, which are: (1) to contribute to narrowing the abovementioned gap between critical thinking research and teaching practice by means of providing pragmatic (i.e. useful) knowledge of students' struggles when solving critical thinking problems, and (2) to investigate whether supposedly effective yet general strategies for teaching critical thinking are effective in an *ecologically valid* (Gehrke, 2018) instructional setting in lower-secondary classrooms. The first one of these two aims is the focus of Article I and Article II, where we explored which aspects of critical thinking students struggled with and their reasoning when facing those aspects. In Article I, we used an *abductive approach* to linking theory and research (Saunders et al., 2016), where the exploration in Phase 1 of the study identified aspects of critical thinking that students seemed to struggle with, and the following data collection and analyses in Phase 2 supported and elaborated on the results from Phase 1. In Article II, we conducted a thematic analysis identifying themes in student reasoning using an *inductive approach* (i.e., we collected data to explore a phenomenon to gain new insights that could be used in generating new theories). The second aim is covered by Article III, where we used a *deductive approach* (i.e., we tested existing theory) to test the project's effect on students' critical thinking skills. Thus, the dissertation as a whole links theory and research using an abductive approach, which has enabled a continuous movement of perspective between theory and data while seeking to develop and add new knowledge to existing theory.

To achieve the aim of the dissertation, I have written three articles—co-authored by my supervisor Professor Stein Dankert Kolstø—each seeking to answer one research question:

Article I

Aspects of critical thinking that are challenging for students – Conflict of interest, observation, and inference

Research question 1 (RQ1):
Which aspects of critical thinking are particularly challenging for lower-secondary school students?

Article II

Students' reasoning when faced with test items of challenging aspects of critical thinking

Research question 2 (RQ2):
How do students reason when faced with test items from these challenging critical thinking categories?

Article III

Large-scale study suggests supposedly effective strategies for teaching critical thinking might be too general

Research question 3 (RQ3):
Do students from schools that are part of the project improve their general critical thinking skills more than students from non-treatment schools?

The results from Article III led us towards a discussion about whether the strategies for critical thinking instruction described in the literature could be too general. Thus, this discussion represents a second main claim which relates to a different question than RQ3.

2. Theoretical framework

Moseley et al. (2004) states that "trying to understand how people think and learn is in some ways an impossible challenge, since we can only try to understand these things by using the very processes that we do not fully understand" (p. 7). However, there are some choices available for the researchers trying to understand how people think. According to Moseley et al. (2004), researchers can focus on measurable aspects of human behavior (e.g., by using a problem-solving test), resort to metaphors with appeal to groups or individuals (e.g., brain as computer), and look for patterns (e.g., by conducting a thematic analysis of verbal reasoning). This dissertation includes all three of these.

In this chapter, I will present the theories on which the work in this dissertation is based on. In particular, I will discuss and compare several relevant definitions of critical thinking and justify my choices concerning the theoretical framework. Section 2.1 discusses different definitions of critical thinking, especially definitions based on cognitive skills and affective dispositions. This section also discusses the role of background knowledge (2.1.1), whether critical thinking is a generic or domain-specific ability (2.1.2), transfer of critical thinking (2.1.3), the presence of critical thinking in some popular terms and concepts from science education (2.1.4), and definitions that oppose those based on skills and dispositions (2.1.5). Section 2.2 discusses how to assess and measure critical thinking. Section 2.3 discusses teaching and learning of critical thinking. Section 2.3.1 discusses some relevant literature concerning dialogue and inquiry. Finally, Section 2.4 briefly summarizes the chapter.

2.1 What is critical thinking?

There are a wide range of definitions of critical thinking, and throughout the literature, the term has been used interchangeably with other terms such as *higher order thinking*, *thinking skills*, *good thinking*, *metacognitive thinking*, *productive thinking*, *creative thinking*, *analytical inquiry*, *analytical thinking*, and *logical thinking* (e.g., Liu et al., 2014; Moon, 2008; Moseley et al., 2004). Traditionally,

definitions of critical thinking can be categorized as coming from cognitive psychology (e.g., Halpern, 1998; Willingham, 2008), philosophy (e.g., Ennis, 1987; Paul, 1992), or education (e.g., Anderson & Krathwohl, 2000; Bloom, 1956). Cognitive psychologists tend to focus on the mental processes or cognitive skills that a good thinker *uses*. Philosophers, on the other hand, tend to focus on the characteristics and qualities that a good thinker *has*, as well as certain standards that good thinking should meet. Variations of Bloom's (1956) taxonomy, particularly the three highest levels (i.e., analysis, synthesis, and evaluation), were previously widely used as critical thinking frameworks in education (Sternberg, 1986). Nevertheless, for at least the last 30 years, the field of education has increasingly been including and integrating definitions and conceptions (i.e., proposals for implementation) of critical thinking from both psychology and philosophy (Abrami et al., 2015; Moseley et al., 2004; Pithers & Soden, 2000).

A simple and enduring definition of critical thinking comes from Robert Ennis (e.g., 1987), which proposes that critical thinking is "reasonable reflective thinking focused on deciding what to believe or do" (p. 10). Alongside his definition, Ennis—a philosopher and leading figure in the field of critical thinking and education—has put forward a list of cognitive skills (e.g., evaluating observation reports, analyzing arguments, etc.) and affective dispositions (e.g., be open-minded, look for alternatives, etc.) that constitute critical thinking (further elaborated in Ennis, 1996a). This view of critical thinking as consisting of skills and dispositions is supported by a confirmatory factor analysis by Taube (1997). This is sometimes referred to as the *two-factor* model of critical thinking (e.g., Clifford et al., 2004), and it has been called the current consensus among researchers (Ku, 2009). Ennis (2016) states that many well-known definitions of critical thinking are not substantially different from his own or from each other, and that these fall within a "mainstream concept of critical thinking" (p. 166). He clarifies the distinction between concept and conception by stating that conceptions are more elaborate proposals for implementation which are built upon a concept. Furthermore, he states that all these definitions are different ways to describe the same concept, and that all of them could

be used to make a conception consisting of a list of skills and dispositions, much like his own list.

A seminal definition of critical thinking that aligns with the mainstream concept of critical thinking and the two-factor theory comes from a panel of experts—including Ennis—that set out to reach a consensus on what critical thinking is (Facione, 1990b). The effort was organized by the American Philosophical Association (APA), and the purpose of the work was to improve assessment and curriculum development at all educational levels. The panel states that critical thinking is a form of purposeful and self-regulatory judgment which results in the use of certain cognitive skills: interpretation, analysis, evaluation, inference, explanation, and self-regulation. Each skill also has several subskills. Furthermore, the panel states that each skill is associated with criteria by which its execution can be evaluated. All six skills can relate to argumentation: One can analyze and evaluate arguments and their components, evaluate the credibility of statements, explain one's inferred conclusions by presenting persuasive arguments, and self-regulate by planning and monitoring one's own process of argumentation. However, a person might possess these skills without being inclined to use them appropriately. Thus, in addition to having the right skills, the ideal critical thinker also has the dispositions to exercise those skills. According to the APA panel (Facione, 1990b), these dispositions of critical thinking represent a "critical spirit, a probing inquisitiveness" (p. 20), and they consist of what seem to be virtues, mental habits, and personality traits. The dispositions include inquisitiveness, concern to be well-informed, open-mindedness, precision, and orderliness in working with complexity. All the skills, subskills, and dispositions are listed in Appendix II.

Other well-known definitions that represent the mainstream concept of critical thinking include those by Siegel (1988) and Kuhn (2015, p. 47), as well as one by Scriven and Paul (1987) stating that "critical thinking is the intellectually disciplined process of actively and skillfully conceptualizing, applying, analyzing, synthesizing, and/or evaluating information gathered from, or generated by, observation, experience, reflection, reasoning, or communication, as a guide to belief and action"

(para. 4). Another noteworthy definition of critical thinking comes from Diane Halpern. She states that critical thinking is "the use of those cognitive skills or strategies that increase the probability of a desirable outcome," and that it is "purposeful, reasoned, and goal-directed—the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions" (Halpern, 2014, p. 8). Moreover, she includes the need for dispositions, like persistence, willingness to plan, and being mindful. In addition to skills and dispositions, Halpern's conception of critical thinking also considers metacognition, intelligence, fast versus slow thinking, and the emotional aspects of thinking. Metacognition in relation to critical thinking refers to monitoring one's thinking process, ensuring that progress is made toward a suitable goal, and deciding the appropriate time and effort to allocate to the task. Concerning intelligence and critical thinking, Halpern makes the case that if people can learn to think better—an idea that is supported by many types of evidence—then they can learn to become more intelligent, at least by everyday definitions of intelligence. Halpern ties in Kahneman's (e.g., 2011) concept of fast versus slow thinking with critical thinking, and she states that while fast and intuitive thinking is likely good thinking when performed by an expert in a given field, one most often have to put in hard and deliberate work of thinking when thinking critically. Emotions also affect thinking as people are far from rational, and Halpern points to evidence that emotion, such as anger, affect people's thinking.

Bailin et al. (1999) find the term "skills" to be inaccurate and confusing as it is easily interpreted as describing psychological processes rather than descriptions of what people can accomplish. Thus, their conception of critical thinking instead consists of *intellectual resources*, such as *operational knowledge of the standards of good thinking* (e.g., considering alternatives), *knowledge of key critical concepts* (e.g., necessary and sufficient conditions), and *heuristics* (e.g., double-checking before concluding). A critical thinker should be familiar with a wide range of key critical concepts, such as argument, premise, and conclusion, in order to recognize and make appropriate distinctions when assessing intellectual products. The standards of good thinking also include standards for judging the credibility of statements made by

authorities and the reliability of reports made by observers. Thus, the intellectual resources seem to be related to the cognitive skills presented in other conceptions based on the mainstream concept of critical thinking (e.g., Ennis, 1996a). As mentioned at the beginning of this chapter, cognitive skills might be impossible to observe directly (Moseley et al., 2004). Thus, the hesitancy of Bailin et al. (1999) to use the term "skills" is understandable, although their underlying concept seems to fall within the mainstream concept of critical thinking. Bailin and Siegel (2003) provide further indications of the possible overlap of this underlying concept and the mainstream concept, as well as of the overlap of the conceptions built on top of the concepts. For example, while Bailin is adamant in avoiding the noun "skills" altogether, Siegel finds it acceptable insofar as skilled thinking refers to thinking that meets relevant criteria. Besides, the conception of Bailin et al. (1999) also includes *habits of mind*, or dispositions, such as respect for reasons and truth, open-mindedness, and an inquiring attitude.

All of these definitions seem to align with the mainstream concept of critical thinking (Ennis, 2016), and each one could be used as the theoretical framework for this dissertation. Knowledge of key critical concepts, presented in Bailin et al. (1999), might be relevant for some of the findings from the studies in this dissertation, such as the identified difficult aspects of critical thinking related to conflict of interest and observation versus inference. Nevertheless, because the *Cornell Critical Thinking Test* (CCTT) *Level X* (Ennis & Millman, 2005) is an important foundation for all the work in this dissertation, Ennis' conception of critical thinking (e.g., Ennis, 1987; Ennis, 1996a) is particularly important for the theoretical framework. However, the closely related conception by Facione (1990b)—which Ennis helped create—is the most important part of the theoretical framework. It is clear and inclusive enough to provide a framework for assessment and instruction of critical thinking in education. Moreover, it is well-known and widely used (e.g., Abrami et al., 2015; Abrami et al., 2008; Ventura et al., 2017), which could indicate that many researchers and educators might be familiar with this conception or closely related variations of it.

2.1.1 The role of background knowledge

Most critical thinking scholars find background knowledge important (Ventura et al., 2017), and McPeck (1990) maintains that to think critically, students need something to think critically about. Similarly, many researchers view background knowledge as essential for practicing critical thinking skills (e.g., Kennedy et al., 1991; Willingham, 2008). Ennis (2016) includes *using existing knowledge as basis for making decisions* as part of his list of critical thinking abilities. He further specifies, "There is absolutely no doubt that background subject matter knowledge in an area calling for critical thinking is essential for doing high-quality critical thinking in that area" (p. 178). However, he emphasizes that "this does not justify not learning to think critically about subject-matter issues that use levels of subject matter that students can understand" (p. 178). The APA expert panel (Facione, 1990b) states that although the critical thinking skills transcend specific disciplines, learning and applying those skills in a variety of contexts often require domain-specific knowledge. This includes understanding methodological principles to perform practices that are essential for reasonable judgment in those specific contexts.

In her book, *Though and Knowledge: An Introduction to Critical Thinking*, Diane Halpern (2014) describes the close relationship between the constructs thought and knowledge. She states that a critical thinker uses skills for learning new techniques and connects new knowledge with previously learned knowledge. Moreover, she touts her favorite definition of critical thinking, published in 1960 by Bertrand Russell. She sums his definition up in an equation where critical thinking equals the sum of three components: attitude, knowledge, and thinking skills.

Bailin et al. (1999) include background knowledge as one of the intellectual resources necessary for critical thinking in their conception. They state that critical thinking always takes place in the context of existing concepts, beliefs, and values. In that regard, the quality of any critical thinking depends to a considerable extent on what the person doing the critical thinking knows or can find out. The authors (i.e., Bailin et al.) find the necessity for background knowledge obvious yet worth noting, as they

point out that there are educators who view teaching critical thinking as a process that is developed completely independent of the development of knowledge.

2.1.2 The generality-versus-specificity debate

There has been a longstanding debate about whether critical thinking is general and applicable across different subject domains or whether it is confined to the context and domain within which it was learned (Ennis, 1989). This debate is closely related to the debate about the importance of background knowledge, discussed in the previous section, and to the debate about the transferability of critical thinking, discussed in the next section. In a seminal piece concerning subject specificity, Ennis (1989) warns that viewing background knowledge as *necessary* for critical thinking does not mean that background knowledge is *sufficient* for critical thinking. Most researchers seem to think that there are at least some generalizable aspects of critical thinking, and the debate is mostly about the usefulness of these aspects for instruction and conceptions of critical thinking.

McPeck (e.g., 1981, 1990) and Willingham (2008) represent the *specifist*-side of the spectrum as they argue that it makes little sense to talk about critical thinking in a generic sense. Instead, McPeck (1990) views critical thinking skills as being more useful the more they are connected to a specific domain, and claims that there is an inverse relationship between the usefulness of critical thinking skills and their generality. In a response to Ennis (1989), McPeck states, “The more general they are, the more trivially obvious they are—for example, not contradicting oneself, not believing everything one hears, and so on” (1990, p. 12). In a reply to McPeck’s response, Ennis (1990) maintains that those skills might be obvious to McPeck but not trivial, and that they are worth teaching because they are so important and because so many people make those types of mistakes (Ennis, 1990).

More towards the *generalist*-side of the spectrum there are researchers that are more positive about the usefulness of viewing critical thinking as general. For example, Halpern (e.g., 2001, 2014) and Robinson (2011) maintain that there is great potential in instruction focused on general critical thinking skills and they list several examples

of successful interventions. Siegel (1988) views certain aspects of critical thinking as general, such as the ability to identify informal fallacies of reasoning. Davies (2013) views the generic sense of critical thinking as the foundation for any supervenient domain-specific sense of critical thinking, and he argues that any such domain-specific sense of critical thinking can be fully explained by the underlying generic sense of critical thinking. Furthermore, van Gelder (2005) argues that critical thinking is "intrinsically *general* in nature" (p. 8) because critical thinking by its definition—roughly speaking, thinking which helps you decide whether to believe something—applies to a wide range of contexts and domains.

Ennis (1989) is among those who maintain that both domain specific and general aspects of critical thinking are important, and he argues that there are differences in reasoning between fields. For example, deductive proof is a common criterion of good reasons in mathematics, while in social sciences statistical significance is a more important consideration. Moreover, in the arts, some subjectivity is acceptable, whereas it is often shunned in the sciences. In the same paper, he also argues that there are commonalities in critical thinking between fields, such as the awareness of how a conflict of interest affects the credibility of a source. Smith (2002) acknowledges the existence of domain-specific thinking skills yet concludes that "there are far more elements of generality in our thinking practices than have heretofore been recognized" (p. 224). In his paper, *Critical Thinking Across the Curriculum: A Vision*, Ennis (2016) also acknowledges both sides of the generality-specificity spectrum by listing general skills and abilities as well as providing suggestions for courses that focus on subject-specific critical thinking abilities, like planning and performing double-blind experiments in medicine. The APA expert panel (Facione, 1990b) has described its list of critical thinking skills as "*pervasive* and *purposeful*" (p. 15)—meaning, in part, that these skills can be used as integrated and essential elements in other endeavors, including programming computers, defending clients, developing a winning sales strategy, or helping a friend find out what might be wrong with his or her car. Moreover, the expert panel states that within the curriculum the goal of learning critical thinking can be differentiated from the

goal of learning subject-specific content. At the same time, the panel does not deny that one of the best ways to learn critical thinking is within the context of a subject.

Although Bailin et al. (1999) emphasize the importance of background knowledge and the need for specific learning and practicing of critical thinking within specific subject-domains, they concede that there are general aspects of critical thinking. For example, they mention general heuristics for critical thinking, such as trying to think of counterexamples, asking for examples to clarify meaning, making a list of reasons for and against an issue, and discussing a problem with a knowledgeable person. Similarly, Flavell (1976) proposes double-checking something before accepting it as fact, and Sternberg (1987) proposes that when solving a problem it is useful to divide it into subproblems. The dispositions of critical thinking, or habits of mind, presented in Bailin et al. (1999) are also generalizable. Bailin and Siegel (2003) agree that both generalism and specifism are correct in several important aspects. For example, some criteria for judging reasons are narrow, while others are broad. Moreover, they agree that the epistemology underlying critical thinking is fully generalizable. This epistemology, implied in the mainstream concept of critical thinking, is represented by a rejection of relativism, a distinction between rational justification and truth, and a recognition that rational justification can be used as a fallible indicator of truth (Bailin & Siegel, 2003).

In sum, although there are disagreements about the usefulness of viewing critical thinking as generalizable, it seems that many researchers agree, at least to some extent, about the generalizability of certain aspects of critical thinking. For the purpose of this study, I adhere to the views held by Ennis (e.g., 2016) and the APA expert panel (Facione, 1990b), acknowledging the importance of both sides of the generality-specificity spectrum as well as the need for background knowledge. This is suitable for the framework of this dissertation because the instrument that was used is a test of general critical thinking skills. Importantly, testing for general critical thinking skills might be useful if we want to find out something about the transfer of such skills to daily life (Ennis, 1989), which is the topic of the next section.

2.1.3 Transfer of critical thinking

The concept of *transfer* addresses the question of whether something that is learned in one context, such as in school, can be successfully applied in a different context, such as in a real-world context like when deciding which profession to choose. This is arguably the ultimate goal of critical thinking instruction and of education in general (Abrami et al., 2008; National Research Council, 2000). It might be useful to think of transfer as "flexible adaptation to new problems and settings" (National Research Council, 2000, p. 77). Such flexible adaptation, or transfer, seems to be supported by deep initial learning, metacognitive monitoring, and learning something across multiple contexts.

The issue of transfer is related to the generality-specificity debate, and those that are on the specificist side of the spectrum are more likely to be skeptical of the possibility of transfer (Ennis, 1989; Lai, 2011). However, although most researchers agree that spontaneous transfer of skills and dispositions to new contexts is rare, there also seems to be agreement about such transfer being possible (Ennis, 1989; Kennedy et al., 1991; Willingham, 2008). Even a hardcore specifist like McPeck (e.g., 1981) notes that his view includes the possibility of transfer to real-world contexts. The APA expert panel also suggests that critical thinking instruction should be aimed at helping students generalize the skills and dispositions to a variety of contexts (Facione, 1990b). Ennis (1989) maintains that transfer becomes likely if there is sufficient practice in several different domains and if the instruction focuses on transfer. Furthermore, he emphasizes the need for clarity in what is meant by terms such as "domain" and what this implies about the *distance* of transfer—is it near or far? Bailin (2002) also emphasizes this need for clarity about what constitutes a domain as well as about what exactly is being transferred.

One example of near-transfer of critical thinking skills within an educational setting comes from Tiruneh et al. (2016). They showed that performance on a domain-specific test of critical thinking skills (i.e., within electricity and magnetism) significantly predicted performance on a domain-general test of critical thinking skills. Moreover, Ennis (1989) states that testing for general critical thinking skills

could be important if we want to find out if transfer occurs—and there are numerous examples of instructional interventions within specific subject-domains that have led to increased scores on tests of general critical thinking skills, like the one used in this study (for reviews and meta-analyses, see Abrami et al., 2015; Abrami et al., 2008).

Halpern (2001, 2014) points to several studies indicating that better thinking can be learned in one context and applied in other contexts, including one example of far-transfer where students spontaneously transferred reasoning skills from an academic context to a real-world context taking place months after the academic instruction (Fong et al., 1986). Moseley et al. (2005) also conclude that instruction in critical thinking can help people think better and that improved thinking will transfer to new contexts. Hernstein et al. (1986) used a randomized controlled experiment with blind grading to show that thinking skills were transferred and used appropriately with novel topics. Moreover, the treatment group showed greater gains on an intelligence test. Another indication of far-transfer is that scores on a test of critical thinking skills have been shown to correlate with fewer negative life outcomes (Butler et al., 2012)—in fact, critical thinking scores were shown to be a better predictor of such outcomes than intelligence (Butler et al., 2017).

In sum, although there is little doubt that spontaneous transfer is rare and difficult to accomplish, most researchers seem to agree that critical thinking can be learned in one context and transferred to another context, including from a school context to a real-world context. To accomplish this lofty goal, instruction should focus on transfer and emphasize the need to practice and use critical thinking skills in many contexts (e.g., Ennis, 1989; Facione, 1990b; Halpern, 2014). Moreover, there is evidence showing that transfer of learning in general might be enhanced when learners are tested during learning (Rohrer et al., 2010). Assessment of critical thinking is covered in Section 2.2.

2.1.4 Critical thinking and popular concepts in science education

There is substantial overlap between the mainstream concept of critical thinking and definitions of popular terms in contemporary science education such as *inquiry*,

scientific literacy, *science practices*, and *scientific practices*. Barbara Crawford (2014) combines historical views by Dewey and Schwab with modern views of the authors of the U.S. K-12 Framework (National Research Council, 2012) and Next Generation Science Standards (NGSS Lead States, 2013) in a definition of inquiry:

> Teaching science as inquiry involves engaging students in using critical thinking skills, which includes asking questions, designing and carrying out investigations, interpreting data as evidence, creating arguments, building models, and communicating findings in the pursuit of deepening their understanding by using logic and evidence about the natural world. (p. 515)

Crawford (2014) states that the K-12 Framework tries to rebrand inquiry as science practices. These *Science and Engineering Practices* (National Research Council, 2012) include analyzing and interpreting data, evaluating information, and planning and carrying out investigations. They overlap with the APA expert panel's conception of critical thinking, and a juxtaposition of the two conceptions can be seen in Appendix III. For example, APA's skills of interpretation, analysis, evaluation, and explanation could encompass the science practices *analyzing and interpreting data* and *obtaining, evaluating, and communicating information*. Similarly, PISA's 2015 list of competencies for scientific literacy (Roberts & Bybee, 2014) includes understanding scientific inquiry, interpreting scientific evidence, and explaining scientific phenomena. Thus, the term scientific literacy also seems to overlap substantially with the APA conception of critical thinking (e.g., the skills of interpretation and explanation). Erduran and Dagher (2014b) focus on three scientific practices that also seem related to the APA conception: classification, observation, and experimentation. Observation and classification could be covered by skills of interpretation and analysis. In the same book, Erduran and Dagher (2014a) present a list of aims and values of science that coincide with commonly mentioned dispositions of critical thinking. A juxtaposition with the APA's list of dispositions can be seen in Appendix IV. Notable overlapping themes include the focus on open-mindedness, respect, and understanding of diverging worldviews; honesty in all

aspects of science, including in facing one's own biases; and valuing evidence-based reasoning.

In sum, comparing these terms from contemporary science education with the mainstream concept of critical thinking as manifested in the APA consensus conception indicates the usefulness of this general conception in other domains—in this case, the domain of science. Critical thinking, in a way, seems to be the glue that permeates through the core of all these concepts, while their unique surface characteristics from the science domain remain intact.

2.1.5 Critiques and alternative concepts of critical thinking

In this section, I briefly discuss some definitions and conceptions of critical thinking that do not align with the mainstream concept. In one of these alternative conceptions, Barbara Thayer-Bacon (2000) critiques how critical thinking has been conceptualized by major theorists and calls for a broader and more inclusive concept, which she calls *constructive thinking*. While acknowledging the work of the previous theorists, her focus goes beyond the traditional tools of reasoning and rationality—including skills and dispositions, which she sees as insufficient. Instead, she emphasizes that constructive thinking is a social practice—she uses the analogy of a quilting bee—where participants with diverse backgrounds and orientations use many different tools as part of the thinking process. These tools can include intuition, emotion, and imagination. On the other hand, Thayer-Bacon problematizes what she views as a focus on individual thinking in the traditional critical thinking lineages, stemming all the way back from Plato to contemporary theorists, and she points out that these lineages represent hierarchies of affluent white males. Rodin's statue, *The Thinker*, is presented as the epitome of the traditional view of the critical thinker—a man in solitude with his thoughts. Drawing on feminist and postmodern theories, Thayer-Bacon seeks to offer a novel and more inclusive platform where women, men, and feminine theorists can engage in her concept of constructive thinking, not least to fight oppression and power.

While Thayer-Bacon (2000) argues that the mainstream concept of critical thinking is not broad and inclusive enough, a more profound objection to the mainstream concept is that its philosophical underpinnings are fundamentally flawed. A special issue of *Studies in Philosophy Education* contains two articles that present this argument, one by Gert Biesta and Geert Stams (2001) and one by James Marshall (2001).

Biesta and Stams (2001) point out that the mainstream concept of critical thinking includes a concept of criticality that uses one or more criteria as the foundation for evaluation when deciding what to believe or do. They refer to this type of criticality as *critical dogmatism* as it is dependent on a somewhat dogmatic assumption about the *truth* of the underlying criteria. Biesta and Stams find this potentially problematic because of the impossibility of justifying the criteria without either dogmatically installing them as a foundation, using circular logic, or through an infinite regress. This is known as the *Münhhausen trilemma* (Albert, 1985), a thought experiment used to demonstrate how any truth, not just related to critical thinking, is theoretically impossible. However, Biesta and Stams see nothing objectionable to this type of criticality (i.e., the one used in the mainstream concept) as long as one acknowledges its dogmatic character. Nevertheless, Biesta and Stams prefer another type of criticality, *Derridean deconstruction*, a postmodernist concept that rejects the possibility of establishing any solid foundation like the criteria in critical dogmatism. While a thorough explanation of deconstruction is beyond the scope of this dissertation, not to mention beyond the scope of my expertise, Biesta and Stams state that deconstruction "tries to open up the system in the name of that which cannot be thought of in terms of the system" and it is "an affirmation of what is wholly other" (Biesta & Stams, 2001, p. 68). Abrami et al. (2015) note that "it is difficult to say exactly how an approach like this would translate into CT [critical thinking] teaching practice" (p. 278). Biesta and Stams state that Derrida refers to the "concern for what is wholly other" as *justice*, and they state that it is from this concept of justice "that deconstruction derives its right to be critical" (p. 68). They conclude from this that social justice is at least as important as rationality as a goal of critical thinking—opposing Siegel's (1995) view that the philosopher's goal is not to bring forth social

justice. This also opposes the APA conception (Facione, 1990b), which rejects that critical thinking has a normative component and maintains that one might be a proficient critical thinker even if one uses critical thinking for unethical purposes.

This theme of social justice also seems to resonate with Marshall's (2001) thoughts. He proposes that critical thinking must include some notion of the self, not just a set of tools like skills and dispositions. Marshall is particularly fond of Foucalt's concept of the self, where the fundamental question for the self is, according to Marshall, "how one practices one's freedom" (p. 83). Consequently, Marshall seeks a concept of critical thinking that is merged with critical theory to ultimately lead students to question social systems and bring forth liberation from oppression.

These alternative concepts of criticality and critical thinking are an important part of the critical thinking discourse and might provide value for critical thinking theory and education. For example, Thayer-Bacon's (2000) focus on critical thinking as a social practice aligns well with popular sociocultural theories of learning, stemming from Vygotsky (Leach & Scott, 2003). Besides, Abrami et al. (2015) argue that these alternative concepts are not totally incommensurable with the mainstream concept of critical thinking. More specifically, although these alternative conceptions might seek a broader type of critical thinking, improving critical thinking skills and dispositions could also enhance students' propensity towards questioning social systems and being less governable, which in turn could lead to more freedom and social justice. Thus, there seem to be important similarities in at least some of the desired outcomes that could be expected to occur from these differing concepts of critical thinking.

In sum, the traditional concept of critical thinking is a good fit with the aims of this dissertation, the context of the ARGUMENT project, and the broader context of education and education research. Furthermore, it is in alignment with my view of how critical thinking can be conceptualized in a useful way and then be taught and measured. On the other hand, it is not yet clear exactly how one might teach and measure critical thinking using the alternative concepts presented in this section, although the expected and desired outcomes could be viewed as at least somewhat

commensurable. All models are wrong, but some are useful—and some are more useful than others, depending, not least, on the context where any particular model is used. Skills, dispositions, and knowledge surely do not capture everything that constitute different concepts of critical thinking, although, as Abrami et al. (2015) argue, to reject these kinds of data completely would in many cases be "tantamount to discarding the good for the sake of a nonexistent and unobtainable better" (p. 304).

2.2 Assessing and measuring critical thinking

Many well-known critical thinking assessments seem to be based on the mainstream concept of critical thinking. However, each assessment should have a clear conception about what exactly it is trying to assess. Moreover, those using these assessments should be clear about the purpose of the assessment. There are several potential purposes for assessing critical thinking, and different tests might serve different purposes. Ennis (1993) lists seven such purposes, including diagnosing levels of students' critical thinking, giving students feedback, motivating students, informing teachers about the success of their efforts in teaching critical thinking, doing research, making decisions about admissions to schools, and evaluating schools.

Most existing assessments of critical thinking treat it as a general concept that is applicable across domains (Liu et al., 2014). Thus, these assessments aim to measure general skills or general dispositions related to critical thinking. Nevertheless, the need for tests that measure domain-specific critical thinking is also recognized, and such tests do exist (Ennis, 1993; Tiruneh et al., 2016). Some of the well-known tests that measure general critical thinking skills and/or dispositions include the *Watson-Glaser Critical Thinking Appraisal* (WGCTA; Watson & Glaser, 1980), the *California Critical Thinking Skills Test* (CCTST; Facione, 1990a), the *California Critical Thinking Disposition Inventory* (CCTDI; Facione & Facione, 1992), the *Halpern Critical Thinking Assessment* (HCTA; Halpern, 2010), the *Ennis-Weir Critical Thinking Essay Test* (Ennis & Weir, 1985), and the *Cornell Critical Thinking Test* (CCTT; Ennis et al., 2005) (for a more comprehensive list, see Liu et al., 2014).

Most of the published tests are designed for adults and college students. However, one version of the CCTT, *Level X*, is aimed at students in Grades 4 through 14 (Ennis & Millman, 2005); the Ennis-Weir Critical Thinking Essay Test is aimed at students in Grades 7 through college; and the Watson-Glaser Critical Thinking Appraisal is aimed at students in Grade 9 through adulthood (Ennis, 1993).

Most of the tests exclusively use selected-response (i.e., multiple-choice) items, while some use a combination of multiple-choice and constructed-response items (e.g., the HCTA)—where test takers respond to a situation in their own words but are also forced to choose between different alternative answers. On the other hand, the Ennis-Weir test requires students to write an essay. Restrictions in testing time limit the number of constructed-response items that can be used in any given assessment. Multiple-choice items are easier to score but more susceptible to guessing, while essay-tests and constructed-response items are more time-consuming to score but can provide a more comprehensive assessment. If a test is to be a more authentic assessment of critical thinking performance—like tests that use constructed-response items—its psychometric quality tends to suffer (Liu et al., 2014). While multiple-choice assessments tend to be objective, efficient, reliable, and low-cost, more authentic assessments tend to provide less information in the same amount of time, cost more, and be less reliable (Lee et al., 2011; Wainer & Thissen, 1992). However, the higher cost of scoring constructed-response items or essays could become less of a concern with the emergence of automated scoring software (e.g., Bridgeman et al., 2012; Leacock & Chodorow, 2003). Tests that use constructed-response items or real-life observational assessments could provide more authentic scenarios because students must exercise critical thinking in ill-structured problems without provided alternative responses, much like in real-world scenarios (Ennis, 1993; Jaarsveld et al., 2012; Liu et al., 2014). Thus, there is a need to strike a balance between the psychometric quality provided by selected-response items and constructed-response items. Using a combination of several item formats is broadly recommended (Butler et al., 2012; Ennis, 1993; Halpern, 2010; Ku, 2009; Liu et al., 2014). Similar to the approach used in the studies included in this dissertation, Ennis (1993) suggests including written justifications with items from the CCTT.

Concerning the psychometric quality of common critical thinking tests, Abrami et al. (2008) note that although a great deal of research has been done to establish the validity and reliability of the different measures, the results are inconsistent. In addition, studies conducted by researchers not affiliated with the test makers tend to report lower psychometric quality of the tests compared to studies conducted by researchers affiliated with the tests (Ku, 2009). Liu et al. (2014) note that most studies exploring the validity of existing critical thinking tests focus on the correlation between critical thinking scores and other cognitive measures, and that this provides support for their validity. Some studies use the relationship between critical thinking test results and life outcomes, behaviors, or job performance as evidence supporting test validity (e.g., Butler et al., 2012; Ejiogu et al., 2006). Liu et al. (2014) point out that these types of studies seldom control for other cognitive measures to make it easier to evaluate the unique contribution of critical thinking in predicting the outcomes. Nevertheless, as mentioned in Section 2.1.3, Butler et al. (2017) did control for intelligence yet still found critical thinking scores to be a better predictor of life outcomes.

Liu et al. (2014) also point out that there are some common problems related to the validity of existing assessments, including insufficient evidence of distinct dimensionality, unclear evidence of differential validity across groups of test takers, and unreliable subscores. Bernard et al. (2008) explored the intercorrelational structure of the subscales of the WGCTA through a factor analysis of 70 studies that had used the test. They concluded that the best interpretation of the WGCTA is represented by the total score instead of the individual scores on the five subscales. In a recent factor analysis of the CCTT Level X, Leach et al. (2020) suggest that there is a need for more clarity concerning the construct validity of the test, especially concerning the different subscales the test is said to measure. However, because critical thinking is multifaceted and complex, it might be difficult to create tests of critical thinking that fit a neat factor structure without making the items unrealistically unidimensional. The psychometric quality of the CCTT Level X, the instrument used in the studies presented in this dissertation, is further discussed in Section 3.1.

In sum, there are several widely used tests of critical thinking skills and dispositions that have been used for both summative and formative assessments across educational domains and age levels. However, there is still a need for continued efforts in developing new and improving existing assessments (Ennis, 1989; Liu et al., 2014), as well as a need for increasing their use in education.

2.3 Teaching and learning critical thinking

People have been learning how to be good critical thinkers for millennia without formal education. There are countless avenues that might have led to improvements in critical thinking, including observing and evaluating sights and sounds in the forest to find food and avoid dangers, becoming aware of how thoughts produce bodily reactions to improve decision making, and discussing politics with a disagreeable uncle to counter his arguments (Facione, 2020). College students' out-of-class experiences can have statistically significant, positive, and unique effects on their gains in critical thinking (e.g., Terenzini et al., 1995; Twale & Sanders, 1999). Moreover, there is evidence that even preschoolers, 3 to 5 years of age, develop some sense of critical thinking (e.g., Jaswal & Neely, 2006; Koenig & Harris, 2005). Luckily, formal instruction can also lead to predictable gains in critical thinking. Thus, in this section of the dissertation, I focus on teaching and learning of critical thinking within educational contexts.

In a meta-analysis of 117 studies, Abrami et al. (2008) found that critical thinking skills and dispositions can be improved through educational interventions at all age levels, with the largest effects seen in secondary education. Furthermore, they showed that effects were found with different types of measures, including standardized tests. A total of 32% of the variance in critical thinking effect sizes were explained by type of critical thinking intervention (i.e., the degree to which critical thinking is an explicit focus of the instruction) and pedagogical grounding (e.g., teacher training was more effective than simply having a curriculum that was focused on critical thinking). The type of intervention can be explained by Ennis's (1989) typology of four critical thinking courses: *general* (i.e., critical thinking is the course content),

infusion (i.e., critical thinking is made explicit and is infused into subject matter content), *immersion* (i.e., immersion in subject matter content where critical thinking is only an implicit goal), and *mixed* (i.e., general critical thinking instruction mixed in with either infusion or immersion). Of these, mixed and infusion were the most effective types of interventions, and the least effective interventions were of the immersion type where critical thinking principles were not explicated. In addition, the analysis shows that collaborative learning conditions had a small but significant effect on critical thinking.

In another meta-analysis, this time of 341 studies, Abrami et al. (2015) found that there are effective strategies for teaching general and content-specific critical thinking skills at all educational levels and across all disciplines. In addition, the authors found that critical thinking dispositions were also improved through educational interventions. The opportunity for dialogue, especially whole-class teacher-led discussions as well as teacher-led group discussions, seems to improve the acquisition of critical thinking skills. Moreover, critical thinking outcomes seem to be improved from exposure to authentic issues that stimulate inquiry, particularly when this includes role-playing and applied problem solving. Finally, while one-on-one mentoring did not generate strong results on its own, it appears that the combination of mentoring with dialogue and authentic issues yields larger effects than either one alone or dialogue and authentic issues combined. Thus, the results suggest that mentoring might serve as a catalyst for critical thinking in combination with other effective strategies.

Based on these two meta-analyses above (Abrami et al., 2015; Abrami et al., 2008), some effective strategies for critical thinking instruction are teacher training, dialogue, inquiry into authentic issues, mentoring, and making critical thinking an explicit part of the instruction. In the following paragraphs, I briefly discuss some research that supports the efficacy of these strategies.

In addition to the findings of Abrami et al. (2008) showing that teacher training has an effect on students' critical thinking, there is a large body of research showing that

professional development programs can affect student learning in general (Blank & Alas, 2009; Egert et al., 2018). Concerning dialogue, there is substantial evidence supporting the effectiveness of dialogic teaching for learning outcomes across various settings and subject matters and with all types of students (Alexander, 2020). Mercer (2008) showed that a certain type of dialogue, *exploratory talk*, yielded greater improvements than other types of dialogue in students' scores on a reasoning test, as well as in math and science. Moreover, gains in critical thinking scores from discussion have also been highlighted through studies on collaborative and cooperative learning (Huang et al., 2017; Loes & Pascarella, 2017). The study by Huang et al. (2017) also explicated domain-specific critical thinking skills in tactical lessons for the sport of basketball. Several researchers recommend using authentic, real-world, or socioscientific issues for learning critical thinking (e.g., Lai, 2011; Zeidler et al., 2009), and for learning in general (e.g., Hmelo-Silver et al., 2007). However, for inquiry into such issues to be successful, educators should provide sufficient structures to help guide the students (Cargas, 2016; Kolstø et al., 2006). In a large randomized controlled trial that included 48 schools, the use of a strategy based on argumentation-focused guided inquiry (*The Science Writing Heuristic*) resulted in significant improvements in students' scores on the CCTT Level X (Hand et al., 2018).

Yang et al. (2008) conducted a study in which a combination of dialogue, authentic issues, and mentoring was used to improve the critical thinking skills of distance veterinary students. This study also made critical thinking explicit through the instructor's modeling of critical thinking. In addition, there was an aspect of teacher training as the instructor discussed the study and its conception of critical thinking with one of the researchers and participated in a pilot study before the main study. In the main study, the students were encouraged to think critically about the course content and about questions and comments from their instructor and their peers. Gains in critical thinking were larger when using structured discussions rather than unstructured discussion, and even larger gains were seen then the teacher modeled critical thinking skills at the beginning of the semester rather than the middle. The

effect sizes were impressive. Nevertheless, the results did not reach statistical significance, probably due to small sample sizes.

Although these general strategies seem to be effective for improving critical thinking, there are several studies that have found no effect or a negative effect from using these strategies (Abrami et al., 2015). Article III of this dissertation discusses potential reasons why some studies have not been able to find gains in critical thinking scores while using dialogue (e.g., Bixler et al., 2015; Garside, 1996; Pellegrino, 2007), authentic issues (e.g., Beavers et al., 2017; Gul & Akcay, 2020), and teacher training (e.g., MacPhail-Wilcox et al., 1990; Zohar & Tamir, 1993).

Many of the authors behind the different conceptions of critical thinking presented in Section 2.1 also propose models for teaching critical thinking, some of which I will present here. Ennis (2016) lays out a comprehensive and ambitious vision for teaching critical thinking across the higher education curriculum. His program starts the first year with a required 30-week, 3-hours-per-week course called “Introduction to Critical Thinking”. The first two-thirds of this course will teach general critical thinking skills and dispositions and help students learn to apply these to various parts of their lives. The final third presents students with case studies calling for critical thinking in subject-specific issues. In addition, general and subject-specific critical thinking will be included in most courses across all fields. The thesis project will be geared more towards critical thinking and include weekly discussions with supervisors and peers to evaluate each other’s work and to give and receive feedback. Ennis recommends two complementary teaching practices: making critical thinking principles explicit and teaching for transfer. Moreover, he proposes two basic teaching methods: lecture-discussion teaching, which is commonly used in higher education already, and problem-based learning, which overlaps with the use of authentic issues mentioned above. In addition, he emphasizes that assessment must happen early and continuously and be both formative and summative.

The APA expert panel (Facione, 1990b) also proposes that critical thinking instruction should include making procedures explicit, as well as explaining and

modeling their appropriate use and justifying their implementation. Moreover, learners should be exposed to situations that require them to exercise critical thinking skills and dispositions. These situations should be artificially simple at first, before evolving into situations that are complex and realistic. Throughout this process, learners should receive constructive feedback and be coached and motivated to become better and more independent critical thinkers. The panel seems to recommend a mixed approach where general critical thinking principles can be taught on their own or be infused into subject-matter courses. Finally, the APA panel suggests that critical thinking should be a part of the curriculum at all levels, from K-12 through college.

Bailin et al. (1999) emphasize the need for teaching students to use the intellectual resources constituting critical thinking in complex practices that include deliberation and discussion. To achieve this goal, they suggest that students must have access to examples of how each principle of critical thinking applies across a wide range of contexts. Moreover, the context of these examples must be rich enough to make clear the purpose of applying the principle. Instead of teaching isolated skills and dispositions, the authors focus on initiating children into communities of practice where good critical thinking practice is encouraged and poor critical thinking practice is discarded. Bailin et al. adhere to the infusion approach yet acknowledge the potential need for general critical thinking courses. Thus, their proposed conception of teaching critical thinking includes mentoring, authentic issues, dialogue, and making critical thinking principles explicit, like the recommendations above.

Halpern (1998) proposes a four-part model for enhancing critical thinking. Her model is based on general principles from cognitive psychology concerning how people learn. An important purpose of the model is to promote transfer across domains of knowledge, enabling people to become better thinkers—and in doing so, making the world better (Halpern, 2014). Halpern's (1998) model consists of a dispositional component, instruction in and practice with critical thinking skills, training activities designed to promote transfer, and a metacognitive component to direct and assess thinking. Halpern suggests that the dispositions and attitudes of a critical thinker

should be made clear to the learners to help them understand that critical thinking is effortful. General skills should also be made clear and include examples of how each skill applies in different situations. The skills could be learned more deeply by elaboration or other techniques that develop interconnected knowledge structures. In addition, students could practice retrieving the skills from memory and think about how to use those skills in novel and complex situations. Halpern suggests that real-world contexts should be used for this practice, aligning with the *situated cognition* viewpoint from cognitive science (Rogoff & Lave, 1984) and with Halpern's extended viewpoint which welcomes recall across domains. She presents examples of thoughtful questions that require learners to focus on structural aspects of a problem and thus help them recognize the need for a particular skill. These questions include asking students to draw a diagram to organize the information; and to explain why they selected a particular multiple-choice answer, which alternative is second best, and why. The final part of Halpern's (1998) four-part model, metacognitive monitoring, points to "the self-awareness and planning functions that guide the use of thinking skills" (p. 454). Halpern states that the metacognitive monitoring skills should also be made explicit. She recommends making the usually implicit thinking process into an explicit one by asking students guiding questions about, for example, how much time and effort a problem is worth, what they already know about the problem, what their goal is for engaging in critical deliberation about the problem, and how they will know then they have reached that goal.

Overall, this section shows that these proposed models for teaching critical thinking have many similarities with each other and with the strategies identified in the meta-analyses by Abrami et al. (2015; 2008). All of them focus on making critical thinking principles an explicit part of instruction, the use of authentic issues, and some form of mentoring. Ennis (2016) and Bailin et al. (1999) also include discussion as part of their models for teaching critical thinking. Thus, their conceptions of how to teach critical thinking are also somewhat commensurable with the focus on critical thinking as a social process in Thayer-Bacon's (2000) conception of constructive thinking. In sum, the general strategies that seem to be effective for teaching critical thinking are making critical thinking principles explicit, the opportunity for dialogue, the use of

authentic issues, mentoring, and teacher training. All of these strategies were as part of the intervention presented in Article III.

2.3.1 A note on dialogue and inquiry

As the ARGUMENT project has been particularly focused on using dialogue and inquiry as strategies for teaching and learning, these strategies warrant some further commentary. First, there is a lot of research on the effectiveness of these strategies that goes beyond research on using them merely for critical thinking instruction. Second, there are different types of dialogue and inquiry that have different purposes and different effects on learning. In this section, I present a selection of research on dialogue and inquiry and some of their empirical foundation as strategies for learning.

According to Howe et al. (2019), there are five commonly recurring themes of proposals about which characteristics classroom dialogue should exhibit to optimize student outcomes. First, instead of limiting their initiations to closed questions (e.g., questions with one correct answer that can be expressed in one sentence), teachers should include asking open questions that allow multiple answers (e.g., Alexander, 2008; Mercer & Littleton, 2007; O'Connor et al., 2015). Second, dialogue participants should contribute extensively by elaborating and expanding on previous contributions (e.g., Alexander, 2008; O'Connor et al., 2015). Third, the reasons on which opinions are based should be revealed through acknowledging, probing, and critiquing differences of opinion (e.g., Alexander, 2008; Lefstein, 2010; Mercer & Littleton, 2007; Reznitskaya & Gregory, 2013). Fourth, participants should pursue an integrated line of inquiry through making links between different contributions (e.g., Alexander, 2008; Mercer & Littleton, 2007; Reznitskaya & Gregory, 2013). Fifth, and finally, participants should take on a metacognitive perspective concerning their dialogue in order to evaluate and reflect on it in light of good practice (e.g., Lefstein, 2010; Reznitskaya & Gregory, 2013). All these dialogic features seem to relate to critical thinking. For example, critiquing differences of opinion includes and requires some skills related to interpretation and analysis of arguments, and taking on a metacognitive perspective to evaluate one's own dialogue have similarities with Halpern's (2014) focus on metacognition and Facione's (1990b) concept of self-

regulation. Because these five dialogic features have mostly been researched in relation to group discussions, Howe et al. (2019) conducted a large-scale study investigating these features in relation to teacher-student dialogue with the whole class, small groups, and individual students. They found that when students participated extensively (e.g., due to sufficient motivation), elaboration and probing (i.e., the second and third features) showed positive correlations with curriculum mastery. In addition, such elaboration was positively correlated with attitudes to school, self as a learner, and relationship with the teacher. The first, second, and third of these dialogic features were of particular focus in the ARGUMENT project through encouraging and supporting teachers to implement several opportunities for students to share their ideas before, during, and after group discussions.

Dialogues that display the abovementioned characteristics tend to be focused on the participants' joint construction of knowledge (e.g., Alexander, 2020). Such concepts of dialogue as co-construction of knowledge can also be viewed as a process of shared (dialogic) inquiry (e.g., Wells, 1999). Moreover, to teach students how to co-construct knowledge through dialogue might be at least as important as teaching students factual knowledge, which tends to change over time (Wegerif et al., 2020). This is also included in what Mercer (2008) calls exploratory talk (mentioned in Section 2.3), which is a "joint, co-ordinated form of co-reasoning, in which speakers share relevant knowledge, challenge ideas, evaluate evidence, consider options, and try to reach agreement in an equitable manner" (p. 95). He found that students in groups utilizing exploratory talk improved reasoning scores both individually and in groups, as well as math and science test scores, more than controls.

Asterhan and Schwarz (2016) note that there is a distinction between dialogues that are characterized by consensual co-construction (i.e., participants expand, elaborate, and explain ideas without challenging or critiquing each other's ideas) and dialogues that are argumentative. Although these dialogue types have many similarities, one important difference is their purpose (i.e., consensual co-construction uses explanation as a means to clarify while argumentation uses explanation to refute or convince). Argumentative dialogue is said to have a special role in higher order

thinking (Osborne & Patterson, 2011), which includes critical thinking. Moreover, explanation-driven discourse, while it may lead to incremental learning, might not be sufficient to radically reorganize conceptual knowledge (De Leeuw & Chi, 2003). Asterhan and Schwarz (2009) conducted an empirical study on dialogue characteristics and found that argumentation predicted dyadic and individual conceptual gains, while consensual co-construction did not. However, there is a distinction between disputative argumentation (i.e., where participants defend a viewpoint and seek to win the argument at the expense of the counterpart) and deliberative argumentation (i.e., where participants listen to and criticize each other's explanations without expressing interpersonal tension). Such deliberate argumentation seems to be somewhat similar to Mercer's (2008) concept of exploratory talk, and the disputative argumentation seems to be similar to what Mercer (2000) calls disputative talk. Mercer's (2000) third type of dialogue is called cumulative talk, and it consists of consensual co-construction where speakers uncritically co-construct a body of knowledge, somewhat similar to the type of dialogue discussed by Asterhan and Schwarz (2016). Of these three types (from Mercer), only exploratory talk is characterized by participants engaging constructively and critically with each other's ideas. This type of exploratory talk, or deliberative argumentation, aligns well with the third feature of effective dialogues (mentioned in the second paragraph of this section), that participants should critique differences of opinion. It also includes the fourth dialogic feature, that participants should pursue an integrated line of reasoning.

According to Mercer (2000), there is observational evidence suggesting that most of the naturally occurring classroom talk between students is cumulative or disputative, with only a few students being involved and with little coverage of the relevant topics. To increase the likelihood of any exploratory talk to occur, he suggests implementing certain conversational ground rules, which could include such things as sharing ideas, listening, not interrupting, giving reasons, questioning ideas, involving everybody, accepting responsibility, talking one at a time, and seeking mutual agreement. Implementation of such ground rules is likely to improve the quality of dialogues and might provide valuable lessons in how to conduct oneself in

discussions (e.g., Mercer et al., 2004). Moreover, some of the examples listed have similarities with critical thinking dispositions (e.g., Facione, 1990b), like for example, giving reasons and questioning ideas. In line with Mercer's (e.g., 2008) findings concerning exploratory talk being more effective for learning of reasoning and content than disputative talk, Asterhan and Schwarz (2016) posit that the effects of deliberative argumentation on promoting content learning are more favorable than those of disputative argumentation. Moreover, their review of findings indicates that argumentation is more effective for learning complex topics than superficial factual knowledge. Interestingly, a meta-analysis also found that learning through collaboration and discussion in small groups might be superior for transfer compared to individual learning (Pai et al., 2015). The ARGUMENT project suggested practices that would teach students to construct their own arguments and to improve those arguments through discussions. Moreover, suggestions for conversational ground rules and how to train students to follow these were presented and discussed at a workshop for the ARGUMENT teachers.

According to Hmelo-Silver et al. (2007), inquiry learning is focused on the learning of "content and discipline-specific reasoning skills and practices (often in scientific disciplines) by collaboratively engaging in investigations" (p. 100). Thus, it includes collaboration and dialogue. Moreover, Kolstø (2018) analyzed six well-known models for inquiry-based science teaching and found that all six suggest the use of dialogue, which further indicates the connection between dialogue and inquiry. Another common feature of inquiry learning is that it is organized around authentic and relevant issues (Hmelo-Silver et al., 2007). Students engage in sense-making concerning such issues, and they create evidence-based explanations and communicate their ideas, while the teacher acts as a facilitator that may provide support as needed. It is important to note that inquiry learning is not the same as unguided discovery learning (Hmelo-Silver et al., 2007), which is not an effective method for learning (Kirschner et al., 2006). Instead, as mentioned in Section 2.3, inquiry learning depends on many types of scaffolding to help guide students. One example of such scaffolds is the use of prompts to use certain reasoning strategies (Derry et al., 2006). Another example is to use argument mapping diagrams to guide

students when learning to distinguish between claims and evidence for those claims (e.g., Toth et al., 2002). The ARGUMENT project suggested use of a somewhat similar scaffold where students filled in components of arguments (e.g., claims and reasons). Providing students with explanations at the right time has also been shown to be an effective scaffold (Schwartz & Bransford, 1998; Schwartz et al., 2011), and allowing students to engage in an inquiry activity before some form of direct instruction could be more effective than beginning with the direct instruction. Scaffolds can also be used to reduce cognitive load, for example by using whiteboards to keep track of the investigation process (Hmelo-Silver, 2004) or by restricting the affordances of a software to make a task more manageable for learners (Quintana et al., 2004). Kolstø (2018) discusses how dialogue has been used to scaffold different parts of the inquiry process, including to help engage students in a problem, elicit prior knowledge, create tentative explanations, create plans to test these explanations, and test the explanations by comparing ideas and interpreting new data.

The particular types of authentic issues of inquiry that were used in the ARGUMENT project are social dilemmas with conceptual or technological connections to science, often called socioscientific issues (e.g., Sadler, 2004; Zeidler, 2014). As such issues are complex and open-ended dilemmas with no definitive answers, they are well suited for learning and practicing argumentation and informal reasoning (Kuhn, 1993) as well as critical thinking (Zeidler & Nichols, 2009). For example, such issues encourage the construction of opposing yet valid arguments that represent different perspectives and values. This is relevant to the informal reasoning and argumentation that is required in a democratic society (Kolstø, 2001). Providing students with opportunities to engage in informal reasoning is important if students are going to learn to think for themselves and construct their own arguments, especially when facing complex issues and their related evidence and data (Driver et al., 2000; Sadler, 2004). Although previous research has suggested that this is not trivial to accomplish, inquiry into socioscientific issues can provide a relevant context for learning argumentation and informal reasoning skills as well as conceptual understanding of science content (for a review, see Sadler, 2004). Thus, there seems to be a well-

founded theoretical rationale and empirical evidence supporting that inquiry (which includes dialogue and argumentation) using socioscientific issues can improve evidence-based decision making, critical thinking, character development, and learning of science content (Zeidler, 2014; Zeidler & Nichols, 2009).

2.4 Summary

This chapter lists several conceptions of critical thinking that are in alignment with "the mainstream concept" of critical thinking, as well as a few that oppose the mainstream concept. Most theorists view critical thinking as consisting of skills and dispositions, and they also view background knowledge as essential. Most researchers also agree that there are generalizable aspects of critical thinking, and many of these researchers believe that it is useful to focus on these aspects. There is substantial evidence supporting that critical thinking can be improved through education. Although transfer of critical thinking from education to other contexts is recognized as a lofty but difficult goal, there is a general agreement among researchers that such transfer is possible, and results from empirical studies support this.

There are several examples of instances where the strategies that are proposed to teach critical thinking has not led to desired outcomes. This points to a gap in the literature and implies a need for research into the specifics of what makes these general strategies effective. For example, the research on inquiry and dialogue, discussed in Section 2.3 and Section 2.3.1, might help inform efficient use of those types of learning strategies concerning critical thinking education. Moreover, it could be useful to know if there are any particular aspects of critical thinking that are more challenging than others and thus could be more important to make explicit through instruction. This aligns with the aims and research questions of this dissertation, presented in Section 1.2. Moreover, Ennis (1989) calls for the study of new approaches and instruments for evaluating critical thinking. This dissertation seeks to answer that call by using multiple-choice items from the Cornell Critical Thinking Test Level X in combination with written justifications.

3. Methodology

Methodology has been defined as the "theory of how research should be undertaken, including the theoretical and philosophical assumptions upon which research is based and the implications of these for the method or methods adopted" (Saunders et al., 2016, p. 719). This dissertation utilizes both quantitative and qualitative research methods. Thus, in this dissertation, I, like many other researchers who have conducted mixed methods research, adhere to the philosophy of pragmatism. Pragmatism is often characterized by seeking a middle ground between different philosophical dogmatisms and by rejecting traditional dualisms, like realism versus antirealism, free will versus determinism, and subjectivism versus objectivism (Johnson & Christensen, 2017). The pragmatist is more concerned about finding a workable solution, independent of the limits of the traditional quantitative and qualitative research paradigms. Rather than expecting to find the final truth, the pragmatist researcher emphasizes the continuous *process* of inquiry in an attempt to provide evidence that meets John Dewey's epistemological benchmark, *warranted assertability* (Johnson et al., 2017). This harmonizes with the *fundamental principle of mixed research*—a driving force of the design of this study—which states that mixed methods researchers should combine components from quantitative and qualitative research in a way that produces complementary strengths and non-overlapping weaknesses (Johnson & Christensen, 2017).

According to mixed methods research terminology, the overall design comprising all three studies of this dissertation could be considered a *complex hybrid design* because there are several phases of both quantitative and qualitative data collection and analyses, both in parallel and in sequence (Schoonenboom & Johnson, 2017). However, using a common notation system in mixed methods research (e.g., Morse & Niehaus, 2009), which has been upgraded by Johnson and Christensen (2017), the overall design of this study can be described as:

QUAN → QUAN + QUAL

This can be called an *equal-status, sequential-concurrent* design. The capitalization indicates that the quantitative methods (QUAN) and the qualitative methods (QUAL) are given equal status or importance. The arrow indicates that there are several sequential phases, and the plus sign indicates that the quantitative and qualitative components in the second phase were performed concurrently. More specifically, we first planned the quantitative component of the overall study which is presented in Article III (i.e., the quasi-experimental study with a pre-post design with control group). The first half of that data collection (i.e., the pretest for the study in Article III) was also used in Phase 1 of Article I (i.e., quantitatively exploring for challenging aspects of critical thinking), and during this analysis it became clear that a subsequent qualitative data collection and analysis could aid the overall aim of the dissertation. Thus, the final design evolved through the emerging (Creswell & Plano Clark, 2011) need for the qualitative component which was used in Article II (i.e., thematic analysis of students' reasoning) and in Phase 2 of Article I (i.e., quantification of qualitative data). In Article I, Phase 2 builds on Phase 1, and the main purpose of mixing quantitative and qualitative data was *triangulation* of results from several methods—a technique where methods, theories, or observers are combined to increase the validity of the results (Greene et al., 1989). In Article II, the purpose of quantifying qualitative data from the main qualitative analysis was *complementarity* (Greene et al., 1989), or what Bryman (2006) calls *illustration*, which means that the quantitative component was added to elaborate and show some patterns and frequencies that would otherwise go unnoted. Such complementarity was also another purpose for quantifying the qualitative data in Article I. Although Article II was born from the conclusion of Article I, the results and findings from these articles complement each other. The purely quantitative data collection and analysis for Article III was conducted in parallel with the data collection for the other two articles. The research questions of all three articles are closely related, and integration of the results and findings serves the overarching aim of the dissertation.

In this chapter, I first discuss the measures, their psychometric quality, and some questions related to their validity. Second, I describe the research participants of the studies. Third, I describe the three studies separately according to their aims, research

questions, designs, methods, and analyses. Finally, I discuss validity, the analytical software, and ethical considerations across the studies. An overview of the methodological aspects of the overall study can be seen in Table 1 at the end of this chapter.

3.1 Measures

3.1.1 The Cornell Critical Thinking Test Level X

The *Cornell Critical Thinking Test* (CCTT) Level X (Ennis & Millman, 2005) was used as the basis for data collection in all three studies. It is a multiple-choice test consisting of 71 items where each item has one keyed answer among three alternatives. The test is created in the form of a story about a crew from Earth that explore a new planet. The crew members represent various professions and have different fields of expertise, and test takers must make decisions based on the available information. The test measures general critical thinking skills, which are divided into *inductive inferences* (Part I), *observations* and *credibility* (Part II), *deductive inferences* (Part III and Part IV), and *assumptions* (Part IV). The results from the study in Article I indicated that Part II of the test was particularly interesting. Thus, after analyzing the test manual's proposed explanations to the keyed answers, we divided the items of Part II into five categories based on their solution strategies: o*bservation versus inference*, *conflict of interest*, *difference in method of observation*, *difference in authority/expertise*, and *other*. This process is described in Article I. Examples of items from each part of the test and from each of the categories within Part II can be seen in Appendix V.

The test makers strived to create the test so that only common everyday knowledge was needed to answer the items, and to write items that call for the employment of the selected critical thinking skills in contexts that most people will understand (Ennis et al., 2005). Moreover, the test makers discussed all items intensely and reached agreement about the correctness of the keyed answers. These last two points are important for the *construct validity* of the test—that is, the validity of inferences

about whether the test adequately measures the construct (i.e., critical thinking skill) it is intended to measure (Shadish et al., 2002). Furthermore, the test makers present eleven types of information that could be used as evidence in support of the construct validity of the CCTT Level X as a test of critical thinking ability. This information includes the rationale behind the test, the degree to which the test seems to fit that rationale, statistical analyses, correlations between the test and other tests that are intended to measure similar outcomes, correlations between the test and other variables, and results of experiments where the test was used to gauge critical thinking outcomes. *Reliability* estimates (i.e., internal consistency of test takers' responses across the items) are also used in support of the construct validity, with estimates ranging from .67 to .90 using Kuder-Richardson Formula 20 and 21 and .76 to .86 using the Spearman-Brown method, for which numbers above 0.70 have been deemed acceptable (de Vet et al., 2017). The internal consistency of the items can also be viewed as support for our chosen method of comparing groups of items to help identify challenging aspects of critical thinking. This implies that the items measure an overall construct of critical thinking. However, Ennis notes that critical thinking is a complex and heterogeneous construct. Thus, statistical methods might not find meaningful and separate factors within these items. But our use of qualitative analyses could, together with subscale scores, extract some meaningful differences. Overall, Ennis et al. (2005) maintain that there is justifiable support for the construct validity of the test as a measure of general critical thinking ability. However, there is less conclusive evidence concerning whether the psychometric quality of the subscales. The correlations between the total score and subscale scores are moderate to high, as is expected because there is overlap between sections. This poses potential challenges concerning interpreting the subscale scores, because such interpretation generally requires a real and reliable distinction between those scores (Liu et al., 2014). However, the test makers (i.e., Ennis et al., 2005) state that the subscale scores can still be used to provide indications about deficiencies of certain skills. Nevertheless, we have not relied solely on subscale scores in any of the studies, although we did include them as part of the evidence in Article I, where we also discuss the potential limitations concerning this. Moreover, the qualitative analysis of

written justifications for answers to test items (see Section 3.1.2 for the modification of the CCTT) could help outweigh the potential limitations related to insufficient psychometric data on the subscale scores (for a relevant discussion, see Norris, 1990).

The CCTT Level X was translated to Norwegian to match the language of the participants in the studies. This poses potential threats to validity. One commonly recommended method for strengthening the validity of a translation is to do a back-translation of the test (e.g., Hambleton et al., 2005), but the translator we used assured us that their method (i.e., no back-translation) was common procedure for academic purposes. In hindsight, insisting on a back-translation could have made the case for the validity of the translation more convincing. However, a pilot test was conducted in a sample that was deemed as being a part of the same population as the participants in the studies (i.e., a classroom in another school in Bergen). Moreover, the translated version was proofread by several researcher colleagues, and we had discussions about further adjustments to the translation. I also observed the administration of the test and took note of any items and wording that the students seemed to misunderstand, which is in line with one of the methods proposed to ensure the quality of a translation (Brislin, 1970).

3.1.2 Modification of the test

In line with my pragmatic stance towards mixing research paradigms and methods, I modified the test to include prompts for written justifications to selected items. Parts III and IV were excluded from the test, and the prompts were added to ten items of Part II that Phase 1 of the study in Article I identified as particularly interesting. The prompts were placed directly after each item. I will refer to this version of the test as the "modified" test or the "CCTT Level XWJ" (where WJ stands for written justifications). The modified test, which was used for the studies described in Article I and Article II, could provide insight into students' reasoning and whether the students provided correct explanations for their answers to the multiple-choice items. Modifying the test has potential implications for its psychometric qualities. However, the insights from qualitatively analyzing students' reasoning on the prompted test items could help strengthen validity as this reveals how students interpret and

understand the items. Moreover, Norris (1990) found that including verbal reports in a multiple-choice test of critical thinking did not alter test performance. In addition, two versions of the test were pilot tested before I decided on which items to prompt and where to place the prompt.

3.2 Participants

All participants were from six lower-secondary schools in Bergen, Norway. The students were eighth, ninth, and tenth graders and between 12 and 16 years old. Three of the schools were part of the ARGUMENT project and the three other schools were selected as non-treatment control schools mainly to help answer RQ3 (i.e., the research question of Article III). However, students from these schools also participated in the cross-sectional studies presented in Article I and Article II. The non-treatment schools were selected based on similarity (with treatment schools) of student scores on nationwide tests (Utdanningsdirektoratet, 2019). In general, schools in Norway seem to be more similar than not, and variables like socioeconomic background have less impact on student outcomes than in other countries (Kjærnsli & Jensen, 2016). The CCTT Level X was administered across all schools in the fall of 2019, and a total of 1353 students completed the test. This sample was used in Phase 1 of Article I, and as the pretest-sample for Article III. In the spring of 2020, the CCTT Level X was administered across all schools again, and a total of 1262 students completed the test. The students who completed both the pretest and the posttest ($N = 779$) were used as the final sample for Article III. The students in classrooms that due to hectic schedules had not taken the test in the fall of 2019 were administered the modified version of the test, CCTT Level XWJ. A total of 284 students completed this test, and these were used as the sample for Phase 2 of Article I and for Article II. Each sample contained students from all three grade levels. The distribution of students across grade levels, sex, and other variables, are described in the three articles.

3.3 Aims, research questions, designs, methods, and analyses

The overarching aim of this dissertation is to provide pragmatic (i.e. useful) knowledge of students' struggles when solving critical thinking problems, as well as to investigate whether supposedly effective yet general strategies for teaching critical thinking were effective in an ecologically valid setting (i.e., in classrooms in schools). The following sections present the aims, research questions, designs, methods, and analyses of each of the three articles that are the base of this dissertation.

3.3.1 Article I: Challenging aspects of critical thinking

Research investigating whether certain aspects of critical thinking could be particularly challenging for students is sparse. Moreover, while several tests of critical thinking offer subscale scores (e.g., Ennis & Millman, 2005; Halpern, 2010; Watson & Glaser, 1980, see Section 2.2), and there are reports of the difficulty of items for some tests (e.g., Leach et al., 2020, for the CCTT Level X), I am not aware of studies investigating whether subscale scores and difficult test items represent challenging aspects of critical thinking. Thus, the aim of the study in Article I was to investigate whether subscale scores and difficult items on the CCTT Level X might represent certain categories of critical thinking strategies that students struggle with. The research question of the study (RQ1) was: Which aspects of critical thinking are particularly challenging for lower-secondary school students?

The study design was cross-sectional, and the study had two phases. In Phase 1, the CCTT Level X was administered to the participants ($N = 1353$). We quantitatively analyzed the test data to look for patterns in the difficulty of individual items and across the four subscales of the test. To do this, we used visual inspection, sorting of items by difficulty, repeated measures ANOVAs, and pairwise comparison *t*-tests. This informed our choice of which items to include in the modified test in Phase 2, where the modified version of the test (CCTT Level XWJ) was administered to participants ($N = 284$) who had not been part of the data collection in Phase 1. We qualitatively analyzed written justifications according to a priori codes concerning

whether the reasons given were correct or incorrect (according to the solutions in the test manual). The prevalence of correct and incorrect reasons alongside correct and incorrect multiple-choice answers added further insight concerning the research question. Thus, Phase 2 included *enumeration*, or *quantification* of qualitative data (Johnson & Christensen, 2017). Using design typology from mixed methods research, the design of the study in Article I can be described as *explanatory sequential* (Fetters et al., 2013), where quantitative data are collected and analyzed first to inform the subsequent collection and analysis of qualitative data. Furthermore, the design can be described with notation from mixed methods research:

QUAN → QUAN + qual

This symbolizes a sequential design where the quantitative component comes first in time, and the capitalization of "QUAN" indicates that the design is *quantitatively driven* (i.e., the quantitative perspective is emphasized, and some qualitative data are added to the study).

3.3.2 Article II: Students' reasoning on test items

Accounts of reasoning related to commonly used assessments of critical thinking are uncommon in the literature. The aim of the study in Article II was to gain insight into how students reason when they are faced with the challenging aspects of critical thinking that were identified in Article I. Such insights could improve instruction and assessment of critical thinking. The research question we tried to answer in this study was (RQ2): How do students reason when faced with test items from these challenging critical thinking categories?

The design was cross-sectional. The data collection was the same as in Phase 2 of Article I. A thematic analysis (Braun & Clarke, 2006) of the participants' ($N = 284$) written justifications on the CCTT Level XWJ constituted the main part of the analysis. In addition, we quantified the qualitative data to show the frequency and distribution of the themes in student reasoning. Thus, the design can be described as a *qualitatively driven concurrent design* (Schoonenboom & Johnson, 2017), notated as:

QUAL + quan

This indicates that the qualitative perspective is emphasized, and that the quantitative component is added to complement the qualitative component and show patterns and frequencies that could otherwise remain elusive.

3.3.3 Article III: Group-comparison of gains from pretest to posttest

Considering the growing demand for improved critical thinking outcomes from the education system, there is potential for broad implementation of effective critical thinking instruction through curricula changes and teacher education. Meta-analyses (Abrami et al., 2015; Abrami et al., 2008) have pointed to several effective strategies for critical thinking instruction. The aim of the study in Article III was to see if the implementation of these general strategies in a teacher professional development project would lead to improved student outcomes on a test of general critical thinking skills. More specifically, the strategies used in the study were teacher training, inquiry into authentic issues, opportunity for dialogue, and making critical thinking an explicit part of instruction. The research question we tried to answer (RQ3) was: Do students from schools that are part of the project improve their general critical thinking skills more than students from non-treatment schools?

To answer this research question, the study employed a quasi-experimental design. The CCTT Level X was administered to the participants ($N = 779$) as a pretest and a posttest, and gains in scores of the two groups were compared. The data were analyzed using *t*-tests, which were deemed more useful than ANCOVA considering that the study's research question asks whether the average gain in score is different for the two groups (Wright, 2006). On the other hand, ANCOVA could be used to answer whether average gain in score, accounting for score on the pre-test, is different for the groups. It is recommended to use ANCOVA sparingly, perhaps only for randomized controlled trials (Fitzmaurice et al., 2011; Oakes & Feldman, 2001), and even then there are recommendations to use a *t*-test of gain in score (Maxwell & Delaney, 1990). In sum, analysis of gain in score seems to be a reliable and unbiased

estimate of true change (Rogosa, 1988). Thus, the analyses of this study, alongside the design and method, were all purely quantitative.

3.4 Validity

From a pragmatic philosophical point of view, validity and truth are related to practical usefulness—for example, a claim can be said to be true if viewing it as true leads to useful practical implications. However, in their widely known book on experimental methodology, Shadish et al. (2002) use the term validity to refer to "the approximate truth of an inference" (p. 34). In their view, validity relates to inferences and is not a property of designs and methods, because the same design can contribute to inferences that differ in their level of validity. Thus, questions of validity are answered through qualitative judgments of inferences about validity, and therefore there is no way to guarantee validity. One can strengthen the validity of a study's inferences by striving to mitigate potential threats to validity by thoughtful design and consideration of each threat to validity in the context of the particular study. This section will discuss potential threats to validity concerning the studies in this dissertation, as well as methodological choices we have made to mitigate these threats and thus strengthen the validity of our inferences.

Shadish et al. (2002) presents a typology of four well-known kinds of validity. *Statistical conclusion validity* is a qualitative judgment of inferences about the correlation between treatment and outcome as well as the strength of that correlation. *Internal validity,* or *causal validity* (Johnson & Christensen, 2017), is a qualitative judgment of inferences about whether the observed correlation between treatment and outcome is due to a causal relationship between them. *Construct validity* is a qualitative judgment of inferences about the abstract constructs that are represented by the particulars (i.e., settings, treatments, observations, and persons) of a study. *External validity*, or *generalizing validity* (Johnson & Christensen, 2017), is a qualitative judgment of inferences about whether the conclusions of a study generalize to other persons, settings, treatment variables, and outcome variables. In addition to these four main types, we will also discuss how to mitigate potential

threats to validity related to qualitative and mixed research, and how these are dealt with in our methodology, results, and inferred conclusions.

Concerning statistical conclusion validity, the large sample sizes used in this research ensured sufficient power to detect relatively small effects. One threat to this type of validity is the lack of standardized implementation of the interventions described in Article III. We observed some classrooms, but not all, and thus we cannot know the degree to which the interventions were implemented according to our vision. Another potential threat to statistical conclusion validity for all three studies is related to the heterogeneity of students' performance on the tests, which to some degree seemed to be due to differences in motivation.

Internal validity is only relevant for the study in Article III, as it is the only study with the appropriate aim and design for making any inferences about a causal relationship between treatment and outcome. Some threats to internal validity are mitigated by the quasi-experimental design with a pretest, a posttest, and a control group. Because of the lack of random sampling in this study, the main threat to internal validity is the potential for *selection effects*—that is, systematic differences between the students in the treatment and control group. However, as described in Section 3.2, students in Norwegian schools are more similar than not. Another threat to this type of validity is that the interventions represented a relatively small amount of the total classroom time of a whole semester, and thus there could be confounding variables related to all the other teaching that occurred in each group. Nevertheless, it seems somewhat unlikely that these variables would shift the outcome towards any particular direction, although they could introduce noise.

Critical thinking is an important construct for all three studies. The main potential threat to construct validity, whether the CCTT seems to measure the construct (or constructs) it is intended to measure, has already been covered (i.e., the validity of the test is covered in Section 3.1, and the conceptualization of critical thinking is covered in Section 2.1). I don't see any obvious threats to construct validity regarding the participants, treatments, and settings. Participant constructs, such as *student*, *girl*, and

boy, should be understandable; the treatment is a thoroughly described construct; and, finally, the setting of *Norwegian lower secondary school* is described as a relatively homogeneous construct (see Section 3.2).

The biggest threats to external validity for the quasi-experimental study in Article III are potential interactions with treatment variations (i.e., the implementation of the interventions) and potential interactions with variations in students (although, as described in Section 3.2, this might not be a large threat).

For the qualitative methods and analyses, there are some specific aspects of validity, or *trustworthiness* (Johnson & Christensen, 2017), that should be discussed. For the thematic analysis presented in Article II, the validity depends on the theoretical and analytical approaches (Braun & Clarke, 2006) which are described in the article. The validity is also strengthened by the included examples of students' written justifications for each theme and by cross-checking the identified themes with the dataset. Moreover, we discussed the coding together with three other educational researchers in seminars. The validity of this coding process is further strengthened by providing a measure of the interrater reliability using both percent agreement and *proportional overlap*, a modified version of Cohen's kappa, proposed by Mezzich et al. (1981) that is suitable for mutually non-exclusive codes (i.e., there were several possible code-combinations for each answer). The interrater reliability for the coding of correct and incorrect reasons in Article I is also provided, both as a percent agreement and a Cohen's Kappa score with excellent agreement (Cicchetti, 1994; Cohen, 1960). For the study in Article I, and for the dissertation as a whole, the use of multiple data-sources (i.e., both versions of the test) strengthens validity. In addition, we strived towards continuous and critical self-reflection of our own biases and assumptions (e.g., from the theoretical framework), a strategy that Johnson and Christensen (2017) call *reflexivity.* We have also tried to rule out alternative explanations, a process which was aided by the anonymous peers that reviewed our articles. All our studies have fairly large sample sizes, especially compared to what is common in qualitative research (Mason, 2010). Thus, the external validity is probably stronger than in other studies using qualitative methods. However, this also means

that the participants are described as a group instead of as individuals with detailed descriptions for each person. Nevertheless, the external validity, or *transferability*, could be judged by the reader, on a case-to-case basis, based on their evaluation of the similarities between the particulars of these studies and their study (Firestone, 1993; Lincoln & Guba, 1985). Moreover, analytic generalization (i.e., generalization of results or cases to a broader theory) of the results can support and/or expand relevant critical thinking theories (Firestone, 1993), and this type of external validity is potentially strengthened by the diversity brought forth by the high number of participants in these studies.

In mixed research, one key point concerning validity, or *legitimation* (Onwuegbuzie & Johnson, 2006), is to address all the pertinent "validities" of the quantitative, qualitative, and mixed aspects of the research. The quantitative and qualitative aspects of validity have been covered. The first pertinent type of validity in mixed research is *inside-outside legitimation*, which is the extent to which the researcher is able to understand and use both the *emic* viewpoint (i.e., the subjective insider view of the participant) and the *etic* viewpoint (i.e., the objective outside view of the researcher). While analyzing students' written justifications, we strived to understand the students' reasoning from their point of view. At the same time, we analyzed both the qualitative and quantitative data in light of our theoretical rationale. Moreover, we have strived to integrate the quantitative and qualitative data, methods, and analyses, and to draw meta-inferences, or conclusions, based on this integration.

3.5 Ethics

The ARGUMENT project was approved by the Norwegian Centre for Research Data (NSD), and this approval included all three studies presented in this dissertation. Schools, students, and parents were informed about the studies (see Appendix VI). Participation was voluntary and no person-identifying information was collected. A self-generated ID-code was used to connect the pretest and posttest used in Article III. The data collection and studies were discussed with The Data Protection Official at the University of Bergen. Our assessment is that the overall study is in alignment

with the principles of beneficence, respect, and justice put forth in *the Belmont Report* (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978). In this regard, our assessment of the treatments, data collection, and studies is that they pose minimal risk to the participants and offer large potential benefits for participants, researchers, education, and society. The teaching strategies we used are research backed and in alignment with the new core curriculum for primary and secondary education in Norway (Ministry of Education and Research, 2017). Furthermore, both teachers and students could benefit from the critical thinking assessment. We provided teachers with reports on how the students scored, including the teachers of the classrooms that participated in the pilot tests. Moreover, the results, and a discussion of the results, were presented for teachers in workshops and at a digital conference.

3.6 Analytical software

The main statistical analyses were conducted in R (R Development Core Team, 2020). Datasets were sorted using spreadsheets, which were also used for the simple coding and counting of correct and incorrect responses for the study in Article I. The thematic analysis in Article II was conducted in NVivo 12 Pro (QSR International, 2018). The power analyses in Article III were conducted in GPower (Faul et al., 2007).

3.7 Overview of methodology

Table 1 – Overview of aspects related to methodology and methods of the overall study and its components

Aim of the dissertation	To provide pragmatic (i.e. useful) knowledge of students' struggles when solving critical thinking problems, as well as to investigate whether supposedly effective yet general strategies for teaching critical thinking are effective in an ecologically valid setting		
Article	I	II	III
Title	Aspects of critical thinking that are challenging for students – Conflict of interest, observation, and inference	Students' reasoning when faced with test items of challenging aspects of critical thinking	Large-scale study suggests supposedly effective strategies for teaching critical thinking might be too general
Research question(s)	**RQ1**: Which aspects of critical thinking are particularly challenging for lower-secondary school students?	**RQ2**: How do students reason when faced with test items from these challenging critical thinking categories?	**RQ3**: Do students from schools that are part of the project improve their general critical thinking skills more than students from non-treatment schools?
Design	Cross-sectional	Cross-sectional	Quasi-experimental
Method	QUAN → QUAN + qual	QUAL + quan	QUAN
Sample size	*Phase 1:* 1353 *Phase 2:* 284	284	779
Data	*Phase 1 (fall 2019):* CCTT Level X *Phase 2 (spring 2020):* CCTT Level XWJ	CCTT Level XWJ (*spring 2020)*	*Pretest (fall 2019):* CCTT Level X *Posttest (spring 2020):* CCTT Level X
Analyses	*t*-test, repeated measures ANOVA, quantification of qualitative data	Thematic analysis, quantification of qualitative data	*t*-test

Note. CCTT Level X: Cornell Critical Thinking Test Level X. CCTT Level XWJ: Cornell Critical Thinking Test Level X modified to include written justifications to selected items.

4. Results

The work towards achieving the overarching aim of this dissertation is presented in the three enclosed articles. In this chapter, I will summarize the results from each of these articles.

4.1 Article I

The aim of the study in Article I was to investigate whether subscale scores and difficult items on the CCTT Level X could represent certain aspects, or categories, of critical thinking strategies that students struggle with. The quantitative analyses (N = 1353) in Phase 1 showed that items in Part II of the *Cornell Critical Thinking Test* (CCTT) Level X had the lowest proportions of correct answers. Furthermore, students' average scores were significantly lower on Part II of the test than on the complete test, $p < .001$, $d = 0.31$. Part II of the test calls for skills related to observations and credibility, and the items require test takers to evaluate the credibility of sources, statements, and observations. We had divided the items of Part II into five categories based on their proposed solution strategies: o*bservation versus inference*, *conflict of interest*, *difference in method of observation*, *difference in authority/expertise*, and *other*. Three categories stood out as particularly challenging. First, in the observation-versus-inference category the proportions of correct answers ranged from 7% to 20%. Second, in the conflict-of-interest category the proportion of correct answers ranged from 27% to 28%. Third, in the method-of-observation category the proportion of correct answers ranged from 38% to 46%. For each of these three categories within Part II, students scored significantly ($p < .001$) lower on average than on Part II as a whole, with small to very large effect sizes (d = [0.35 – 1.45]; Sawilowsky, 2009).

In Phase 2 of the study, we coded students' (n = 284) written justifications on the modified CCTT as either "right reason" or "wrong reason". Quantification of these data provided insight into students' reasoning, and the results supported that the items in the three identified categories were in fact difficult due to the skills and knowledge

needed to solve them. First, independent of their chosen answer being right or wrong, less than one third of the students touched upon the correct reason. Second, on seven of the ten items that asked for written justifications, less than one third of the responses that contained the right answer did not also contain the right reason. This could indicate that most students who chose the correct answers to these items were guessing or unable to explain or justify their answers.

These results suggest that there are challenging aspects of critical thinking related to evaluating the credibility of observations and statements. More specifically, the most challenging categories of critical thinking items identified in Article I require that test takers are able to discern observations from inferences, recognize conflicts of interests, and use these skills when evaluating the credibility of statements.

4.2 Article II

The aim of the study in Article II was to gain insight into how students reason when they are faced with the challenging aspects of critical thinking that were identified in Article I. The thematic analysis of students' (N = 284) written justifications to items representing these aspects identified 21 lower-level themes encompassed by six higher-level themes of reasoning (seen in Table 2 of Article II). More than a quarter of the responses (28%) expressed strong inductive logic but were based on irrelevant or incorrect premises. That is, when the students seemingly failed to identify the correct reason yet still wrote a reason, they used alternative evidence as premises (14% of responses), or they changed the context (14% of responses) either by misunderstanding the information or by using their own experience, generalizations, or made-up information so their reasoning matched their chosen answer. Only 3% of the responses did not seem to represent strong inductive logic. Most of these responses (81%) expressed the belief that an inference is just as, or more, believable than an observation, and around 14% of these responses expressed a recognition of the difference between observations and inferences without it impacting the choice of answer. For example, on one of the observation-versus-inference items in Part II of the test (see Appendix V for examples of items), students were asked to identify

which of two statements were most believable, or if they were equally believable. One of the statements was an observation, while the other was an inference based on that observation. One student wrote, “A pinpoints something, B uses what A says to figure out that he must be some sort of leader. Therefore C is the correct answer.” Thus, the student seems to recognize that there is a difference between the two statements, and that one statement is inferred from the other, without being able to arrive at the correct answer. These types of reasons were probably making logical sense from the students’ points of view, indicating a potential lack of knowledge or skills related to judging whether observations or inferences are more believable. In 29% of the responses, students did not give a substantial reason for their answer. Moreover, 26% of responses expressed a lack of basis to choose an answer (i.e., choose alternative A or B). Finally, 13% of the responses contained a correct reason, and 72% of these responses were also related to choosing the correct answer.

4.3 Article III

The aim of the study in Article III was to see if supposedly effective strategies for teaching critical thinking, that were implemented in a teacher professional development project, would lead to improved student outcomes on a test of general critical thinking skills. The strategies used in the study were teacher training, inquiry into authentic issues, opportunity for dialogue, and making critical thinking an explicit part of instruction. Although the total sample ($N = 779$) did increase its average critical thinking score significantly from pretest to posttest, $p < .001$, $d = 0.15$, we did not detect a significant difference in gain in scores between the treatment group and the non-treatment control group. For the individual grade levels in the total sample, the only significant improvement in critical thinking score from pretest to posttest was found in the eighth grade, $p < .001$, $d = 0.24$. These effect sizes range from very small to small, according to Sawilowsky (2009).

The test in this study was considered a low stakes test because it would not affect students’ grades. This could affect motivation and performance. Self-efficacy measured at pretest significantly predicted a small portion of the variance in posttest

scores, $\beta = 0.27$, $p < .001$, $R^2 = 0.07$. However, using low-end cut-offs for time spent and score to remove unserious answers did not affect the significance of the results of our analyses.

These results warrant a discussion (see Article III) about whether the strategies for critical thinking instruction described in the literature could be too general. Furthermore, this implies a need for more detailed knowledge concerning the specifics that make these strategies effective in some cases and not in others.

5. Discussion

This chapter starts with a discussion of the results considering the research questions and existing research. Next, I propose some potential contributions of this dissertation, followed by a discussion of implications for instruction. I also cover some strengths and limitations of the work presented in this dissertation and reflect on potential aims for future research. Finally, I leave the reader with some concluding remarks.

5.1 Discussion of results

The challenging (i.e., for these students) aspects of critical thinking identified in Article I are related to the items in Part II of the *Cornell Critical Thinking Test* (CCTT) Level X. These items require the skill of evaluation, defined by Facione (1990b) as being able to "assess the credibility of statements" and to "assess the logical strength of the actual or intend inferential relationships among statements" (p. 15). Furthermore, the identified difficult aspects of critical thinking—observation versus inference, conflict of interest, and, to a lesser extent, method of observation—require concrete sub-skills of evaluation. For example, the items in the observation-versus-inference category require that test takers are able to assess the degree to which a statement is inferential or observational and include the result of this assessment when deciding how believable that statement is. It is important to note that observations, in general, can be theory-laden, as perception is affected by experiences and knowledge (Estany, 2001). Thus, observations might not always be more believable than inferences. However, the type of difference (i.e., between observations and inferences) found in the items of the CCTT Level X is between observation descriptions and inferences that are potential causes or implications of those observations. Moreover, the items in the observation-versus-inference category do not simultaneously contain any obvious conflict of interest or difference in expertise or authority among the sources of the statements. It therefore seems reasonable to assume that observations are more believable than inferences in the given context, like it is stated in the solution in the test manual (Ennis et al., 2005).

The items in the conflict-of-interest category require an awareness of cognitive biases (e.g., conflict of interest), the ability to recognize such biases in the source of a statement, and the ability to take this into account when judging the believability of that statement. As discussed in Section 1.1, these skills (i.e., those related to observation versus inference and conflict of interest) are important for digital literacy, and previous research has indicated that students struggle with tasks requiring these types of skills. For example, students often fail to recognize biases, like conflicts of interests, in online news, ads, and websites (Nygren & Guath, 2019; Wineburg et al., 2016) and struggle to evaluate the credibility of online news (Ku et al., 2019). Concerning skills related to observation and inference, Abd-El-Khalick et al. (2002) state that novice learners tend to struggle with distinguishing between observations and inferences. I have not seen any research where items from the CCTT have been explored in relation to conflict of interest or observation versus inference. Nor have I seen any research that has identified these as particularly challenging aspects of critical thinking, like the study presented in Article I has done. However, Leach et al. (2020) present mean scores for items on the CCTT Level X based on data from 2265 fifth graders, and if one ranks these mean scores from highest to lowest it becomes clear that there is substantial overlap with the most difficult items identified in Article I.

Article II investigated how students reason when facing these difficult items and discusses potential barriers to successfully solve them. Thus, Article II is to an extent also based on the findings from Article I and these two articles complement each other—Article I identified possible challenging aspect of critical thinking and Article II analyzed the students' reasoning when facing these aspects. It is important to note that the test results would not impact students' grades. Thus, this was a low stakes test context. There are also indications that the motivation for taking the test was low. For example, the most common themes of responses were related to not choosing alternative A or B and not giving any particular reason for the chosen answer. Moreover, some students wrote that they were unmotivated and tired. However, most of these types of responses were removed from the dataset. Low motivation in combination with low stakes contexts can affect performance on tests, in general

(Cole & Osterlind, 2008), as well as performance on tests of critical thinking skills and dispositions (Bensley et al., 2016; Dehghani et al., 2011; Kim et al., 2015). Thus, the motivation and context could be the main barriers, and many students would perhaps express higher levels of critical thinking in a different context. With this caveat in mind, the critical thinking and reasoning that was expressed in the context of the study, and possible barriers related to this, can be discussed.

In addition to motivation and context, Article II also points to barriers related to skills, knowledge, and dispositions. As the items clearly require evaluation skills, not being proficient enough in such skills could be one potential barrier to solve the items. Furthermore, as many students seemed to misunderstand the available information, not being proficient enough in the skill of interpretation is another potential barrier. For example, concerning the conflict-of-interest items, many students seemed to overlook the fact that only one person had something to gain and instead wrote that both the involved parties had a conflict of interest. In addition, the finding that many students used alternative evidence in their reasoning could indicate a lack in the skill of inference and its subskills querying evidence, conjecturing alternatives, and drawing conclusions (Facione, 1990b). In addition to barriers related to skills, there could be barriers related to a lack of knowledge of key critical concepts (Bailin et al., 1999). For example, knowing that an observation is generally more probable to be true than an inference, and knowing how a conflict of interest affects credibility, could be helpful or necessary for solving the items in the identified difficult categories. Moreover, some would argue that such pieces of knowledge can be expanded into procedural knowledge, making it part of a process somewhat similar to general skills (Smith, 2002). For example, the knowledge of what an observation and an inference is could be expanded into a more general skill by describing it as a process of recognizing observations and inferences, and then judging whether the observation is more believable than the inference. Thus, when students learn and practice critical thinking skills it could be useful to also focus on operational knowledge, such as principles and step-by-step procedures, abstracted from good critical thinking. Such knowledge could enable or support the learning of critical thinking skills through deliberate practice. Finally, although the CCTT is not

designed to measure dispositions, the high number of responses expressing that there was no basis to choose either alternative A or alternative B over the other could indicate a lack of certain dispositions of critical thinking, like having self-confidence in one's own ability to reason. Nevertheless, this could also be related to motivation and test context, like discussed in the preceding paragraph.

Article II also points out that students seemed to use the solution strategy from an example item at the start of Part II of the test on several subsequent items that had similar surface characteristics but called for a different solution strategy. This type of fixed thinking, often called a *mental set* (Schultz & Searleman, 2002), is another potential barrier to solving the items. Moreover, once someone has made a decision on which answer to choose, they tend to justify and elevate that answer until it becomes the dominant option in their mind—a process which is called *dominance structuring* (Facione, 2020). This could also help explain, for example, why students used alternative evidence when justifying their chosen answer. One common type of response where students used alternative evidence was to base the reasoning on the education, experience, or expertise of the people involved in the items. For example, 144 responses claimed that the medical expert's statements were most believable. This type of reasoning, appeal to authority, is listed both as a valid argument and a fallacious argument depending on the context (Walton, 2008). In fact, for some items on this test, appeal to authority is the proposed correct reasoning. However, appeal to authority is not the proposed correct reasoning for any of the chosen items that included prompts for written justifications. Here, appeal to authority is not necessarily fallacious, but students often seemed to misunderstand the context in some way or seemed to fail to notice other more relevant evidence.

In cases where students were not able to notice the conflict of interest or the difference between the observation and the inference, they might have answered intuitively by using fast thinking, which could lead to incorrect answers as these students are not expert critical thinkers (Kahneman, 2011). After these students chose their answers, they were asked to justify them. This could lead to some sort of (backwards) rationalization (e.g., Evans & Wason, 1976; Kahneman, 2011), which

could also help explain why some students used alternative evidence, or made-up evidence (i.e., changed the context), as premises in their reasoning. They might have been searching, after the fact, for a premise to an intuitively reached conclusion.

Another interesting theme of reasoning identified in Article II, *logical fallacy*, is related to the observation-versus-inference items. One example of such an item is seen in Appendix V. It shows one statement that represents an observation and one statement that is an inference related to that observation (i.e., one worker says, "Five times now the person in the blue jacket has talked to someone and pointed, and immediately they have run off in the direction he pointed," and another worker says, "He must be the leader"). If the inferential statement (i.e., "He must be the leader") is true, the observational statement should also probably be true. However, the observational statement could be true without the inferential statement being true. Thus, answering that the inference statement is more or as believable as the observation statement could be seen as a logical fallacy having similarities with the *conjunction fallacy* (Tversky & Kahneman, 1983), which is the logical fallacy of judging the probability of a conjunction to exceed the probability of its constituents. Article II found that 66 written justifications contained this type of fallacy. According to Tversky and Kahneman (1983), committing this type of fallacy could indicate that students were using intuitive heuristics—that is, they were thinking fast instead of more deliberatively evaluating the problem at hand.

I have not seen any published research using written justifications alongside the multiple-choice items of the CCTT or any other test of critical thinking. However, Ennis (1996b) presents some promising yet preliminary results from using a similar method to reveal students' dispositions. His results from using such a method provide one potential solution to what he sees as some fundamental problems of assessing critical thinking dispositions, namely, that they are not directly observable and that "a disposition is something we want students to evidence on their own—without being pushed or prompted to evidence it" (p. 175). Moreover, Norris (1990) combined multiple-choice critical thinking test items with verbal reports. However, his goal was to investigate whether the inclusion of verbal reports influenced test performance,

which it did not. Thus, Norris posits that inclusion of verbal reports might provide useful validation data for critical thinking tests. This matches my observations when exploring the dataset—I found no clear indications of differences between performance on the modified test and the original test. Besides, there is evidence to suggest that using this type of method can improve learning. One study found that combining written justifications with multiple-choice questions of physics concepts in a peer instruction setting increased the likelihood of choosing the correct answers (Koretsky et al., 2016). This could indicate that there is potential in using items from the CCTT in a peer instruction setting with the aim of improving critical thinking, something which could be explored by future research.

The results from the study in Article III indicate that teacher training, dialogue, inquiry into authentic issues, and making critical thinking principles explicit through instruction are not necessarily enough to improve students' critical thinking outcomes more than in controls. These strategies have been shown to be effective in numerous other studies (for reviews, see Abrami et al., 2015; Abrami et al., 2008). Thus, the results of the study could imply that these strategies are too general and that there is a need for more insight regarding what makes the strategies effective in some cases and not in others.

A large body of evidence shows that teacher training in the form of professional development programs can affect student learning in general (Blank & Alas, 2009; Egert et al., 2018), as well as student critical thinking (Abrami et al., 2008). There are also studies that have found no effect on CCTT scores, specifically, after training instructors for critical thinking interventions (e.g., MacPhail-Wilcox et al., 1990; Zohar & Tamir, 1993). Article III discusses that discrepancies in the effectiveness of teacher training for critical thinking outcomes could be related to variations in the duration of teacher training and points to studies that have investigated how the duration of teacher training affects different learning outcomes. For example, a meta-analysis of experimental and quasi-experimental studies found that an average of 49 hours per year of professional development improved student achievement by 21 percent (Yoon et al., 2007). Furthermore, programs with durations totaling from five

to 14 hours showed no statistically significant effect on student learning. Banilower et al. (2006) showed that consistent student achievement results in science and math were found when teachers received over 100 hours of professional development.

Although there is clear evidence for the effectiveness of dialogic teaching in general (e.g., Alexander, 2020) and for critical thinking (Abrami et al., 2015), there are several studies that report no gains in critical thinking scores from using dialogue (Bixler et al., 2015; Garside, 1996; Pellegrino, 2007). Notably, these studies do not provide detailed accounts of how group work and discussions were conducted. Thus, readers do not get insight into the type of dialogue that were used and how discussions were structured. Some dialogue types are more effective for learning than others (e.g., exploratory talk; Mercer, 2008), and unstructured discussions can easily lead to aimless sharing of personal experiences without trace of critical thinking (Angeli et al., 2003). On the other hand, there is evidence to suggest that structured and teacher-led discussions lead to more favorable critical thinking outcomes (Abrami et al., 2015; Yang et al., 2008).

For inquiry to be an effective strategy for teaching and learning, the inquiry must be highly scaffolded to provide sufficient guidance for learners (Hmelo-Silver et al., 2007; Kolstø et al., 2006). However, some studies that have found no effect on critical thinking skills from using inquiry into authentic issues seem to provide scaffolds and guidance for aspects not directly related to critical thinking. For example, in a study by Gul and Akcay (2020) the participants were provided with guidance on theories regarding the socioscientific issues that served as the objects of inquiry, while learning critical thinking skills seemed to be left as an implicitly expected outcome. Moreover, students in a study by Beavers et al. (2017) got feedback on their planning and implementation of their authentic practice task, but did not seem to get any direct guidance for practicing critical thinking, nor was there any explication of critical thinking principles. Thus, it could be a good idea to include scaffolds and guidance regarding learning critical thinking when using inquiry into authentic issues, and in general.

Some of the strategies that are proven to be effective for teaching critical thinking, such as dialogue and inquiry, are employed in contexts of collaboration and social interactions. In such contexts, students get to share ideas and give and receive feedback on issues that (ideally) trigger engagement. Social interaction, for example through dialogue, might engage students and thus increase the motivation for critical thinking (e.g., Kolstø, 2018). This contrasts the testing context of the CCTT, which is purely individual. The test is designed as a story, but it may or may not engage all the students, and students don't necessarily care about making an effort on tests (Zamarro et al., 2016). This possible tension between the context of learning and practice and the context of the test raises an interesting question: Are the students able to express their critical thinking skills to the same degree in each of these two different contexts? One could speculate that the individual context of the test might limit at least some aspects of critical thinking skills that the students would be able to express through dialogue and inquiry in an authentic context. Thus, the test might not reveal the students' full potential for critical thinking. Nevertheless, the research that has identified dialogue and inquiry as effective strategies for critical thinking is largely based on these types of individual tests (Abrami et al., 2015), indicating that students are able to transfer some critical thinking skills between collaborative learning contexts and individual testing contexts. Moreover, the choice of test, the interpretation of the findings, the findings' inferred implications for instruction, and the congruence between these elements, are affected by the choice of critical thinking conception. Using another concept or conception (the distinction between concept and conception can be seen in Section 2.1) of critical thinking, more focused on the social aspects—such as the concept by Thayer-Bacon (2000) and, to some extent, the conception by Bailin et al. (1999)—could very well lead to differences in methodology, findings, and implications for instruction.

Concerning the study in Article III, explication of critical thinking principles did occur and was observed in lessons, but in the observed lessons it occurred to a substantially smaller degree than the other strategies. Abrami et al. (2008) found that making critical thinking principles explicit was one of the most important means for making instruction in critical thinking effective. As discussed in this section, the

results from the studies in Article I and Article II indicate several skills and knowledge components that could be made explicit during instruction.

5.2 Implications for instruction

The discussion in Article III suggests some characteristics that could be important for ensuring the efficacy of supposedly effective (yet perhaps too general) strategies for teaching critical thinking—which could provide guidance for instruction and future projects with similarities to the ARGUMENT project. Notably, critical thinking principles should be made explicit in instruction through scaffolding, structured dialogue, and modeling, for example when working with authentic issues. This should probably also be emphasized in teacher training programs. The results from Article I and Article II complement this suggestion as well as each other in the following ways.

Article I points to certain knowledge components and skills that might be required to solve items representing challenging aspects of critical thinking. For example, knowing what a conflict of interest is and how it could affect credibility, and knowing the difference between an observation and an inference, could be important or essential to solve some of these types of items. Thus, this knowledge should be made explicit in instruction and could be implemented into skill practice with the aim of improving sub-skills of evaluation, such as recognizing a conflict of interest, discerning an observation from an inference, and taking this into account when deciding about the credibility of a source or statement. Teachers can explicate this knowledge and how to use these skills, for example through dialogue, when working with socioscientific (i.e., authentic) issues, especially considering that such authentic issues tend to include several conflicts of interest. Relevant contexts to focus on skills and knowledge related to observations and inferences include science experiments and subsequent writing of lab reports, as well as topics concerning the nature and history of science. In both these contexts, there is a need to distinguish between observations (e.g., in lab experiments) and inferences (e.g., what inferences scientists historically have made based on available data).

Article II indicates that students might struggle with using the correct evidence when facing items requiring these types of skills and knowledge components (i.e., skills and knowledge related to conflict of interest and observation versus inference). The students often used alternative evidence (14%) or changed the context of the problems (14%). Although the incidence of these types of reasoning might be lower in a more authentic context than in the test context, these findings could have implications for instruction. Inquiry into authentic issues could provide situations that require evaluation of conflicts of interests as well as observations and inferences. When using authentic issues as part of instruction—as such issues also tend to contain a lot of contrasting evidence—teachers could support learning of critical thinking by guiding students to use subskills of evaluation that could mitigate struggles related to using alternative evidence or changing the context. Some examples of such subskills are weighing alternative solutions and evidence, gathering enough evidence before inferring a conclusion, and understanding and interpreting the problem and its context.

5.3 Contributions

This section contains an overview of some potentially important contributions from this dissertation. The results from the quantitative and qualitative analyses of Article I and Article II identify some difficult aspects of critical thinking as well as students' reasoning when facing these aspects, which in turn point to potential barriers related to solving items representing these difficult aspects of critical thinking. These results represent empirical contributions—not least due to their implications for instruction, discussed in Section 5.2 above. Furthermore, Articles I and II provide empirical and methodological contributions concerning how to interpret results from multiple-choice tests of critical thinking in relation to validity, as both our results and our methods can help unveil how test takers (mis)interpret the items of such tests. Concerning (mis)interpretation of the items, our results showed that students sometimes make up their own evidence, or in other ways change the information provided by the context, to use as evidence in their reasoning. Concerning the

method, written justifications can reveal instances where students mark the correct answer but not due to the correct reason, and instances where students mark an incorrect answer yet provide a correct reason. Thus, our method also enables test administrators to base their scoring on a rationale of their choosing, and not just on the rationale of the solution in the test manual. For example, one could give credit to responses showing that test takers weigh different alternatives, even if the marked answer is incorrect according to the test manual. Our method of combining written justifications with multiple-choice items brings more nuanced insight into students' thinking. Thus, another methodological contribution is related to the development of new or improvement of existing critical thinking tests. One such possibility is to develop tests that identify and score other aspects of critical thinking (e.g., dispositions and intuition) than the skill aspect measured by the original test, through analysis of the written justifications.

The study in Article III achieved results that were not expected based on data from the literature concerning the teaching strategies that were used. The large number of participants and the use of a standardized test means that these data could be useful to include in future meta-analyses. Moreover, the article's ensuing discussion contributes by pointing out aspects of the teaching strategies that could be important for the effectiveness of said strategies. This could provide some ideas when designing interventions in future research, as well as when designing instruction using the teaching strategies in question. In addition, the translation and digitalization of the test, and its use in students across six Norwegian lower-secondary schools represent a few potentially important contributions. First, our work in pilot testing and ensuring the quality of the translation strengthens validity. In addition, the inclusion of written justifications allows for further qualitative validation of the test, a topic which we have touched upon several places (e.g., in Article II). Moreover, the data we have gathered could be used for further quantitative validation of the test in this type of setting. The data will be made available to interested researchers at dataverse.no. Second, our work has in some ways laid the foundation for easy implementation of this critical thinking test in Norwegian schools, either for summative or formative purposes—as well as for research.

5.4 Strengths

The mixing of quantitative and qualitative methods in study design, data collection, and analyses is a strength of the research in this dissertation (Johnson & Christensen, 2017). Furthermore, as far as I am aware, the studies in Article I and Article II are the first published studies that combine written justifications with items from the CCTT, something which has potential implications that are discussed in Section 5.2. Moreover, the relatively high number of participants in all three studies strengthens external validity and transferability of our inferences based on the results. The sample sizes in the first two articles are larger than what is typically seen in qualitative research (Mason, 2010)—and concerning Article III, the fact that most of the students in six schools participated in the project warrants considering this a large-scale study. In addition, the external validity of the research project is strengthened due to using ecologically valid units, treatments, and settings (Shadish et al., 2002). Finally, both versions of the test are easy to implement by other researchers and instructors in Norway (especially considering that we have translated the test and put the CCTT Level X and Level XWJ into a digital survey solution) and other countries.

5.5 Limitations

While there is justifiable support for the construct validity of the CCTT Level X as a measure of general critical thinking ability, there is less conclusive evidence for the psychometric quality of the subscales (Ennis et al., 2005). However, the test makers (i.e., Ennis et al.) maintain that the subscale scores can still be used to give indications about deficiencies in certain skills. Potential limitations concerning the validity of the test, including its translation, are discussed in Section 3.1.

Phase 1 of the study in Article I was exploratory and did not contain any *a priori* hypotheses or systematic assessment of potential reasons for items being difficult. Thus, although the results from Phase 2 of the study strengthen our conclusions, caution is advised when interpreting these results and conclusions until further research is conducted.

The large sample sizes of the first two articles comes at the cost of the length and depth of the qualitative data (i.e., the written justifications were mostly one to a few sentences long). Although this length was enough to identify several themes of reasoning, longer responses could have provided deeper insight into why students reasoned the way they did.

The lack of random sampling of student participants represents a potential limitation (see Section 3.4), especially for the quasi-experiment in Article III. Moreover, we have little knowledge of what happened in control schools, except that they were not part of the ARGUMENT project or any similar project. In addition, most intervention studies risk being limited by implementation fidelity. This might also be the case in this study as our insight into the implementation mainly comes from observations of around one classroom per grade level in each of the three project schools, as well as teacher reports at follow-up workshops. Implementation fidelity might have been more affected in the second half of the implementation due to lockdowns related to COVID-19. Finally, the relatively short duration of each module of the intervention from the study in Article III is another possible limitation.

5.6 Future research

More research is needed to support the tentative conclusions from Article I concerning whether observation versus inference and conflict of interest represent challenging aspects of critical thinking. For example, future research should investigate if the results of Article I can be replicated if the subscales are administered in a different order, thus accounting for fatigue towards the end of the test. Moreover, studies could try to replicate our findings by using the same method across different settings and populations, but also by creating more authentic and ill-structured assessment situations. For example, students may or may not have the same struggles with taking conflicts of interests into account when evaluating the credibility of statements in a more complex setting, such as in a discussion of an authentic issue. Students also may or may not have the same struggles in a setting that triggers stronger engagement, which a discussion of an authentic issue could do.

Moreover, their reasoning might differ from the themes that we identified in Article II.

Future research should also explore alternative testing methods that account for conceptions of critical thinking as a collective activity where dialogue and collaborative inquiry stimulate thoroughness—especially as thoroughness and motivation seemed to be potentially missing for some of the participants of the studies presented here. There might be other types of tests that are more suited to other concepts and conceptions of critical thinking than the individual type of test used here.

Future research should investigate whether there is a correlation between the critical thinking expressed in this type of well-structured testing context and in other more ill-structured real-life contexts. Furthermore, there is a need to investigate whether learning to master these difficult aspects of critical thinking in a well-structured context will lead to spontaneous transfer to more authentic contexts.

Future research should also explore how administering the test with higher stakes for students affects the results. This, as well as random sampling, could reduce potential effects of differences between students (e.g., in motivation and reading ability).

Future research should gather more extensive qualitative data concerning students' justifications to selected difficult items of the CCTT Level X, for example through interviews. This could provide further insights into why students reasoned the way they did, which in turn could have further implications for instruction.

An aim of future research should be to get more detailed insight into what makes the strategies discussed in Article III effective in some but not in all cases. Moreover, these strategies could be combined in several ways. For example, research should investigate the effects of training teachers to become competent in making critical thinking principles explicit in their teaching in different contexts and subjects. Furthermore, research should investigate the effect of preparing teachers to guide

students by highlighting critical thinking principles as relevant situations to do so occur, such as when having discussions concerning authentic issues.

Importantly, future research should investigate whether the results from Article I and Article II, combined with ideas from the discussion of Article III, could improve critical thinking instruction (see Section 5.2 for suggestions on how to implement this in instruction). For example, instruction that explicates the knowledge and skills related to the difficult aspects of critical thinking that were identified in Article I might be effective in improving critical thinking.

5.7 Concluding remarks

One of my earliest goals with this work was to assess how the ARGUMENT project affected students' critical thinking skills. In addition, I wanted to investigate how teachers' use of teaching materials from the ARGUMENT project (e.g., the critical thinking poster, seen in Appendix I) would affect critical thinking as expressed in classroom and group discussions. However, neither the poster nor other critical thinking materials was initially implemented by the teachers. Thus, my goal shifted further towards investigating students' critical thinking and reasoning, and potential barriers they faced in these regards, through variants of the Cornell Critical Thinking Test. The studies presented in Articles I and II were done within the context of the ARGUMENT project, although these articles focus only on exploring one aspect of the project's aims, critical thinking. Article III took on a more outside view of the project and measured its effects on critical thinking compared to controls.

Although all findings and conclusions are still tentative (e.g., due to low stakes testing context, inconclusive validity data regarding subscale scores, etc.), I maintain that the conclusions are justified by the available data. Hopefully, and especially after further validation and insight from future work (e.g., as suggested in Section 5.6), these findings can contribute to and guide instruction and the design of other similar projects in the future. The identified difficult aspects of critical thinking and students' reasoning when facing these aspects could indicate some directions for research, and

possibly also, with time, which aspects of critical thinking skills and knowledge instruction should focus on.

References

Abd-El-Khalick, F., Lederman, N. G., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire (VNOS): Toward valid and meaningful assessment of learners'conceptions of nature of science. *Journal of Research in Science Teaching*, *39*(6), 497-521. https://doi.org/10.1002/tea.10034

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, *85*(2), 275-314. https://doi.org/10.3102/0034654314551063

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research*, *78*(4), 1102-1134. https://doi.org/10.3102/0034654308326084

Albert, H. (1985). *Treatise on critical reason*. Princeton University Press. http://www.jstor.org/stable/j.ctt7ztkd9

Alexander, R. (2008). *Towards dialogic teaching: Rethinking classroom talk*. Dialogos.

Alexander, R. (2020). Dialogic pedagogy in a post-truth world. In N. Mercer, R. Wegerif, & L. Major (Eds.), *The Routledge international handbook of research on dialogic education* (pp. 672-686). Routledge.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2000). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. Pearson.

Angeli, C., Valanides, N., & Bonk, C. (2003). Communication in a web-based conferencing system: The quality of computer-mediated interactions. *British Journal of Educational Technology*, *34*, 31-43. https://doi.org/10.1111/1467-8535.d01-4

Association of American Colleges and Universities. (2011). *The LEAP vision for learning: Outcomes, practices, impact, and employer's views*.

Asterhan, C. S. C., & Schwarz, B. B. (2009). Argumentation and explanation in conceptual change: Indications from protocol analyses of peer-to-peer dialog. *Cognitive Science*, *33*(3), 374-400. https://doi.org/10.1111/j.1551-6709.2009.01017.x

Asterhan, C. S. C., & Schwarz, B. B. (2016). Argumentation for learning: Well-trodden paths and unexplored territories. *Educational Psychologist*, *51*(2), 164-187. https://doi.org/10.1080/00461520.2016.1155458

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

Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing critical thinking. *Journal of Curriculum Studies*, *31*(3), 285-302. https://doi.org/10.1080/002202799183133

Bailin, S., & Siegel, H. (2003). Critical thinking. In N. Blake, P. Smeyers, R. Smith, & P. Standish (Eds.), *The Blackwell guide to the philosophy of education* (pp. 181-193). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470996294.ch11

Banilower, E., Boyd, S., Pasley, J., & Weiss, I. (2006). Lessons from a decade of mathematics and science reform: A Capstone report for the local systemic change through teacher enhancement initiative. *Horizon Research, Inc.*

Beavers, E., Orange, A., & Kirkwood, D. (2017). Fostering critical and reflective thinking in an authentic learning situation. *Journal of Early Childhood Teacher Education*, *38*(1), 3-18. https://doi.org/10.1080/10901027.2016.1274693

Bellaera, L., Weinstein-Jones, Y., Ilie, S., & Baker, S. T. (2021). Critical thinking in practice: The priorities and practices of instructors teaching in higher education. *Thinking Skills and Creativity*, *41*, Article 100856. https://doi.org/10.1016/j.tsc.2021.100856

Bensley, D. A., Rainey, C., Murtagh, M. P., Flinn, J. A., Maschiocchi, C., Bernhardt, P. C., & Kuehne, S. (2016). Closing the assessment loop on critical thinking: The challenges of multidimensional testing and low test-taking motivation. *Thinking Skills and Creativity*, *21*, 158-168. https://doi.org/10.1016/j.tsc.2016.06.006

Bernard, R. M., Zhang, D., Abrami, P. C., Sicoly, F., Borokhovski, E., & Surkes, M. A. (2008). Exploring the structure of the Watson–Glaser Critical Thinking Appraisal: One scale or many subscales? *Thinking Skills and Creativity*, *3*(1), 15-22. https://doi.org/10.1016/j.tsc.2007.11.001

Biesta, G. J. J., & Stams, G. J. J. M. (2001). Critical thinking and the question of critique: Some lessons from deconstruction. *Studies in Philosophy and Education*, *20*(1), 57-74. https://doi.org/10.1023/A:1005290910306

Bixler, G. M., Brown, A., Way, D., Ledford, C., & Mahan, J. D. (2015). Collaborative concept mapping and critical thinking in fourth-year medical students. *Clinical Pediatrics*, *54*(9), 833-839. https://doi.org/10.1177/0009922815590223

Bjønness, B., & Kolstø, S. (2015). Scaffolding open inquiry: How a teacher provides students with structure and space. *NorDiNa*, *11*, 223-237. https://doi.org/10.5617/nordina.878

Blank, R. K., & Alas, N. d. l. (2009). Effects of teacher professional development on gains in student achievement: How meta analysis provides scientific evidence useful to education leaders. *Society for Research on Educational Effectiveness*.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Longmans, Green.

Bouygues, H. L. (2018). *The state of critical thinking: A new look at reasoning at home, school, and work*. https://reboot-foundation.org/wp-content/uploads/_docs/REBOOT_FOUNDATION_WHITE_PAPER.pdf

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77-101. https://doi.org/10.1191/1478088706qp063oa

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27-40. https://doi.org/10.1080/08957347.2012.635502

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*(3), 185-216. https://doi.org/10.1177/135910457000100301

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research*, *6*(1), 97-113. https://doi.org/10.1177/1468794106058877

Butler, H. A., Dwyer, C. P., Hogan, M. J., Franco, A., Rivas, S. F., Saiz, C., & Almeida, L. S. (2012). The Halpern Critical Thinking Assessment and real-world outcomes: Cross-national applications. *Thinking Skills and Creativity*, *7*(2), 112-121. https://doi.org/10.1016/j.tsc.2012.04.001

Butler, H. A., Pentoney, C., & Bong, M. P. (2017). Predicting real-world outcomes: Critical thinking ability is a better predictor of life decisions than intelligence. *Thinking Skills and Creativity*, *25*, 38-46. https://doi.org/10.1016/j.tsc.2017.06.005

Cáceres, M., Nussbaum, M., & Ortiz, J. (2020). Integrating critical thinking into the classroom: A teacher's perspective. *Thinking Skills and Creativity*, *37*, 100674. https://doi.org/10.1016/j.tsc.2020.100674

Cargas, S. (2016). Honoring controversy: Using real-world problems to teach critical thinking in honors courses. *Honors in Practice*, *12*, 123-137.

Case, R., & Wright, I. (1997). Taking seriously the teaching of critical thinking. *Canadian Social Studies*, *32*(1), 12-19.

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*.

Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instrument in psychology. *Psychological Assessment*, *6*, 284-290. https://doi.org/10.1037/1040-3590.6.4.284

Clifford, J. S., Boufal, M. M., & Kurtz, J. E. (2004). Personality traits and critical thinking skills in college students: Empirical tests of a two-factor theory. *Assessment*, *11*(2), 169-176. https://doi.org/10.1177/1073191104263250

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Cole, J., & Osterlind, S. (2008). Investigating differences between low- and high-stakes Test performance on a general education exam. *The Journal of General Education*, *57*, 119-130. https://doi.org/10.1353/jge.0.0018

Crawford, B. A. (2014). From inquiry to scientific practices in the science classroom. In *Handbook of Research on Science Education*. Routledge. https://doi.org/10.4324/9780203097267.ch26

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* SAGE Publications.

Dam, G., & Volman, M. (2004). Critical thinking as a citizenship competence: Teaching strategies. *Learning and Instruction*, *14*, 359-379. https://doi.org/10.1016/j.learninstruc.2004.01.005

Davies, M. (2013). Critical thinking and the disciplines reconsidered. *Higher Education Research and Development*, *32*, 529-544. https://doi.org/10.1080/07294360.2012.697878

De Leeuw, N., & Chi, M. T. H. (2003). The role of self-explanation in conceptual change learning. In G. Sinatra & P. Pintrich (Eds.), *Intentional conceptual change* (pp. 55-78). Lawrence Erlbaum.

de Vet, H. C. W., Mokkink, L. B., Mosmuller, D. G., & Terwee, C. B. (2017). Spearman–Brown prophecy formula and Cronbach's alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, *85*, 45-49. https://doi.org/10.1016/j.jclinepi.2017.01.013

Dehghani, M., sani, H. J., Pakmehr, H., & Malekzadeh, A. (2011). Relationship between students' critical thinking and self-efficacy beliefs in Ferdowsi

University of Mashhad, Iran. *Procedia - Social and Behavioral Sciences*, *15*, 2952-2955. https://doi.org/10.1016/j.sbspro.2011.04.221

Derry, S. J., Hmelo-Silver, C. E., Nagarajan, A., Chernobilsky, E., & Beitzel, B. D. (2006). Cognitive transfer revisited: Can we exploit new media to solve old problems on a large scale? *Journal of Educational Computing Research*, *35*(2), 145-162. https://doi.org/10.2190/0576-R724-T149-5432

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, *84*(3), 287-312. https://doi.org/10.1002/(SICI)1098-237X(200005)84:3<287::AID-SCE1>3.0.CO;2-A

Egert, F., Fukkink, R. G., & Eckhardt, A. G. (2018). Impact of in-service professional development programs for early childhood teachers on quality ratings and child outcomes: A meta-analysis. *Review of Educational Research*, *88*(3), 401-433. https://doi.org/10.3102/0034654317751918

Ejiogu, K. C., Yang, Z., & Trent, J. D. (2006). *Understanding the relationship between critical thinking and job performance* [Poster presentation]. 21st annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching Thinking Skills: Theory and Practice.* (pp. 9-26). Freeman.

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. (1996a). *Critical thinking*. Prentice-Hall.

Ennis, R. H. (1996b). Critical thinking dispositions: Their nature and assessability. *Informal Logic*, *18*(2), 165-182. https://doi.org/10.22329/il.v18i2.2378

Ennis, R. H. (2016). Critical thinking across the curriculum: A vision. *Topoi*, *37*(1), 165-184. https://doi.org/10.1007/s11245-016-9401-4

Ennis, R. H., & Millman, J. (2005). Cornell Critical Thinking Test: Level X In (5 ed.). Seaside, CA: Critical Thinking Company.

Ennis, R. H., Millman, J., & Tomko, T. N. (2005). *Cornell critical thinking tests level X & level Z manual* (5 ed.). Critical thinking.

Ennis, R. H., & Weir, E. E. (1985). *The Ennis-Weir critical thinking essay test: An instrument for teaching and testing*. Midwest Publications.

Erduran, S., & Dagher, Z. R. (2014a). Aims and values of science. In *Reconceptualizing the nature of science for science education*. Springer. https://doi.org/10.1007/978-94-017-9057-4_3

Erduran, S., & Dagher, Z. R. (2014b). Scientific practices. In *Reconceptualizing the nature of science for science education*. Springer. https://doi.org/10.1007/978-94-017-9057-4_4

Estany, A. (2001). The thesis of theory-laden observation in the light of cognitive psychology. *Philosophy of Science*, *68*(2), 203-217. https://doi.org/10.1086/392873

Evans, J. S. B. T., & Wason, P. C. (1976). Rationalization in a reasoning task. *British Journal of Psychology*, *67*(4), 479-486. https://doi.org/10.1111/j.2044-8295.1976.tb01536.x

Facione, P. A. (1990a). *The California Critical Thinking Skills Test: College level experimental validation and content validity*.

Facione, P. A. (1990b). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction* [ERIC Document Reproduction Service No. ED315423]. American Philosophical Association.

Facione, P. A. (2020). *Critical thinking: What it is and why it counts* (M. R. LLC, Ed. 9th ed.). Measured Reasons LLC and Insight Assessment.

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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs-principles and practices. *Health services research*, *48*(6 Pt 2), 2134-2156. https://doi.org/10.1111/1475-6773.12117

Firestone, W. (1993). Alternative Arguments for Generalizing From Data as Applied to Qualitative Research. *Educational Researcher*, *22*, 16-23. https://doi.org/10.3102/0013189X022004016

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis*. Wiley. https://books.google.no/books?id=qOmxRtdNJpEC

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The Nature of Intelligence* (pp. 231-235). Erlbaum.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, *18*(3), 253-292. https://doi.org/10.1016/0010-0285(86)90001-0

Garside, C. (1996). Look who's talking: A comparison of lecture and group discussion teaching strategies in developing critical thinking skills. *Communication Education*, *45*(3), 212-227. https://doi.org/10.1080/03634529609379050

Gehrke, P. J. (2018). Ecological validity. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. Thousand Oaks,, California.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*(3), 255-274. https://doi.org/10.2307/1163620

Gul, M. D., & Akcay, H. (2020). Structuring a new socioscientific sssues (SSI) based instruction model: Impacts on pre-service science teachers' (PSTs) critical thinking skills and dispositions. *International Journal of Research in Education and Science*, *6*, 141-159. https://doi.org/10.46328/ijres.v6i1.785

Gunawardena, M., & Wilson, K. (2021). Scaffolding students' critical thinking: A process not an end game. *Thinking Skills and Creativity*, *41*, 100848. https://doi.org/10.1016/j.tsc.2021.100848

Guyton, E. M. (1988). Critical thinking and political participation: Development and assessment of a causal model. *Theory and Research in Social Education*, *16*(1), 23-49. https://doi.org/10.1080/00933104.1988.10505554

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, *53*(4), 449-455. https://doi.org/10.1037/0003-066X.53.4.449

Halpern, D. F. (2001). Assessing the effectiveness of critical thinking instruction. *The Journal of General Education*, *50*, 270 - 286. https://doi.org/10.1353/jge.2001.0024

Halpern, D. F. (2010). Halpern Critical Thinkning Assessment. In: SHUHFRIED (Vienna Test System).

Halpern, D. F. (2014). *Thought and knowledge: An introduction to critical thinking* (5th ed.). Psychology Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Lawrence Elbaum, Mahwah.

Hand, B., Shelley, M. C., Laugerman, M., Fostvedt, L., & Therrien, W. (2018). Improving critical thinking growth for disadvantaged groups within elementary school science: A randomized controlled trial using the Science Writing Heuristic approach. *Science Education*, *102*(4), 693-710. https://doi.org/10.1002/sce.21341

Hernstein, R. J., Nickerson, R. S., de Sánchez, M., & Swets, J. A. (1986). Teaching thinking skills. *American Psychologist*, *41*(11), 1279-1289. https://doi.org/10.1037/0003-066X.41.11.1279

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review*, *16*(3), 235-266. https://doi.org/10.1023/B:EDPR.0000034022.16470.f3

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark. *Educational Psychologist*, *42*(2), 99-107. https://doi.org/10.1080/00461520701263368

Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher–student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences*, *28*(4-5), 462-512. https://doi.org/10.1080/10508406.2019.1573730

Huang, M.-Y., Tu, H.-Y., Wang, W.-Y., Chen, J.-F., Yu, Y.-T., & Chou, C.-C. (2017). Effects of cooperative learning and concept mapping intervention on critical thinking and basketball skills in elementary school. *Thinking Skills and Creativity*, *23*, 207-216. https://doi.org/10.1016/j.tsc.2017.01.002

Jaswal, V. K., & Neely, L. A. (2006). Adults don't always know best: Preschoolers use past reliability over age when learning new words. *Psychological Science*, *17*(9), 757-758. https://doi.org/10.1111/j.1467-9280.2006.01778.x

Johnson, R. B., & Christensen, L. B. (2017). *Educational research: Quantitative, qualitative, and mixed approaches* (5th ed.). SAGE Publications, Inc.

Johnson, R. B., de Waal, C., Stefurak, T., & Hildebrand, D. L. (2017). Understanding the philosophical positions of classical and neopragmatists for mixed methods research. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *69*(2), 63-86. https://doi.org/10.1007/s11577-017-0452-3

Jaarsveld, S., Lachmann, T., & van Leeuwen, C. (2012). Creative reasoning across developmental levels: Convergence and divergence in problem creation. *Intelligence*, *40*(2), 172-188. https://doi.org/10.1016/j.intell.2012.01.002

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus, and Giroux.

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

Kim, H., Lee, E., & Park, S.-Y. (2015). Critical thinking disposition, self-efficacy, and stress of Korean nursing students. *Indian Journal of Science and Technology*, *8*. https://doi.org/10.17485/ijst/2015/v8i18/76710

Kirschner, P., Sweller, J., & Clark, R. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*. https://doi.org/10.1207/s15326985ep4102_1

Kjærnsli, M., & Jensen, F. (2016). *Stø kurs. Norske elevers kompeanse i naturfag, matematikk og lesing i PISA*. Universitetsforlaget.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, *76*(6), 1261-1277. https://doi.org/10.1111/j.1467-8624.2005.00849.x

Kolstø, S. D. (2001). Scientific literacy for citizenship: Tools for dealing with the science dimension of controversial socioscientific issues. *Science Education*, *85*(3), 291-310. https://doi.org/10.1002/sce.1011

Kolstø, S. D. (2018). Use of dialogue to scaffold students' inquiry-based learning. *Nordic Studies in Science Education*, *14*(2), 154-169. https://doi.org/10.5617/nordina.6164

Kolstø, S. D., Bungum, B., Arnesen, E., Isnes, A., Kristensen, T., Mathiassen, K., . . . Ulvik, M. (2006). Science students' critical examination of scientific information related to socioscientific issues. *Science Education*, *90*(4), 632-655. https://doi.org/10.1002/sce.20133

Koretsky, M. D., Brooks, B. J., & Higgins, A. Z. (2016). Written justifications to multiple-choice concept questions during active learning in class. *International Journal of Science Education*, *38*(11), 1747-1765. https://doi.org/10.1080/09500693.2016.1214303

Ku, K. Y. L. (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

Ku, K. Y. L., Kong, Q., Song, Y., Deng, L., Kang, Y., & Hu, A. (2019). What predicts adolescents' critical thinking about real-life news? The roles of social media news consumption and news media literacy. *Thinking Skills and*

Creativity, *33*(September 2019), 1-12. https://doi.org/10.1016/j.tsc.2019.05.004

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education*, *77*(3), 319-337. https://doi.org/10.1002/sce.3730770306

Kuhn, D. (2015). Thinking together and alone. *Educational Researcher*, *44*(1), 46-53. https://doi.org/10.3102/0013189x15569530

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

Leach, J., & Scott, P. (2003). Individual and sociocultural views of learning in science education. *Science & Education*, *12*(1), 91-113. https://doi.org/10.1023/A:1022665519862

Leach, S. M., Immekus, J. C., French, B. F., & Hand, B. (2020). The factorial validity of the Cornell Critical Thinking Tests: A multi-analytic approach. *Thinking Skills and Creativity*, *37*(September 2020), 1-14. https://doi.org/10.1016/j.tsc.2020.100676

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389-405. https://doi.org/10.1023/A:1025779619903

Lee, H.-S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education*, *24*(2), 115-136. https://doi.org/10.1080/08957347.2011.554604

Lefstein, A. (2010). More helpful as problem than solution: Some implications of situating dialogue in classrooms. In K. Littleton & C. Howe (Eds.), *Educational dialogues: Understanding and promoting productive interaction* (pp. 170-191). Routledge.

Li, Y., Li, K., Wei, W., Dong, J., Wang, C., Fu, Y., . . . Peng, X. (2021). Critical thinking, emotional intelligence and conflict management styles of medical students: A cross-sectional study. *Thinking Skills and Creativity*, *40*, 100799. https://doi.org/10.1016/j.tsc.2021.100799

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage Publications.

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

Loes, C. N., & Pascarella, E. T. (2017). Collaborative learning and critical thinking: Testing the link. *The Journal of Higher Education*, *88*(5), 726-753. https://doi.org/10.1080/00221546.2017.1291257

MacPhail-Wilcox, B., Dreyden, J., & Eason, E. (1990). An investigation of Paideia program effects on students' critical thinking skills. *Educational Considerations*, *17*(2), 61-67.

Marshall, J. D. (2001). A critical theory of the self: Wittgenstein, Nietzsche, Foucault. *Studies in Philosophy and Education*, *20*(1), 75-91. https://doi.org/10.1023/A:1005243027145

Mason, M. (2010). Sample size and saturation in PhD studies using qualitative interviews. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *11*(3). https://doi.org/10.17169/fqs-11.3.1428

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Wadsworth/Thomson Learning.

McPeck, J. E. (1981). *Critical thinking and education* (Vol. 30). St. Martin's Press.

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

Mercer, N. (2000). *Words and minds: How we use language to think together*. Routledge.

Mercer, N. (2008). Talk and the development of reasoning and understanding. *Human Development*, *51*(1), 90-100. https://doi.org/10.1159/000113158

Mercer, N., Dawes, L., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science [10.1080/01411920410001689689]. *British Educational Research Journal*, *30*(3), 359-377. https://doi.org/https://doi.org/10.1080/01411920410001689689

Mercer, N., & Littleton, K. (2007). *Dialogue and the development of children's thinking: A sociocultural approach* (1st ed.). Routledge.

Mezzich, J. E., Kraemer, H. C., Worthington, D. R. L., & Coffman, G. A. (1981). Assessment of agreement among several raters formulating multiple diagnoses. *Journal of Psychiatric Research*, *16*(1), 29-39. https://doi.org/10.1016/0022-3956(81)90011-X

Ministry of Education and Research. (2017). *Core curriculum – values and principles for primary and secondary education*.

Moon, J. (2008). *Critical thinking: An exploration of theory and practice*. Routledge.

Morse, J., & Niehaus, L. (2009). *Mixed method design: Principles and procedures*. Lef Coast Press.

Moseley, D., Baumfield, V., Elliott, J., Gregson, M., Higgins, S., Miller, J., & Newton, D. (2005). *Frameworks for thinking: A handbook for teaching and learning*. Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511489914

Moseley, D., Baumfield, V., Higgins, S., Lin, M., Miller, J., Newton, D., . . . Gregson, M. (2004). *Thinking skill frameworks for post-16 learners: An evaluation. A research report for the Learning and Skills Research Centre*.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1978). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*.

National Research Council. (2000). *How people learn: Brain, mind, experience, and school*. The National Academies Press. https://doi.org/https://doi.org/10.17226/9853

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press. http://public.eblib.com/choice/publicfullrecord.aspx?p=3378982

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. National Academies Press. http://site.ebrary.com/id/10863742

Norris, S. P. (1990). Effect of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement*, *27*(1), 41-58. https://doi.org/10.1111/j.1745-3984.1990.tb00733.x

Nygren, T., & Guath, M. (2019). Swedish teenagers' difficulties and abilities to determine digital news credibility. *Nordicom Review*, *40*, 23 - 42. https://doi.org/10.2478/nor-2019-0002

O'Connor, C., Michaels, S., & Chapin, S. H. (2015). 'Scaling down' to explore the role of talk in learning: From district intervention to controlled classroom study. In L. B. Resmick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 111-126). American Educational Research Association.

Oakes, J. M., & Feldman, H. A. (2001). Statistical power for nonequivalent pretest-posttest designs. The impact of change-score versus ANCOVA models. *Eval Rev*, *25*(1), 3-28. https://doi.org/10.1177/0193841x0102500101

OECD. (2015). *Education 2030 project proposal*. OECD Publishing.

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the schools*, *13*(1), 48-63. https://www.proquest.com/scholarly-journals/validity-issue-mixed-research/docview/211030483/se-2?accountid=8579

Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education*, *95*(4), 627-638. https://doi.org/10.1002/sce.20438

Pai, H.-H., Sears, D. A., & Maeda, Y. (2015). Effects of small-group learning on transfer: A meta-analysis. *Educational Psychology Review*, *27*(1), 79-102. https://doi.org/10.1007/s10648-014-9260-8

Pal, A., & Banerjee, S. (2019). Understanding online falsehood from the perspective of social problem. In I. E. Chiluwa & S. A. Samoilenko (Eds.), *Handbook of Research on Deception, Fake News, and Misinformation Online*. IGI Global. https://doi.org/10.4018/978-1-5225-8535-0.ch001

Paul, R. (1992). Critical thinking: What, why, and how. *New Directions for Community Colleges*, *1992*(77), 3-24. https://doi.org/10.1002/cc.36819927703

Pellegrino, A. (2007). *The manifestation of critical thinking and metacognition in secondary American history students through the implementation of lesson plans and activities consistent with historical thinking skills* (Publication Number UMI No. 3282653) [Doctoral dissertation, ProQuest Dissertations and Theses database.

Pérez-Escolar, M., Ordóñez-Olmedo, E., & Alcaide-Pulido, P. (2021). Fact-checking skills and project-based learning about infodemic and disinformation. *Thinking Skills and Creativity*, *41*, Article 100887. https://doi.org/10.1016/j.tsc.2021.100887

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

QSR International. (2018). *NVivo 12 Pro*. In QSR International.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., . . . Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences*, *13*(3), 337-386. https://doi.org/10.1207/s15327809jls1303_4

R Development Core Team. (2020). *R: A language and environment for statistical computing*. In R Foundation for Statistical Computing. https://www.R-project.org/

Reznitskaya, A., & Gregory, M. (2013). Student thought and classroom language: Examining the mechanisms of change in dialogic teaching. *Educational Psychologist*, *48*(2), 114-133. https://doi.org/10.1080/00461520.2013.775898

Roberts, D. A., & Bybee, R., W. . (2014). Scientific literacy, science literacy, and science education. In N. G. Ledermann & S. K. Abell (Eds.), *Handbook of*

Research on Science Education (Vol. II). Routledge. https://doi.org/10.4324/9780203097267.ch27

Robinson, S. R. (2011). Teaching logic and teaching critical thinking: Revisiting McPeck. *Higher Education Research & Development*, *30*(3), 275-287. https://doi.org/10.1080/07294360.2010.500656

Rogoff, B., & Lave, J. (1984). *Everyday cognition: Its development in social context*. Harvard University Press.

Rogosa, D. (1988). Myths about longitudinal research. In *Methodological issues in aging research.* (pp. 171-209). Springer Publishing Company.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233-239. https://doi.org/10.1037/a0017678

Rowe, M. P., Gillespie, B. M., Harris, K. R., Koether, S. D., Shannon, L.-J. Y., & Rose, L. A. (2015). Redesigning a general education science course to promote critical thinking. *CBE life sciences education*, *14*(3), ar30. https://doi.org/10.1187/cbe.15-02-0032

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching*, *41*(5), 513-536. https://doi.org/10.1002/tea.20009

Saunders, M., Lewis, P., & Thornhill, A. (2016). *Research methods for business students*. Pearson.

Sawilowsky, S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*, 597-599. https://doi.org/10.22237/jmasm/1257035100

Schoonenboom, J., & Johnson, R. B. (2017). How to construct a mixed methods research design [journal article]. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *69*(2), 107-131. https://doi.org/10.1007/s11577-017-0454-1

Schultz, P. W., & Searleman, A. (2002). Rigidity of thought and behavior: 100 years of research. *Genet Soc Gen Psychol Monogr*, *128*(2), 165-207.

Schulz, H., & Fitzpatrick, B. (2016). Teachers' understandings of vritical and higher order thinking and what this means for their teaching and assessments. *Alberta Journal of Educational Research*, *62*(1), 61-86. https://doi.org/10.11575/ajer.v62i1.56168

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction*, *16*(4), 475-522. https://doi.org/10.1207/s1532690xci1604_4

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning

and transfer. *Journal of Educational Psychology*, *103*(4), 759-775. https://doi.org/10.1037/a0025140

Scriven, M., & Paul, R. (1987). *Presentation at the 8th Annual international conference on critical thinking and education reform.* Retrieved October 7 from http://www.criticalthinking.org/pages/defining-critical-thinking/766

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Siegel, H. (1988). *Educating Reason: Rationality, Critical Thinking, and Education*. Routledge.

Siegel, H. (1995). What price inclusion? *Teachers College Record*, *97*, 6-31.

Smith, G. (2002). Are there domain–specific thinking skills? *Journal of Philosophy of Education*, *36*, 207-227. https://doi.org/10.1111/1467-9752.00270

Sternberg, R. J. (1986). *Critical Thinking: Its Nature, Measurement, and Improvement*. N. I. o. Education. https://files.eric.ed.gov/fulltext/ED272882.pdf

Sternberg, R. J. (1987). Teaching intelligence: The application of cognitive psychology to the improvement of intellectual skills. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching Thinking Skills: Theory and Practice* (pp. 182-218). Freeman.

Taube, K. T. (1997). Critical thinking ability and disposition as factors of performance on a written critical thinking test. *The Journal of General Education*, *46*(2), 129-164. www.jstor.org/stable/27797335

Terenzini, P. T., Springer, L., Pascarella, E. T., & Nora, A. (1995). Influences affecting the development of students' critical thinking skills. *Research in Higher Education*, *36*(1), 23-39. https://doi.org/10.1007/BF02207765

Thayer-Bacon, B. J. (2000). *Transforming Critical Thinking: Thinking Constructively*. Teachers College Press.

Throndsen, I., Carlsten, T. C., & Björnsson, J. K. (2019). *TALIS 2018: Første hovedfunn fra ungdomstrinnet*. https://nifu.brage.unit.no/nifu-xmlui/bitstream/handle/11250/2601320/TALIS2018_rapport_juni2019.pdf?sequence=1

Tiruneh, D. T., Weldeslassie, A. G., Kassa, A., Tefera, Z., De Cock, M., & Elen, J. (2016). Systematic design of a learning environment for domain-specific and domain-general critical thinking skills [journal article]. *Educational Technology Research and Development*, *64*(3), 481-505. https://doi.org/10.1007/s11423-015-9417-2

Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). "Mapping to know": The effects of representational guidance and reflective assessment on scientific inquiry. *Science Education*, *86*(2), 264-286. https://doi.org/10.1002/sce.10004

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293-315. https://doi.org/10.1037/0033-295X.90.4.293

Twale, D., & Sanders, C. S. (1999). Impact of non-classroom experiences on critical thinking ability. *NASPA Journal*, *36*(2), 133-146. https://doi.org/10.2202/1949-6605.1078

Utdanningsdirektoratet. (2019). *Nasjonale prøver 8. og 9. trinn - resultater*. Utdanningsdirektoratet (Norwegian Ministry of Education)

Retrieved June 9 from https://www.udir.no/tall-og-forskning/statistikk/statistikk-grunnskole/nasjonale-prover-8.-og-9.-trinn/

van Gelder, T. (2005). Teaching critical thinking: Some lessons from cognitive science. *College Teaching*, *53*(1), 41-48. https://doi.org/10.3200/CTCH.53.1.41-48

Ventura, M., Lai, E., & DiCerbo, K. (2017). *Skills for Today: What We Know about Teaching and Assessing Critical Thinking*. Pearson.

Wainer, H., & Thissen, D. (1992). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103-118. https://doi.org/10.1207/s15324818ame0602_1

Walton, D. (2008). *Informal Logic: A Pragmatic Approach*. Cambridge University Press.

Watson, G., & Glaser, E. M. (1980). *Watson-Glaser critical thinking appraisal*. The Psychological Corporation.

Wegerif, R., Mercer, N., & Major, L. (2020). Introduction to the Routledge International Handbook of Research on Dialogic Education. In N. Mercer, R. Wegerif, & L. Major (Eds.), *The Routledge international handbook of research on dialogic education* (pp. 1-8). Routledge.

Wells, G. (1999). *Dialogic inquiry: Towards a sociocultural practice and theory of education*. Cambridge University Press.

Willingham, D. T. (2008). Critical thinking: Why is it so hard to teach? *Arts Education Policy Review*, *109*(4), 21-32. https://doi.org/10.3200/AEPR.109.4.21-32

Wineburg, S., McGrew, S., Breakstone, J., & Ortega, T. (2016). *Evaluating information: The cornerstone of civic online reasoning*. Stanford Digital Repository. Available at: http://purl.stanford.edu/fv751yt5934.

Wright, D. (2006). Comparing groups in a before-after design: When t test and ANCOVA produce different results. *The British journal of educational psychology*, *76*, 663-675. https://doi.org/10.1348/000709905X52210

Yang, Y.-T. C., Newby, T., & Bill, R. (2008). Facilitating interactions through structured web-based bulletin boards: A quasi-experimental study on promoting learners' critical thinking skills. *Computers & Education*, *50*(4), 1572-1585. https://doi.org/10.1016/j.compedu.2007.04.006

Yoon, K., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. (Issues & Answers Report, REL 2007-No. 033). Washington, DC: US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest

Zamarro, G., Hitt, C., & Méndez, I. (2016). When students don't care: Reexamining international differences in achievement and non-cognitive skills. *EDRE Working Paper No. 2016-18*. https://doi.org/10.2139/ssrn.2857243

Zeidler, D. (2014). Socioscientific issues as a curriculum emphasis: Theory, research and practice. In N. G. A. Lederman, S. K. (Ed.), *Handbook of Research on Science Education, volume II* (pp. 697-726). Routledge.

Zeidler, D. L., & Nichols, B. H. (2009). Socioscientific issues: Theory and practice. *Journal of Elementary Science Education*, *21*(2), 49-58. https://doi.org/10.1007/BF03173684

Zeidler, D. L., Sadler, T. D., Applebaum, S., & Callahan, B. E. (2009). Advancing reflective judgment through Socioscientific Issues. *Journal of Research in Science Teaching*, *46*(1), 74-101. https://doi.org/10.1002/tea.20281

Zohar, A., & Tamir, P. (1993). Incorporating critical thinking into a regular high school biology curriculum. *School Science and Mathematics*, *93*(3), 136-140. https://doi.org/10.1111/j.1949-8594.1993.tb12211.x

Appendices

Appendix I: Critical thinking posters for eight grade and for ninth and tenth grade

Vi holder øye med vår egen kritiske tenkning!

- Greier vi å unngå at vår kritiske tenkning blir påvirket av et bestemt svar vi håper på?
- Har vi holdt oss til fakta?
- Har vi gjort vårt beste for å vurdere alle sider av saken, også de sidene vi ikke liker eller er uenige i?

ARGUMENT BERGEN KOMMUNE Høgskulen på Vestlandet UNIVERSITETET I BERGEN

Appendix II: Consensus conception of critical thinking

The American Philosophical Association's consensus conception of critical thinking

Cognitive skills and subskills	Affective dispositions
Interpretation Categorization Decoding Significance Clarifying Meaning **Analysis** Examining Ideas Identifying Arguments Analyzing Arguments **Evaluation** Assessing Claims Assessing Arguments **Inference** Querying Evidence Conjecturing Alternatives Drawing Conclusions **Explanation** Stating Results Justifying Procedures Presenting Arguments **Self-regulation** Self-examination Self-correction	**Approaches to life and living in general** • Inquisitiveness with regard to a wide range of issues • Concern to become and remain generally well-informed • Alertness to opportunities to use CT • Trust in the processes of reasoned inquiry • Self-confidence in one's own ability to reason • Open-mindedness regarding divergent world views • Flexibility in considering alternatives and opinions • Understanding of the opinions of other people • Fair-mindedness in appraising reasoning • Honesty in facing one's own biases, prejudices, stereotypes, egocentric or sociocentric tendencies • Prudence in suspending, making or altering judgments • Willingness to reconsider and revise views where honest reflection suggests that change is warranted **Approaches to specific issues, questions, or problems** • Clarity in stating the question or concern • Orderliness in working with complexity • Diligence in seeking relevant information • Reasonableness in selecting and applying criteria • Care in focusing attention on the concern at hand • Persistence though difficulties are encountered • Precision to the degree permitted by the subject and the circumstance

Appendix III: Juxtaposition of critical thinking skills and science practices

The American Philosophical Association's consensus of critical thinking skills	**K-12 Framework Science and Engineering Practices**
Interpretation Categorization Decoding Significance Clarifying Meaning **Analysis** Examining Ideas Identifying Arguments Analyzing Arguments **Evaluation** Assessing Claims Assessing Arguments **Inference** Querying Evidence Conjecturing Alternatives Drawing Conclusions **Explanation** Stating Results Justifying Procedures Presenting Arguments **Self-regulation** Self-examination Self-correction	Asking questions (for science) and defining problems (for engineering) Developing and using models Planning and carrying out investigations Analyzing and interpreting data Using mathematics and computational thinking Constructing explanations (for science) and designing solutions (for engineering) Engaging in argument from evidence Obtaining, evaluating, and communicating information

Appendix IV: Juxtaposition of critical thinking dispositions and aims and values of science

The American Philosophical Association's consensus of critical thinking dispositions	**Aims and Values of Science**
Approaches to life and living in general • Inquisitiveness with regard to a wide range of issues • Concern to become and remain generally well-informed • Alertness to opportunities to use CT • Trust in the processes of reasoned inquiry • Self-confidence in one's own ability to reason • Open-mindedness regarding divergent world views • Flexibility in considering alternatives and opinions • Understanding of the opinions of other people • Fair-mindedness in appraising reasoning • Honesty in facing one's own biases, prejudices, stereotypes, egocentric or sociocentric tendencies • Prudence in suspending, making or altering judgments • Willingness to reconsider and revise views where honest reflection suggests that change is warranted **Approaches to specific issues, questions, or problems** • Clarity in stating the question or concern • Orderliness in working with complexity • Diligence in seeking relevant information • Reasonableness in selecting and applying criteria • Care in focusing attention on the concern at hand • Persistence though difficulties are encountered • Precision to the degree permitted by the subject and the circumstance	• Seeking neutrality and avoiding bias • Searching for new explanations • Ensuring that explanations are accurate • Basing claims on sufficient, relevant and plausible data • Giving reasons to justify claims • Recognizing opposite ideas and responding to objections • Taking opposition to own ideas seriously • Changing own ideas in light of evidence • Considering and respecting human needs • Making sure nobody controls ideas to favor particular group biases • Being honest and acting honestly in all aspects of scientific activities • Respecting ideas as long as they are evidence based irrespective of whose ideas they are

Appendix V: Analogous examples of items from the Cornell Critical Thinking Test Level X

Part of test (category in part II)	**Analogous examples of items**	**Solution with explanation**
Part I: Induction	You go into the first hut. Everything is covered by a thick layer of dust. A. This fact **supports** the medical expert's idea that everyone in the first group is dead. B. This fact **goes against** the health officer's idea. C. **Neither**: this fact does not help us decide.	A. The dust is well explained by the hypothesis that everyone in the first group is dead.
Part II: Observation and credibility (Observation versus inference)	(Which underlined statement is more believable?) A. A worker says, "Five times now the person in the blue jacket has talked to someone and pointed, and immediately they have run off in the direction he pointed." B. Another worker says, "He must be the leader." C. A and B are equally believable.	A. The first worker is doing less, if any, inferring. The other infers leadership.
Part II: Observation and credibility (Conflict of interest)	(Which underlined statement is more believable?) If the man on the left is Captain James, the reward goes to the geologist. If not, it goes to the pilot. A. The medical expert looks through his field glasses at the one on the left. "That is not Captain James," he says. B. The geologist looks through his field glasses and replies, "Yes, it is." C. A and B are equally believable.	A. The geologist has a conflict of interest.
Part II: Observation and credibility (Method of observation)	(Which underlined statement is more believable?) A. One worker counts the people as they move around the town square. He reports, "Only 42 people came back from the river." B. Another worker says, "You must have missed two. I counted them as they walked past the big rock, and 44 came back. I don't believe any of them came back another way." C. A and B are equally believable.	B. The second worker specified a better method for counting.
Part II: Observation and credibility (Authority/expertise)	(Which underlined statement is more believable?) A. The construction worker investigates the stream by the village and reports, "The water is not safe to drink." B. The medical expert says, "We can't tell yet if the water is safe to drink." C. A and B are equally believable.	B. The medical expert should know more about whether the water is safe to drink than the mechanic.

	(Which underlined statement is more believable?)	
Part II: Observation and credibility (Other)	A. The worker, looking through his binoculars, says, "I think there are 40 of them." B. The geologist, looking through his binoculars, says, "No, I think there are only 37." C. A and B are equally believable.	C. There is no basis for preferring one over the other.
Part III: Deduction	The worker says, "It these beings are people from Earth, then they will welcome us. Certainly they are people from Earth." Which follows? A. These beings will not welcome us. B. These beings are not from Earth. C. These beings will welcome us.	C. If what the worker says is true, then C must also be true. (Affirmation of the antecedent.)
Part IV: Assumptions	"The explorers can't escape, because they can't tear down the walls of the stone hut." Which one of the following is probably taken for granted? A. The explorers can jump out the window. B. The guards are alert. C. All ways of escape, except through the walls, are impossible.	C. Among the choices, C helps the reasoning the most.

Appendix VI: Information letter to students and parents

TIL ELEVER OG FORESATTE

ARGUMENT (Allmenndannende Realfag Gjennom Utforsking Med Ekte og Nære Tall)

Allmenndannende Realfag
Gjennom Utforsking
Med Ekte og Nære Tall
= ARGUMENT

Bergen kommune har fra i fjor ledet forskningsarbeidet ARGUMENT i samarbeid med forskere fra Høyskolen på Vestlandet (HVL) og Universitetet i Bergen (UiB). Hovedmålet vårt er å øke elevers motivasjon og læring i matematikk og naturfag, samt deres evne til faglig utforskning og kritisk tenkning.

Vi er interesserte i å finne ut hvordan elever lærer når de arbeider med tall, og bruker argumentasjon og kritisk tenkning i forbindelse med samfunnsaktuelle problemstillinger. Elevene arbeider med egne målinger og større datamengder, for eksempel fra bergensveret.no. ARGUMENT gjennomfører klasseromsforskning der lærere og forskere arbeider tett sammen. Resultatene skal bli satt inn i Bergen kommunes systematiske program for skoleutvikling og bli tilgjengelig for alle via oppgaver som utvikles på ektedata.uib.no.

Din skole spiller en viktig rolle
Flere skoler i Bergen kommune er med på utviklingen av ARGUMENT. Noen klasser, elever, og lærere blir fulgt opp tettere, og flere har allerede fått informasjon om hva det innebærer og blitt invitert til å delta i forskjellig grad. Nå får alle elever ved din skole muligheten til å delta i forskningen i prosjektet. Derfor går dette skrivet ut til alle elever og foresatte ved de aktuelle skolene. Deltagelsen gir prosjektgruppen bedre informasjon om elevers kritiske tenkning og effekten av undervisning i kritisk tenkning. I tillegg får vi muligheten til å sammenligne på tvers av skoler. Dette er nyttig for utvikling og forbedring av undervisning i kritisk tenkning, og for å få bedre forskningsresultater. Vårt mål og vår plan er at dette skal gagne alle ungdomsskolene i Bergen på sikt.

Hva innebærer det å delta?
Vi ønsker å gjennomføre en test av kritisk tenkning i starten av høstsemesteret 2019, og en i slutten av vårsemesteret 2020. Lærere i matematikk og naturfag vil bli bedt om å svare på et spørreskjema omtrent samtidig som testene. Testen av kritisk tenkning, inkludert instruksjon og å svare på et kort spørreskjema, tar omtrent en time å gjennomføre. Testen er internasjonalt anerkjent som en av de beste og mest brukte testene av kritisk tenkning for ungdomsskoleelever. Den er lagt opp som en historie der man underveis skal ta stilling til påstander og gjøre egne vurderinger. Spørreskjemaene til lærere tar 10-15 minutter å svare på. Vi håper at så mange som mulig av elever og lærere i matematikk og naturfag deltar.

Tillatelse til datainnsamling for forskningsbruk
Forskerne har kontaktet Norsk Senter for forskningsdata (NSD) og Universitetet i Bergen sitt personvernombud for å sørge for at prosjektet følger retningslinjene for forskning og personvern. Den endelige vurderingen er at våre datainnsamlingsplaner er innenfor retningslinjene. Siden vi ikke samler inn navn eller direkte personidentifiserende informasjon fra elever, trenger vi ikke samtykke fra foresatte. Alle data blir selvsagt

behandlet og oppbevart trygt. I god forskningsetisk ånd ønsker vi likevel å informere både foresatte, elever, og lærere med dette skrivet.

Hvordan finne ut mer?

Hvis du har spørsmål til studien kan du ta kontakt med:

- Praktisk ansvarlig for forskningen: Vegard Havre Paulsen, stipendiat ved Institutt for fysikk og teknologi, Universitetet i Bergen, telefon 99 35 35 81, e-post Vegard.Paulsen@uib.no
- Forskningsleder: Stein Dankert Kolstø, professor ved Institutt for fysikk og teknologi, Universitetet i Bergen, telefon 55 58 48 39 eller 92 64 21 36, e-post Stein.Dankert.Kolstoe@uib.no
- Prosjektleder: Janneke Tangen, Rådgiver i Etat for skole, Bergen kommune, telefon 55 56 24 78, e-post Janneke.Tangen@bergen.kommune.no
- Universitetet i Bergen sitt personvernombud: Janecke Helene Veim, telefon 55 58 20 29, e-post Janecke.Veim@uib.no
- NSD (Norsk Senter for Forskningsdata AS), telefon 55 58 21 17, e-post personverntjenester@nsd.no

Vennlig hilsen, på vegne av ARGUMENT-gjengen

Janneke Tangen

Rådgiver Etat for skole - Leder ARGUMENT

PART TWO | ARTICLES

ARTICLE I

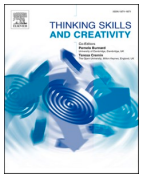
ARTICLE II

Thinking Skills and Creativity 43 (2022) 100969

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: www.elsevier.com/locate/tsc



Students' reasoning when faced with test items of challenging aspects of critical thinking

Vegard Havre Paulsen *, Stein Dankert Kolstø

Department of Physics and Technology, University of Bergen, Norway

ARTICLE INFO

Keywords:
Critical thinking
Reasoning
Test
Assessment
Education

ABSTRACT

The growing focus on education for critical thinking has increased the need to determine which specific aspects of critical thinking students find challenging. Insight into how students reason when facing problems related to these aspects could improve instruction. While there are several assessments of critical thinking, accounts of reasoning related to these assessments are lacking. Here we present a thematic analysis of 284 secondary students' written justifications when facing selected challenging items from a modified version of the Cornell Critical Thinking Test Level X. We identified 21 lower-level themes encompassed by six higher-level themes of reasoning. More than a quarter of the responses (28%) expressed strong inductive logic yet were incorrect reasons because the premises used were based on alternative evidence (14%), or because the premises used were made up by students who changed elements of the context (14%). Few responses (3%) did not seem to represent strong inductive logic, most of them being that the students seemed to believe an inference to be just as, or more, believable than an observation. We discuss potential barriers that students faced in this study, and particularly how these barriers relate to skills, dispositions, and knowledge.

1. Introduction

It is impossible for any individual to keep up with all new knowledge and information. However, it is possible to improve one's ability to determine what is relevant information, evaluate its credibility, and draw reasonable conclusions about what one should believe or do. This is part of what constitutes critical thinking. Thus, it seems reasonable that fostering critical thinking skills that students can apply independent of subject matter has been proposed as an important goal for education in the 21st century (Dam & Volman, 2004; NGSS Lead States, 2013). Critical thinking test results have been shown to correlate with fewer negative life outcomes across a wide range of domains, including health, education, finance, and interpersonal relationships (Butler et al., 2012). Furthermore, critical thinking might be a better predictor of such outcomes than intelligence (Butler, Pentoney & Bong, 2017). Critical thinking is also associated with emotional intelligence (Li et al., 2021). Educators, policymakers, and governments focus on the importance of educating for critical thinking (OECD, 2015; Ventura, Lai & DiCerbo, 2017), yet there is concern that educational efforts have not sufficiently improved critical thinking outcomes (Bouygues, 2018; Case & Wright, 1997), and there are indications that adolescents struggle with evaluating evidence and source credibility (Ku et al., 2019). Nevertheless, there is substantial evidence supporting that critical thinking can be improved through education across many educational levels and disciplines, and instruction

; CT, Critical thinking; CCTT, Cornell Critical Thinking Test.

* Corresponding author.
E-mail address: vegard.paulsen@uib.no (V.H. Paulsen).

https://doi.org/10.1016/j.tsc.2021.100969
Received 25 March 2021; Received in revised form 22 September 2021; Accepted 25 October 2021
Available online 10 November 2021

with explicit focus on critical thinking seems to be most effective (for meta-analyses, see Abrami et al., 2015, 2008). Thus, to make such instruction more focused and effective, it might be useful to have insight into which particular aspects of critical thinking students struggle with, as well as how they reason when faced with these aspects.

Through a quantitatively driven exploration of some of the lowest scoring items on the Cornell Critical Thinking Test (CCTT) Level X and secondary school students' written justifications of their answers to those items, the authors of the present study found indications, presented in Paulsen and Kolstø (2021), that students struggle with certain categories of critical thinking test items. The students struggled with evaluating the credibility of statements where there was a need to discern between an observation and an inference, a need to recognize a conflict of interest, and, to a lesser degree, a need to discern between methods of observation. In the present study, we further investigated students' responses to these items by qualitatively analyzing parts of the same dataset in order to answer one main research question: How do students reason when faced with test items from these challenging critical thinking categories? In light of the results, we will discuss possible barriers that students face in order to successfully solve the items from these categories, and particularly how these barriers relate to skills, dispositions, and knowledge.

1.1. What is critical thinking?

Researchers have offered a wide range of definitions of critical thinking (for reviews, see Abrami et al., 2015, 2008). Robert Ennis, a thought-leader in the field of critical thinking, have been proposing for more than 30 years that critical thinking is "reasonable reflective thinking focused on deciding what to believe or do" (1987, p. 10). Alongside this definition Ennis has put forward a list of cognitive abilities and affective dispositions that constitute critical thinking (further elaborated in Ennis, 1996a). This view of critical thinking as a synthesis of abilities and dispositions seems to be the consensus among researchers (Ku, 2009). Moreover, Ennis (2016) states that many well-known definitions of critical thinking are not significantly different from his own or each other, and that all of these fall within a "mainstream concept of critical thinking" (p. 166). He goes on to state that all these definitions could be used to create a list of skills and dispositions, like his own list. One seminal definition that has done just that comes from a Delphi consensus panel of experts, including Ennis (Facione, 1990). The panel posits that critical thinking is a form of purposeful and self-regulatory judgment requiring both cognitive skills and dispositions. The main skills listed by the panel are interpretation, analysis, evaluation, inference, and explanation, as well as self-regulation, as a metacognitive component. More specifically, the skills are often related to evaluation and analysis of arguments with the purpose of judging their credibility. According to the panel's consensus definition, the dispositions that are part of critical thinking consist of personality traits, mental habits, and virtues, such as inquisitiveness, fairness in evaluation, open-mindedness, and striving to be well informed.

Bailin, Case, Coombs and Daniels (1999) conceptualize critical thinking as a set of *intellectual resources*, as they find skills to be an inaccurate and confusing term. These intellectual resources include knowledge of certain standards of good thinking (e.g., considering alternatives), knowledge of key critical concepts (e.g., necessary and sufficient conditions), and heuristics (e.g., double-checking before concluding). Dispositions, or habits of mind, are also part of this conceptualization. Although the term skills is not included, this conceptualization seems closely related to the two-factor model of skills and dispositions, and Ennis' vision of the "mainstream concept of critical thinking" (Ennis, 2016, p. 166). The distinction of skills and knowledge is not a discussion suited for this article. However, we find Bailin et al. (1999) focus on knowledge of generalizable key critical concepts potentially relevant for this study.

For the purposes of this study, we deem the "mainstream concept of critical thinking" (Ennis, 2016, p. 166) as elaborated by Facione (1990) and Ennis (e.g., 1996a) a suitable framework, with added elements from Bailin et al. (1999). One important reason for this is that the instrument used to measure critical thinking skills in this study, the CCTT Level X, is based on Ennis' conception of critical thinking (Ennis & Millman, 2005). The test is designed to minimize the need for domain-specific background knowledge for that not to be a limiting factor for test takers. For example, there should be no need to know anything about math, science, or history to complete the test, and all contextual information is provided in the test itself. Nevertheless, both Ennis and Facione emphasize the importance of background knowledge, in addition to skills and dispositions, as part of good critical thinking. The question about the need for domain-specific background knowledge is closely tied to the question of whether critical thinking is generalizable and applicable across a range of domains or if it is domain-specific. Ennis (1989) states that most researchers view critical thinking as at least somewhat generalizable, and that they view domain-specific background knowledge as necessary, but not sufficient. Several publications indicate that this could still be a prevailing view in this century, and that the specificity-generality debate has shifted towards a synthesis of both sides of the spectrum (Davies, 2013; Robinson, 2011; Smith, 2002). In other words, the choice between generalist and specifist approaches towards teaching and assessing critical thinking is not black or white. Davies (2013, p. 536) argues for a "bottom-up view" of critical thinking, where the generic sense of critical thinking undergirds, and explains "without residue", any supervenient domain-specific sense of critical thinking. Robinson (2011) argues that general critical thinking courses are of value to students and throws doubt on the hardcore specifist position forwarded by McPeck (1981). Finally, Smith (2002) characterizes domain-specific thinking skills, yet maintains that "there are far more elements of generality in our thinking practices than have heretofore been recognized" (p. 224).

1.1.1. Learning critical thinking in school and applying it outside of school

The discussion of generality versus specificity is related to the concept of *transfer,* which addresses the question of whether something that is learned in one setting, such as in school, can be successfully applied in a different setting, such as a real-world context like when deciding what profession to choose. This is arguably the ultimate purpose of critical thinking instruction and education in general (Abrami et al., 2008; National Research Council, 2000). It might be more useful to think of transfer as "flexible adaptation to new problems and settings" (National Research Council, 2000, p. 77). Such *flexible adaption* seems to be supported by deeper initial

learning, metacognitive monitoring, and learning something in multiple contexts. With regards to critical thinking, there is abundant evidence of it being successfully taught in education, using a wide range of conceptualizations and measures, and across a wide range of educational settings, and for all age-levels (Abrami et al., 2015, 2008; Halpern, 2001; Lai, 2011). Moreover, there are indications of *near-transfer* of critical thinking skills within educational settings. For example, Tiruneh et al. (2016) showed that performance on a domain-specific critical thinking test (i.e., within electricity and magnetism) significantly predicted performance on a domain-general critical thinking test. Furthermore, Halpern (2001) points to numerous studies indicating that better thinking can be learned in one context and applied in other contexts, including a study where students spontaneously transferred reasoning skills from an academic context to a real-world context taking place several months after the academic instruction. However, while other studies have indicated that the evidence for transfer of critical thinking is mixed and less conclusive, many researchers seem to believe that transfer is possible even though some are more skeptical of the possibility than others (for a review, see Lai, 2011). Moreover, researchers from both sides in the generality-specificity debate maintain that transfer of critical thinking from educational to real-world contexts is probably possible, including hardcore specifists like McPeck, 1981

1.2. Assessment of critical thinking

Existing critical thinking assessments range from multiple-choice, to essay-only, to a combination of multiple choice and constructed response (for a list of tests, see Liu, Frankel & Roohr, 2014). A combination is often recommended because multiple-choice tests are easier to score and essay tests provide a more detailed assessment (e.g., Ennis, 1993; Ku, 2009). This study uses a modified version of the CCTT Level X that includes written justifications to chosen items, an approach recommended by Ennis (1993). While most tests are designed for adults and college students, the CCTT Level X is aimed at grades 4 through 14 (Ennis, Millman & Tomko, 2005), making it a suitable choice for this study, with participants from lower secondary school. The CCTT is designed to test general critical thinking skills in a context that is independent from subject-specific domains, as are most other assessments (Liu et al., 2014). Thus, if critical thinking skills are learned in a specific subject-domain (e.g., math) and then applied in a domain-independent test of general critical thinking skills like the CCTT, an increase in scores (e.g., from pretest to posttest) could indicate that some sort of transfer of these skills has occurred.

A great deal of research has been done to establish the validity and reliability of commonly used critical thinking tests, yet the results are inconsistent (Abrami et al., 2008). However, after reviewing the available research, Ennis et al. (2005) holds that there is justifiable support for the construct validity of the CCTT Level X as a test of critical thinking ability. The test makers base their tentative conclusion on at least eleven types of information that could be pertinent to evaluate construct validity, including the rationale behind the test, the degree to which the test seems to fit that rationale, statistical analyses, correlations between the test and other tests that are intended to measure approximately similar outcomes, reliability, correlations between the test and other variables, factor analyses, and results of experiments where the test was used to gage critical thinking ability. Nevertheless, a more recent factor analysis by Leach, Immekus, French and Hand (2020) maintains that more clarity is needed concerning the concept validity of the test, especially relating to the subscale-scores for the different aspects that the test incorporates. However, it can be argued that identifying a neat factor-structure might not be feasible due to the heterogeneity and overlap of the aspects and thus also the items. Ennis et al. (2005) do not claim that the test is a be-all and end-all test directly assessing all aspects of critical thinking, but they hold that the test does seem to adequately measure critical thinking skills, particularly related to induction, deduction, and evaluation of credibility (e.g., of claims based on others' observations). Moreover, the members of the Illinois Critical Thinking Project reached agreement concerning that these items do call for the indicated aspects of critical thinking skills. It is important to note that the test makers strived to write items that call for the employment of these skills in contexts that most people will understand (i.e., contexts independent from specific subject-domains), and to write each item such that test takers should generally possess "all the other things required to get it right" (Ennis et al., 2005, p. 22). Furthermore, they intensely discussed all individual items and reached universal agreement on the correctness of the keyed answers. However, the members acknowledged that there could be other correct answers to some items, particularly if test takers had different auxiliary assumptions (than, for example, the assumption that a layer of dust will probably develop in a hut that is not lived in). The recommended stance to account for this is to deem any score on Part I and Part II with greater than 85% agreement with the key to be indicative of mastery. Finally, our addition of written justifications to the test poses a potential risk to its psychometric qualities, like any other modification would. However, one study found that including verbal reports in a multiple-choice test of critical thinking did not alter the test performance, indicating that test validity and reliability might be preserved in spite of our modifications (Norris, 1990).

We are also aware that there are objections to the validity of skill-based tests in general. Critics maintain that results from such tests only show performance on the test, and that these tests perhaps only measure students' disposition to do what they are told, or, at best, a few skills that are only a very limited part of any conceptualization of critical thinking that these critics might subscribe to. However, these conceptualizations of critical thinking might differ from those used as rationale for most tests. For example, Barbara Thayer--Bacon (2000) argues for a transformation of the skills-and-dispositions conceptualization of critical thinking, which she sees as insufficient, into what she calls constructive thinking, where she includes intuition, emotion, and imagination. She also draws on feminist and post-modern theory and sees critical thinking as relating to power and oppression. This highlights the need for research on how to improve and broaden conceptualizations of critical thinking, as well as assessments of different aspects of critical thinking. Nevertheless, as we have argued in Section 1.1, there seems to be merit to the idea that certain general critical thinking skills are able to capture at least a part of the complexity of what critical thinking is, and as Abrami et al. (2015) argues, to reject these kinds of data completely "would be tantamount to discarding the good for the sake of a nonexistent and unobtainable better" (p. 304). As mentioned, the current study uses the CCTT Level X (and its theoretical rationale) and has modified it to include written justifications

to selected items. Thus, the findings of the study can contribute to insights into how students reason when working with these items, which might have implications for critical thinking instruction, current beliefs about test validity, interpretation of test results, and development of other tests. These contributions are discussed in Section 4.3.

2. Method

The study presented here gathered data from lower-secondary students ($N = 284$) from six schools in the second largest city in Norway, Bergen. These schools were part of a teacher development project that the researchers are a part of, and we used the opportunity to conduct this study (which was not directly related to the project's main aims) on students in some classrooms in these schools. Students in these schools had fairly similar scores on nationwide tests (Utdanningsdirektoratet, 2019). Moreover, results from the 2015 PISA-test indicate that differences between Norwegian schools are small in general, and that factors like socioeconomic background have less impact on student scores than in other countries (Kjærnsli & Jensen, 2016). In this section, we will present the instruments that were used, and the process of collecting and analyzing the data. The project was approved by the Norwegian centre for Research Data. No person-identifying information was collected. Parents and participants were informed about the study, and participation was voluntary.

2.1. Instruments and data collection

The CCTT Level X is a 71-item multiple-choice test presented as a story about people from Earth travelling to and exploring a new planet (Ennis & Millman, 2005). The crew consists of people from different professions, and test takers get to join the action and must make decisions based on the available information. Each item has three alternatives and one keyed answer. The project team had the test translated to Norwegian by a professional translator, and then put it into an online survey solution for ease of administration and analysis. We had also administered a pilot test in another school in the area and adjusted the language, and the instructions for test administrators. The online solution offered the possibility of full anonymization.

Paulsen and Kolstø (2021) identified the lowest scoring items from the initial critical thinking test in the project ($N = 1353$) as well as which part of the test had the most low-scoring items. Most of the low scoring items were from Part II, which according to the test manual covers *observation* and *credibility* (Ennis et al., 2005). All the items in Part II require test takers to evaluate the credibility of observation reports. We analyzed the manual's proposed explanations of the keyed answers to the items of Part II, and subsequently divided these items into five categories: Observation versus inference (items 28, 43, 44, and 46), difference in method of observation (items 41, 42, 45, 47, 48, and 50), conflict of interest (items 36 and 38), difference in authority/expertise (items 31, 32, and 39), and other (items 29, 30, 33, 34, 35, 37, 40 and 49). The first three categories contained items that had among the lowest scores (Paulsen & Kolstø, 2021). Thus, we decided to modify the test by adding prompts to justify the answers to chosen items from these first three categories. In order to keep intact the validity of the test, the nature of the items cannot be disclosed. Thus, we have changed surface features of test items and the examples presented in Table 1 are analogous to the actual test items.

Written justifications, we thought, would give insight into student reasoning, particularly relating to wrong answers. Furthermore, some students could have valid reasons for choosing a different answer than a keyed answer, which would not be given credit for in a multiple-choice-only test. A pilot test with four different versions of the modified test was administered to another school in the area. The versions had either six or ten prompted items, and the prompts were placed either directly after each item or at the end of the test.

Table 1
Analogous examples of items from the explored categories of items from the Cornell Critical Thinking Test Level X.

Category	Item #	Analogous examples of items(which underlined statement is more believable?)	Solution with explanation
Observation versus inference	28 43 44 46	A A worker says, "Five times now the person in the blue jacket has talked to someone and pointed, and immediately they have run off in the direction he pointed." B Another worker says, "He must be the leader." C A and B are equally believable.	A. The first worker is doing less, if any, inferring. The other infers leadership.
Conflict of interest	36 38	If the man on the left is Captain James, the reward goes to the geologist. If not, it goes to the pilot. A The medical expert looks through his field glasses at the one on the left. "That is not Captain James," he says. B The geologist looks through his field glasses and replies, "Yes, it is." C A and B are equally believable.	A. The geologist has a conflict of interest.
Method of observation	41 42 45 47* 48* 50*	A One worker counts the people as they move around the town square. He reports, "Only 42 people came back from the river." B Another worker says, "You must have missed two. I counted them as they walked past the big rock, and 44 came back. I don't believe any of them came back another way." C A and B are equally believable.	B. The second worker specified a better method for counting.

* No prompt to write response on item.

The students seemed to handle ten prompted items well (i.e., most students responded to all ten), and the versions with prompts directly after each item seemed to give satisfactory written responses. The two last parts of the test were removed so that only Part I and Part II remained in the modified test. We were interested in Part II, and Part I was needed to give students the full story leading up to Part II. This allowed the total testing time to remain around one hour. The modified test consisted of 47 items, ten of which included prompts for written justifications. Table 1 lists the nine selected items (item 37 was prompted just to keep the flow from item 36 to 38). Focusing on a test for one hour was more than enough for many students and keeping the administration time low was more practical for teachers. The test administrators were instructed to make sure that students delivered the test in the allotted time, 50 min for the test, and around 10 min for a survey beforehand.

At the end of the spring semester of 2020, the modified test was administered to classes in the six schools that for some reason had not taken the unmodified version of the CCTT Level X that related to another data collection in the project. A total of 284 students from all six schools completed the modified test, of which 170 were 8th graders, 94 were 9th graders, and 20 were 10th graders.

2.2. Data analysis

To gain insight into the reasoning students had for choosing a given answer we conducted a thematic analysis by following the guidelines proposed by Braun and Clarke (2006). NVivo 12 Pro-was used for the analysis. Each response to an item was defined as a segment for analysis. Thus, for 284 students and nine written-justification items, there were a total of 2556 segments. The length of the written responses typically ranged from one word to a couple of sentences. After reading through the dataset several times, the first author generated a set of semantic codes based on what the students wrote in their response to each item. This initial coding was done until no new codes seemed to be needed. The authors and three other educational researchers discussed this coding process, and overlapping codes were merged. The resulting semantic codes are those called *lower-level themes* in Table 2, and Section 3.1 presents examples of direct quotes from students for each of these semantic lower-level themes. After landing the lower-level themes, the search for more latent themes (i.e., the *higher-level themes*, seen in Table 2) started, using a data-driven, inductive approach. Concept maps were developed by piling the lower-level themes (i.e., semantic codes) that seemed to have something in common together in groups that could possibly form overarching themes. For example, responses coded as *the person's mode of expression; better method or procedure of observation; majority agreement, repetition, order;* and *the person's experience, education, or knowledge* all seemed to focus on using evidence that was unrelated to the test manual's correct solution. Thus, these were grouped together in the higher-level theme called *using alternative evidence,* as seen in Table 2. The continuous discussions between the authors and the three other educational researchers led to minor revisions of codes and themes throughout the analysis. Finally, we landed on the two levels of themes presented in Table 2. The 21 lower-level themes (or semantic codes) were closest to the dataset, and these grouped together into the six overarching, or higher-level, themes. Each response could be coded as several overlapping higher-level themes, and the responses within each higher-level theme could be coded as several lower-lever themes. Ultimately, the themes emerged from the data as viewed

Table 2
Overview of themes from the thematic analysis of written justifications to selected items of the Cornell Critical Thinking Test Level X.

Higher-level themes	Lower-level themes
1 Changing context	a Misunderstanding context
	b Giving an answer that does not address what was asked
	c Adding extra information
	d Using their own experience, common conceptions, and generalizations
2 Using alternative evidence	a The person's experience, education, or knowledge
	b Majority, agreement, repetition, order
(alternative compared to correct solution)	c Better method or procedure of observation
	d The person's mode of expression
3 Lack of basis to choose A or B	a Observation statement and inference statement perceived as similar
	b Lack of information or evidence, or available evidence is similar for both statements
	c Alternative C because it is impossible to be 100% certain
4 Logical fallacy	a Inference more or as probable as an observation
	b Other logical fallacies
5 Lacking reason	a Because / no reason
	b Can't be bothered
	c Irrelevant (words, sentences, or drawings not answering the item)
	d Nonsense / gibberish (words or signs with no apparent meaning)
	e Guessing
	f Don't know (or not selecting an answer)
6 Right reason (see the online supplementary material for elaboration on what entails right reason for each category)	a Right reason, wrong answer
	b Right reason, right answer

through the lens of the research questions, the theoretical framework of critical thinking, and the conceptual basis of the CCTT. Our aim when conducting the analysis was to find the essence of what students expressed in their reasons and how these reasons led to their chosen answers. We assumed that there were certain similarities in the ways the students expressed their reasoning and that these were possible to observe in the data. At the same time, we acknowledged that there was an inevitable subjectivity in the analysis, and thus decided to test interrater reliability.

After landing the final set of themes, the first author coded a portion of the dataset again. Another coder, MK, after a test round of coding, coded 252 randomly chosen responses (i.e., around 10% of the dataset). Because the themes were mutually non-exclusive (i.e., there were several possibilities of theme-combinations for each answer) using Cohen's (1960) original Kappa formula might have been problematic. Thus, we calculated an interrater reliability score based on *proportional overlap,* a modified version of Cohen's Kappa, proposed by Mezzich, Kraemer, Worthington and Coffman (1981). The observed proportional overlapping agreement (*Po*) and expected agreement by chance (*Pe*) were $Po = 0.81$ and $Pe = 0.28$ for the six higher-level themes, and $Po = 0.69$ and $Pe = 0.09$ for the 21 lower-level themes. The resulting Mezzich's Kappa values are $K = 0.74$ (SE = 0.13) for the higher-level themes and $K = 0.65$ (SE = 0.04) for the lower-level themes, which were deemed as good. (See the supplementary material for elaboration of this process). The first author then proceeded to code the rest of the dataset.

Throughout the process of analyzing the data it became clear that there were students that did not give serious responses to the items. Thus, after the thematic analysis, we removed (from the dataset) all students that had three or more responses coded as 5b, 5c, or 5d, as this seemed to effectively target students that obviously did not make an effort while leaving in those who wrote serious responses. A total of 44 students were removed, leaving 240 students and a total of 2160 responses for the outcome presented in the results section.

3. Results

To answer the question of how students reason when faced with the items from categories representing challenging aspects of critical thinking, we here present the lower-level and higher-level themes (see Section 2.2 and Table 2) that emerged from the thematic analysis, including examples of students' reasoning. We found 21 lower-level themes encompassed by six higher-level themes. These results also suggest what barriers students might be most likely to face when trying to answer items from these categories of challenging aspects of critical thinking, and these potential barriers will be covered in the discussion.

3.1. Explanation of themes with examples of students' reasoning

The hierarchy of themes can be seen in Table 2, and we present examples of coding and student reasoning in this section. Analogous examples of the items in each of the explored categories (observation vs inference, conflict of interest, and method of observation) can be seen in Table 1. To match the changed surface features of the items, we have also changed the words in the presented examples of student reasoning that refer to these surface features (i.e., the words in brackets).Note that the sum of responses from all lower-level themes might be larger than the number of responses that are reported for the overarching higher-level theme, like is the case for Theme 1 and Theme 2. This is because each response could be coded as several lower-level themes within the same higher-level theme. Each response could also be coded as several higher-level themes. Also note that we use the terms "correct" and "incorrect" as they relate to the solutions proposed in the test manual, and not as any other judgment of the expressed reasoning (e.g., personal opinions about the logic and rationale of the responses).

Theme 1 – Changing context: Four lower-level themes were used for responses that seemed to miss given information, give an answer that did not address what was asked, add extra information, or were based on personal experience, common conceptions, or generalizations. These lower-level themes seemed to group together into one higher-level theme. This theme, *changing context*, encompassed 360 responses.

Of these, 122 responses (34%) were coded as 1a (*misunderstanding context*). Two items required students to recognize a conflict of interest and use this as a basis for deciding which person or statement was more believable. One of these items (seen in Table 1) had two opposing statements by two different people, where only one of them would get a large reward if their statement turned out to be correct. A student who misunderstood the context, and answered that both statements were just as believable, wrote: "I answer this because both can lie to get the reward."

Furthermore, 79 responses (22%) were coded as 1b (*giving an answer that does not address what was asked*). Most of these (68 responses / 86%) were related to one of the items (i.e., item 28) requiring students to discern between an observation and an inference and decide which of the two statements was more believable. A soldier made the statement, "The water looks clear." And a medical expert, after testing a sample, stated, "The water is safe to drink." The correct answer is that the soldier's statement is more believable because it is an observation about the appearance of the water, while the medical expert infers that the water is safe to drink. This item seems to be particularly difficult, something which we attend to in the discussion. The following example shows a student who touched upon the correct reason, but ultimately her answer failed to address which statement was more believable and instead she focused on whether the water was safe to drink or not. Thus, she answered that the medical expert's statement was more believable "[*b*]ecause the [soldier] only says what he sees, not that it is safe. The [medical expert] says it is safe." Another student reasoned, "Because a [medical expert] knows a lot more about what is dangerous and not, than the others."

A total of 108 responses (30%) were coded as 1c (*adding extra information*). One student, on the item described in the previous paragraph, reasoned that the soldier was more believable "because I think that [medical expert] did not want them to live or something so that he finally could be correct for once." This is an example of a student who marked the correct alternative, but for a different

reason than what is proposed by the test makers. The implications of such instances are addressed briefly in the discussion.

Finally, 64 responses (18%) were coded as 1d (*using their own experience, common conceptions, and generalizations*). There are a few examples of students' reasons on items where two people are counting and reporting different numbers. On these items, the correct reason is that one of them has a better procedure for counting, for example by counting people as they walk by a landmark along a path versus counting a crowd. Here are two examples of students who seemed to base their answers on generalizations of their own experience and, in this case, chose the answer with the highest count: "[T]hinking it is most probable to fail to notice something than to actually see something that is not there," and "I think this is correct because it is pretty easy to fail to notice someone when one counts from a long distance."

Theme 2 – Using alternative evidence: Four lower-level themes where students seemed to fail to notice the correct solution and instead used alternative evidence were grouped into one higher-level theme. Theme 2 (*using alternative evidence*) encompassed 352 responses where students used different kinds of alternative evidence as the premises for their reasoning.

Within this higher-level theme, the most common lower-level theme was 2a (*the person's experience, education, or knowledge*), which was used for 186 responses (53%). In 144 of these responses, students reasoned that the medical expert's statements were most believable, for example by writing, "The [medical expert] knows best," or "The [medical expert] has training in counting." Another example comes from a student who, on one of the items related to conflict of interest, seemingly failed to notice the conflict of interest and answered that both were equally believable. She reasoned, "The [geologist] and [medical expert] are educated and both have to be meticulous in their work. They probably watch in a meticulous and detailed manner to be certain that their answer is correct."

A total of 35 responses (10%) were coded as 2b (*majority, agreement, repetition, order*). On one item related to conflict of interest (seen in Table 1), one student who did not mention the conflict of interest chose that statement A was more believable "Because it was more people that were agreeing that it was [Captain James] to the right." Another student, on an item where the correct reason was related to the method of counting, answered that the order of counting was key: "The first [worker] counted them last which proves that there are only 32."

Moreover, 85 responses (24%) were coded as 2c (*better method or procedure of observation*). Most of these, 51 responses, were related to the item described in the section presenting Theme 1b (i.e., item 28), and one student wrote, "Because the [medical expert] have taken tests of the water and should then know if it is safe."

Finally, 76 responses (22%) were coded as 2d (*the person's mode of expression*). Some examples of this are, "Both describe a lot," "The mechanic was unsure," and "[I]t is most correct to believe the one that is certain."

Theme 3 – Lack of basis to choose A or B: Responses related to answering alternative C (statements A and B are equally believable) fell into Theme 3 (*lack of basis to choose A or* B), which encompassed 644 responses.

A total of 107 responses (17%) were coded as 3a (*observation statement and inference statement perceived as similar*). Naturally, all of these were related to items requiring students to discern between observation and inference. In these responses, students explicated that they viewed the statements as similar by writing, for example, "They said the same," and "Both mean the same."

A total of 518 responses (80%) were coded as 3b (*lack of information or evidence, or available evidence is similar for both statements*). Reasons for choosing alternative C (i.e., statement A and B are equally believable) were given this code, such as, "There is too little information to say who is right," "Because they disagree," "These parties are both soldiers," and "I think both answers are equally believable because everyone uses the same type of field glasses from the same distance and none of them have particularly better eyes than others."

Only 19 responses (3%) were coded as 3c (*alternative C because it is impossible to be 100% certain*). These stood out from the other codes under Theme 3 because these responses expressed students' seeming unwillingness to make a choice unless they could be completely sure. Students wrote reasons such as, "Nobody can be one hundred percent certain," and "[O]ne is not 100% sure in a case like this, and then it is almost always 50% true and 50% don't know."

Theme 4 – Logical fallacy: A total of 81 responses that did not seem to represent strong inductive logic were coded as 4 (*logical fallacy*). This contrasts Themes 1 and 2 which seem to represent strong inductive logic from the students' point of view, at least after they changed the context or used alternative evidence. That is, if the premises the students used are true, then their conclusions were likely true (i.e., inductively strong).

Of the 81 responses that constitute the higher-level theme *logical fallacy*, 66 responses (81%) were coded as 4a (*inference more or as probable as an observation*). This code has similarities with the conjunction fallacy (Tversky & Kahneman, 1983), the logical fallacy of judging the probability of a conjunction to exceed the probability of its constituents. This code showed up on items where there is one observational statement and one inferential statement, which might or might not be based on the observation. Table 1 shows an example of such an item. Students are given two statements, A and B, and are asked to decide which one is more believable, or if they are equally believable. Statement A might be: A worker says, "Five times now the person in the blue jacket has talked to someone and pointed, and immediately they have run off in the direction he pointed." Statement B: Another worker says, "He must be the leader." Here, the worker in statement B definitely infers more than the worker in statement A. Thus, statement A is more believable, according to the solution in the test manual. For this specific item, 38 students believed that the inference "he is the leader" is more or equally believable than the observation. One student wrote, "I think he is the leader because he is the only one with [a jacket] that stands out and he commands people around." Another wrote, "It seems right that he is the leader because there is one of him and everyone else goes where he points, like he is giving them orders." A total of 11 responses (14%) expressed that the students seemed to recognize a difference between a statement that was an observation and a statement that was an inference without considering that the observation was more believable. Thus, these were also coded as 6 (*right reason*). For example, one student wrote, "A pinpoints something, B uses what A says to figure out that he must be some sort of leader. Therefore C is the correct answer."

Responses that seemed not to make logical sense yet did not fit into 4a, a total of 15 responses, were coded as 4b (*other logical*

fallacies). These encompassed cases where the chosen answer did not match the reason or had reasons that in their expressed form seemed to be logically inconsistent with the chosen answer and the available information. On one of conflict-of-interest items, where the medical expert and the geologist still is in a word-against-word situation about whether the person is Captain James or not (i.e., like in the other conflict-of-interest item seen in Table 1), a student who answered alternative A wrote, "If they both say that they have paid attention, they have paid attention and it seems that they have kind of the same opinion." The reason does not seem to logically support choosing alternative A—it seems more supportive of alternative C—and further, contrary to expressing the same opinion, the item's underlined statements from the geologist and medical expert express opposite opinions. Generally, the examples of this theme require a more intricate description of the particular items as well as the items leading up to them. Thus, we propose that those interested in a few more examples see the supplementary materials, and the CCTT Level X.

Theme 5 – Lacking reason: In total, 729 responses did not contain any reason and thus were grouped together as Theme 5 (*lacking reason*). Of these, 429 responses (59%) were coded as 5a (*because / no reason*), where students typically wrote, "Because," "It felt right," and so forth. Furthermore, responses where students gave reasons such as, "[C]an not stand to spend time," and "[N]ow I am tired," were coded as 5b (*can't be bothered*). Most of these (94%) were removed from the dataset, only one remained, and thus the one single response coded as 5b that remained represented only 0.1% of all responses in Theme 5. A total of 10 responses (1%) were coded as 5c (*irrelevant*), including profane insults, creative use of letters and signs to form images, and other expressions not related to the test, such as, "[I] love food." Also, 14 responses (2%) were coded as 5d (*nonsense / gibberish*), such as "uhybyb." The code 5e (*guessing*), used 11 times (2%), was given to responses explicating that the students were just guessing. Finally, 264 responses (36%) were coded as 5f (*don't know*), which was used when students did not select an answer to the item or explicated that they did not know, such as, "[D]on't know," or "[D]idn't answer."

This theme was affected the most by cleaning the dataset. Before removal of 44 students, Theme 5 encompassed 1103 responses, 38% of the total number of uses of all higher-level themes. As Table 3 shows, this was reduced to 29% after the cleaning. This did not alter the ranking of the themes. Before the cleaning, Themes 5b, 5c, and 5d, were used 25, 120, and 228 times, respectively.

Theme 6 – Right reason: A total of 329 responses contained the correct reason and thus were grouped together as Theme 6 (*right reason*). Of these, 238 responses (72%) were also connected to the correct answer and thus coded as 6b (*right reason, right answer*), whereas 91 responses (28%) were connected to a wrong answer and coded as 6a (*right reason, wrong answer*). On the observation-versus-inference item seen in Table 1, a student gave both the right answer and the right reason: "[Worker] A has seen it himself and [worker] B is just guessing from the information he got from [worker] A." The coding criteria for right reason for different items can be seen in the supplementary material.

3.2. Distribution of themes and categories

The most common themes in student responses could indicate what barriers students most often face when answering items related to the categories of challenging critical thinking aspects (identified in our related study; Paulsen & Kolstø, 2021). As seen in Table 3, the most common themes for all categories were 3 (*lack of basis to choose A or B*) and 5 (*lacking reason*), indicating that the students might fail to notice the correct solution, or even any solution at all. The distribution of themes was fairly equal across the categories, with a few notable differences. First, 6 (*right reason*) was seen more in the method of observation category than in the two others, which makes sense considering that the items in this category had a higher rate of right answers (Paulsen & Kolstø, 2021). Second, the observation versus inference category had more logical fallacies than the other two because most logical fallacies were coded as 4a (*inference more or as probable as an observation*) which should only be used for this category. These two differences can be seen clearly in Table 4, which shows the distribution of categories within each theme, corrected for number of items in each category (i.e., by dividing the number of times a theme was used by the number of prompted items from each category).

4. Discussion

Based on the empirical data from secondary school students' written justifications to their answers on items of the *Cornell Critical Thinking Test* (CCTT) Level X, we found 21 lower-level themes encompassed by six higher-level themes of student reasoning for the chosen items. The results revealed that more than a quarter (28%) of the responses given expressed strong inductive logic but were based on irrelevant or incorrect premises. That is, when the students seemingly failed to notice the correct reason yet still wrote a reason, they used alternative evidence (14% of responses), or changed the context (14% of responses) either by misunderstanding the context or the item, or by using their own experience, generalizations, or made-up information so that their reasoning matched their

Table 3
Distribution of higher-level themes of student ($n = 240$) reasoning within each category and across all categories of critical thinking items.

	Observation versus inference	Conflict of interest	Method of observation	All categories	
Higher-level theme	%	%	%	#	%
Changing context	14%	20%	11%	360	14%
Using alternative evidence	16%	8%	16%	352	14%
Lack of basis to choose A or B	28%	30%	20%	644	26%
Logical fallacy	6%	2%	0%	81	3%
Lacking reason	30%	31%	27%	729	29%
Right reason	6%	9%	26%	329	13%

Table 4
Distribution of categories within each theme of student ($n = 240$) reasoning, corrected for numbers of items in each category.

Category of critical thinking items	Higher-level theme 1. Changing context	2. Using alternative evidence	3. Lack of basis to choose A or B	4. Logical fallacy	5. Lacking reason	6. Right reason
Observation versus inference	32%	41%	37%	73%	36%	16%
Conflict of interest	44%	21%	39%	26%	35%	23%
Method of observation	24%	38%	25%	1%	29%	61%

conclusion. Only 3% of the responses given did not seem to represent strong inductive logic, most of them (81%) being that the students seemed to believe an inference to be just as, or more, believable than an observation. Around 14% of the responses that fell within this theme, *logical fallacy*, expressed a recognition of the difference between observations and inferences without it impacting their choice of answer. From the students' points of view, their reasons were probably making logical sense, indicating that they perhaps lack knowledge or skills related to judging whether observations or inferences are more probable to be true. In 29% of the responses, students gave no substantial reason for their answers. Moreover, 26% of the responses expressed a lack of basis to choose a particular answer (i.e., choose alternative A or B). Finally, students touched upon the correct reasons in 13% of the responses, and 72% of these responses were also related to choosing the correct answer.

4.1. Possible barriers that students faced

Here we will discuss some possible barriers that students face in order to successfully solve the items from these critical thinking categories, and particularly how these barriers relate to skills, dispositions, and knowledge. We think it is important to note the context the students were put in by teachers and researchers. First, this was a low-stakes test because the students knew that the test results would not impact their grades. Second, there are indications that the general motivation for taking the test was low, such as Themes 5 (*lacking reason*) and 6 (*lack of basis to choose A or B*) being the most common themes for the responses. Some students also expressed in their written justifications that they were unmotivated and tired. Motivation has been shown to affect low-stakes test performance (Cole & Osterlind, 2008), and specifically, tests of critical thinking skills and dispositions (Bensley et al., 2016; Dehghani, sani, Pakmehr & Malekzadeh, 2011; Kim, Lee & Park, 2015). Thus, it is possible that many students would express higher levels of critical thinking in a different context, and that motivation and context are some of the main barriers.

For the rest of this section, we return our focus to the reasoning and critical thinking that was expressed within the context of this study, and the possible barriers related to this. The incorrect but logically strong reasons were evenly distributed between the higher-level Themes 1 (*changing context*) and 2 (*using alternative evidence*). Within these, the lower-level theme, 2a (*the person's experience, education, or knowledge*), was used more than the others. The responses within Theme 2a were appeals to authority, where students mostly used the medical expert and his expertise as evidence for one statement being more believable than another. We see one possible test-related reason why this argument was so common, representing almost 7% of all responses and 20% of the logically strong yet incorrect reasons from Theme 1 and Theme 2. At the start of Part II of the test, students read an example item where the reasoning behind the correct answer was that the medical expert should know better regarding the issue in question. Consequently, this might have led students to try out the same solution for subsequent items because it worked in the past. This type of fixed thinking represents a form of rigidity, often called a mental set, which might make it difficult to come up with alternative solutions (Schultz & Searleman, 2002). Thus, one barrier that students faced might be that their thinking was fixed on the same solution that was used in the example item of Part II. Appeal to authority is listed both as a valid argument and a fallacious argument depending on the context (Walton, 2008). In fact, for some items on this test, appeal to authority is the correct reasoning. However, appeal to authority is not the proposed correct reasoning (according to the test manual) for any of our chosen items that included prompts for written justifications. Here, appeal to authority is not necessarily fallacious, but students often misunderstood the context in some way or seemed to fail to notice other more relevant evidence.

Within Theme 1 (*changing context*), 22% of the responses were coded as Theme 1b (*giving an answer that does not address what was asked*). Most (86%) of the responses coded as 1b were related to item 28. For this item, one reason why most answers did not address what was asked could be that this seems to be a very difficult item with certain characteristics of a trick question. Only 7% of students answered this item correctly, and only 11% of those who answered this item correctly also gave the correct reason (Paulsen & Kolstø, 2021). The surface structure of the item is also similar to the example item that the students were shown at the beginning of Part II of the test. The example item involves a soldier and a medical expert that both are making inferences about water safety. The correct solution for the example item is that the medical expert's inference about water safety is more believable than the soldier's inference about water safety because of the medical expert's relevant expertise. This makes it easy to apply the same logic to item 28 and to miss the difference between the statements, where the medical expert took a larger leap in his inference than the soldier's minimally, if at all, inferential statement about the appearance of the water. In sum, we think that there could be grounds for removing item 28 from the test, although one weak item probably does not affect overall test validity.

Because the CCTT tests for critical thinking skills, it seems reasonable to assume that being less proficient in these skills than what is demanded to successfully solve the test items is a possible barrier that students face. However, it is possible to be proficient in the skills of critical thinking without being disposed to use those skills (Facione, 1990). Thus, dispositions are also important, and not being

disposed to think critically in the testing context might also be a barrier for exhibiting critical thinking to the best of one's abilities. Furthermore, even though the CCTT is created to minimize the requirements for background knowledge, there might be some general or specific knowledge that is either needed or useful to solve the test. Not possessing this knowledge might also be viewed as a barrier for critical thinking in this context. We will discuss skills, dispositions, and knowledge in the following paragraphs.

Part II of the CCTT, which deals with evaluating the credibility of observation reports, clearly corresponds well with the cognitive skill of evaluation from Facione's (1990) consensus definition of critical thinking. Thus, lacking in this skill might be a barrier for students when working with these items. The skill of evaluation includes being able to recognize factors relevant to assessing the degree of credibility with which to credit a source of information, which also directly relates to recognizing a conflict of interest and taking that into account when deciding what to believe about the believability of a statement or the credibility of a source. Furthermore, evaluation skills include being able to assess the degree of certainty one can place in the probability that a statement holds true, including judging whether a statement is purely an observation or if it is an inference, and in that case, to what degree it is inferential. We recognize that observations might not always be more believable than inferences. For example, it has been argued that observations in general are theory-laden, as perception is guided by past experience and knowledge (Estany, 2001). However, the kind of difference involved in the CCTT Level X is between descriptions of observations and inferences that are possible causes or implications of observation. The implicit claim about observations being more believable than inferences therefore seems reasonable in the given context.

Furthermore, because many students misunderstood the information they were given, they might not be proficient enough in the skill of interpretation (Facione, 1990). Again, this might be context specific, or related to dispositions, and it is possible that these students possess a higher level of interpretation skills than what is expressed in their test answers. However, expressed by their written responses, these students did not seem to comprehend the meaning or significance of certain statements and other information in the test, and thus failed to find the correct solution for some items. Regarding conflict of interest, for example, several students seemingly failed to notice that only one person was likely to have such motives behind their actions and instead believed that both persons involved had a conflict of interest.

Students also seemed to not be proficient enough in the skill of inference (Facione, 1990) and the sub-skills querying evidence, conjecturing alternatives, and drawing conclusions. Students seemingly did not gather enough evidence and alternative explanations and ended up drawing conclusions too quickly based on incorrect evidence.

In addition to the purely skill-based conceptualizations of critical thinking, some argue the importance of knowledge of key critical concepts (e.g., Bailin et al., 1999). Two examples of such key critical concepts could be to know that an observation is generally more probable to be true than an inference, and to know how a conflict of interest affects credibility, both of which one could argue would be helpful when solving our selected test items. Others have described how pieces of knowledge can be expanded into a process to increase the procedural knowledge substance and become more like general skills (Smith, 2002). For example, the knowledge of what a conflict of interest is could be expanded into a more general skill by describing it as a process of mapping out all potential actions, the potential benefits of all parties involved, and then decide who has what to gain from the choice of actions. Thus, when learning and practicing critical thinking skills it might be useful to also focus on operational knowledge (e.g., principles and step-by-step procedures) abstracted from good critical thinking. Consequently, such knowledge might support or enable the development of the required critical thinking skills through practice. Nevertheless, if someone does not know the difference between an observation and inference (e.g., expressed by the responses under Theme 3a), or have no concept of what a conflict of interest is, answering certain test items might be difficult. This raises an interesting question (which is beyond the scope of this study to answer) regarding the distinction between general skills and general knowledge. However, given that cognitive skills include, or at least relate to, relevant generic knowledge, one could argue that the ability to apply these insights are in fact the general skills that the test is supposed to measure.

It should be noted that the test makers maintain that the CCTT Level X is valid in *standard conditions* (Ennis et al., 2005), which includes adequate reading comprehension, full competence with language and concepts used in the test, desire to do well, sufficient sleep, among other variables. Thus, there are potential threats to test validity depending on whether standard conditions were met, for how many students these were met, and to what degree these were met. We have already discussed the motivation these students had for this test and the motivation seen in low-stakes tests in general, which is related to desire (or lack thereof) to do well. The need for standard conditions probably also applies to our modified test, and the modification could bring some insight concerning whether standard conditions were met. This insight from the written responses helped us clean the dataset, which would hopefully contribute to mitigating the lack of desire to do well. Through pilot testing we mitigated, at least partly, issues related to language comprehension. However, it seems likely that weaker students would struggle with understanding certain words, or with the sheer volume of the text. Future research should focus on removing uncertainty related to standard conditions by, for example, increasing the stakes of the test, controlling for reading level, and doing random sampling.

Moreover, it should be noted that the current study does not make any inferences concerning whether the students' critical thinking performance on the test reveals anything about their critical thinking performance in other contexts, particularly real-life contexts. One could speculate that the performance could generalize to other contexts insofar as they are similar to the one in the study. However, searching for a correlation with critical thinking in real-life contexts, and investigating transfer, could be aims of future research.

In addition to the skills put forward by Facione (1990), the dispositions that motivate the use of those skills are also important. For example, if one lacks self-confidence in one's own ability to reason, then one might be quick to choose alternative C (i.e., A and B are equally believable), and reason that there is no basis to choose one over the other like many students (26%) were. Nevertheless, the potential for any connection of our results and dispositions is highly speculative as the CCTT Level X is not designed to measure critical thinking dispositions. Ennis (1996b, p. 175) highlights that some fundamental problems of assessing critical thinking dispositions are

that they are not directly observable, and "that a disposition is something we want students to evidence on their own—without being pushed or prompted to evidence it." The same paper, however, suggests that combining written justifications with selected items from the CCTT Level X (i.e., a similar method as in the current study) could be used for developing a measure of critical thinking dispositions, and the paper presents preliminary yet promising results. We, like the authors of that paper, invite others to try a similar approach, and we think that this is an interesting aim for future research.

4.2. Implications for critical thinking instruction

Some students (i.e., at least the 44 that were removed from the dataset) seemed not to take the test seriously and were not motivated to make an effort. Instruction in critical thinking, like in other areas, should be engaging to be effective. In general, engaging teaching methods like dialogue, authentic instruction, and mentorship are also effective methods for teaching critical thinking (Abrami et al., 2015). Moreover, critical thinking is most effectively taught when it is an explicit focus of instruction (Abrami et al., 2008). Students seemed to struggle with using the correct evidence on items requiring considering a conflict of interest and the difference in credibility of an inference and an observation. Explicit knowledge of these concepts and how they affect credibility could improve performance in these aspects of critical thinking and provide a good basis for increasing evaluation skills for similar tasks. This study found that students often used alternative evidence (14%) or changed the context of the problems (14%). Thus, one way to improve students' critical thinking could be to work with authentic and interesting problems that require critical thinking while teachers support learning by explicating how skills, dispositions, and knowledge are used—particularly those related to weighing alternative solutions and evidence, gathering enough evidence before inferring a conclusion, and understanding and interpreting the problem and its context. Previous research has suggested that one way of achieving this goal is to initiate students into an environment where critical thinking is valued and where real-life issues are handled (Bailin et al., 1999; Snyder & Snyder, 2008).

4.3. Contributions of the study

We see three main contributions of this study. First, the written justifications, and the identified themes of student reasoning, provide insight into how students reason when they are facing these types of critical thinking test items. As discussed in Section 4.2, this could also be an important contribution concerning how to teach students critical thinking. Second, our findings represent a contribution concerning how to interpret results from these types of tests in relation to validity, as both our method and our findings can help reveal more of how test takers interpret and (mis)understand these types of items. For example, our findings show that students sometimes make up their own premises to use as evidence for their reasoning. In addition, the written justifications used in this study have revealed instances where students marked the correct answer but not due to the correct reason, and instances where students marked an incorrect answer yet gave a correct reason. Thus, the results of this study suggest that the inclusion of written justifications enables test administrators to base their scoring on a different rationale than the test manual. For example, one could give credit to a response that expresses that the test taker weighs different alternatives even though the marked answer is "incorrect" according to the test manual. As the method in this study helps reveal more nuanced insights into students' thinking, a third contribution relates to the development of new or improvement of existing tests of critical thinking. For example, it might be possible to identify and score other aspects of critical thinking (e.g., dispositions and intuition) through analysis of the written justifications than what is possible with the original test.

4.4. Strengths and limitations

As far as we are aware, this is the first study to include written justifications in the CCTT, opening up for potential insight into student reasoning for the selected items. We consider this a strength of the study. Moreover, we see it as a strength that we have mixed qualitative and quantitative methods in the data collection, analysis, and study design of this exploratory study. The benefits of this include complementarity, or illustration (Bryman, 2006), where the written justifications can elaborate on students' reasoning for the selected challenging items that Paulsen and Kolstø (2021) identified. The relatively high number of students in this study ($n = 240$, after cleaning the dataset) strengthens the external validity, or transferability of the results, while the written justifications provide qualitative data with more details. A potential limitation is that we have not statistically validated the Norwegian translation of the test and its use in a Norwegian setting. However, we did pilot tests in similar samples and several rounds of language revision. Finally, as mentioned at the start of Section 4.1, student motivation is known to affect test performance on low-stakes tests, which could limit how well the results generalize to non-testing or high-stakes situations.

Funding

This work was supported by the Research Council of Norway under grant number 275835.

Credit author statement

Vegard Havre Paulsen: Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Visualization. **Stein Dankert Kolstø:** Writing – Review & Editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors have no competing interests.

Acknowledgements

We would like to thank Mildrid Kyte for coding a part of the dataset. We would also like to thank Mildrid Kyte, Matthias Gregor Stadler, and Vegard Gjerde for valuable discussions of the thematic analysis. In addition, we thank the two anonymous reviewers providing thoughtful feedback that have been of great value in improving the quality of this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tsc.2021.100969.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically:A meta-analysis. *Review of Educational Research, 85*(2), 275–314. https://doi.org/10.3102/0034654314551063

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., et al. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research, 78*(4), 1102–1134. https://doi.org/10.3102/0034654308326084

Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing critical thinking. *Journal of Curriculum Studies, 31*(3), 285–302. https://doi.org/10.1080/002202799183133

Bensley, D. A., Rainey, C., Murtagh, M. P., Flinn, J. A., Maschiocchi, C., Bernhardt, P. C., et al. (2016). Closing the assessment loop on critical thinking: The challenges of multidimensional testing and low test-taking motivation. *Thinking Skills and Creativity, 21*, 158–168. https://doi.org/10.1016/j.tsc.2016.06.006

Bouygues, H.L. (.2018). *The state of critical thinking: A New Look at Reasoning at Home, School, and Work*. Retrieved from https://reboot-foundation.org/wp-content/uploads/_docs/REBOOT_FOUNDATION_WHITE_PAPER.pdf.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77–101. https://doi.org/10.1191/1478088706qp063oa

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research, 6*(1), 97–113. https://doi.org/10.1177/1468794106058877

Butler, H. A., Dwyer, C. P., Hogan, M. J., Franco, A., Rivas, S. F., Saiz, C., et al. (2012). The halpern critical thinking assessment and real-world outcomes: Cross-national applications. *Thinking Skills and Creativity, 7*(2), 112–121. https://doi.org/10.1016/j.tsc.2012.04.001

Butler, H. A., Pentoney, C., & Bong, M. P. (2017). Predicting real-world outcomes: Critical thinking ability is a better predictor of life decisions than intelligence. *Thinking Skills and Creativity, 25*, 38–46. https://doi.org/10.1016/j.tsc.2017.06.005

Case, R., & Wright, I. (1997). Taking seriously the teaching of critical thinking. *Canadian Social Studies, 32*(1), 12–19.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cole, J., & Osterlind, S. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education, 57*, 119–130. https://doi.org/10.1353/jge.0.0018

Dam, G., & Volman, M. (2004). Critical thinking as a citizenship competence: Teaching strategies. *Learning and Instruction, 14*, 359–379. https://doi.org/10.1016/j.learninstruc.2004.01.005

Davies, M. (2013). Critical thinking and the disciplines reconsidered. *Higher Education Research and Development, 32*, 529–544. https://doi.org/10.1080/07294360.2012.697878

Dehghani, M., sani, H. J., Pakmehr, H., & Malekzadeh, A. (2011). Relationship between students' critical thinking and self-efficacy beliefs in Ferdowsi university of Mashhad, Iran. *Procedia - Social and Behavioral Sciences, 15*, 2952–2955. https://doi.org/10.1016/j.sbspro.2011.04.221

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. (1993). Critical thinking assessment. *Theory Into Practice, 32*(3), 179–186. Retrieved from http://www.jstor.org/stable/1476699.

Ennis, R. H. (1996a). *Critical thinking*. Upper Saddle River, NJ: Prentice-Hall.

Ennis, R. H. (1996b). Critical thinking dispositions. *Their Nature and Assessability. Informal Logic, 18*(2), 165–182.

Ennis, R. H. (2016). Critical thinking across the curriculum: A vision. *Topoi, 37*(1), 165–184. https://doi.org/10.1007/s11245-016-9401-4

Ennis, R. H., & Millman, J. (2005). *Cornell critical thinking test: Level x in*. Seaside, CA: Critical Thinking Company, 5 ed.

Ennis, R. H., Millman, J., & Tomko, T. N. (2005). *Cornell critical thinking tests level x & level z manual*. Seaside, CA: Critical thinking, 5 ed.

Estany, A. (2001). The thesis of theory-laden observation in the light of cognitive psychology. *Philosophy of Science, 68*(2), 203–217. Retrieved from http://www.jstor.org/stable/3081064.

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. Newark, DE: American Philosophical Association.

Halpern, D. (2001). Assessing the effectiveness of critical thinking instruction. *The Journal of General Education, 50*, 270–286.

Kim, H., Lee, E., & Park, S.-Y. (2015). Critical thinking disposition, self-efficacy, and stress of Korean nursing students. *Indian Journal of Science and Technology,* 8. doi: 10.17485/ijst/2015/v8i18/76710.

Kjærnsli, M., & Jensen, F. (2016). *Stø kurs. norske elevers kompeanse i naturfag, matematikk og lesing i pisa*. Oslo: Universitetsforlaget.

Ku, K. Y. L. (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

Ku, K. Y. L., Kong, Q., Song, Y., Deng, L., Kang, Y., & Hu, A. (2019). What predicts adolescents' critical thinking about real-life news? The roles of social media news consumption and news media literacy. *Thinking Skills and Creativity, 33*, 1–12. https://doi.org/10.1016/j.tsc.2019.05.004, 100570.

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

Leach, S. M., Immekus, J. C., French, B. F., & Hand, B. (2020). The factorial validity of the cornell critical thinking tests: A multi-analytic approach. *Thinking Skills and Creativity, 37*, 1–14. https://doi.org/10.1016/j.tsc.2020.100676, 100676.

Li, Y., Li, K., Wei, W., Dong, J., Wang, C., Fu, Y., et al. (2021). Critical thinking, emotional intelligence and conflict management styles of medical students: A cross-sectional study. *Thinking Skills and Creativity, 40*, Article 100799. https://doi.org/10.1016/j.tsc.2021.100799

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

McPeck, J.E. (.1981). Critical thinking and education (Vol. 30): St. Martin's Press.

Mezzich, J. E., Kraemer, H. C., Worthington, D. R. L., & Coffman, G. A. (1981). Assessment of agreement among several raters formulating multiple diagnoses. *Journal of Psychiatric Research, 16*(1), 29–39. https://doi.org/10.1016/0022-3956(81)90011-X
National Research Council. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: The National Academies Press.
NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: National Academies Press.
Norris, S. P. (1990). Effect of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement, 27*(1), 41–58. https://doi.org/10.1111/j.1745-3984.1990.tb00733.x
OECD. (2015). *Education 2030 project proposal*. Paris: OECD Publishing.
Paulsen, V.H., .& Kolstø, S.D. (.2021). *Conflicts of interest, observations, and inferences – Challenging aspects of critical thinking [Submitted manuscript]. Department of Physics and Technology, University of Bergen.*
Robinson, S. R. (2011). Teaching logic and teaching critical thinking: Revisiting McPeck. *Higher Education Research & Development, 30*(3), 275–287. https://doi.org/10.1080/07294360.2010.500656
Schultz, P. W., & Searleman, A. (2002). Rigidity of thought and behavior: 100 years of research. *Genetic, Social, and General Psychology Monographs, 128*(2), 165–207.
Smith, G. (2002). Are there domain–specific thinking skills? *Journal of Philosophy of Education, 36*, 207–227. https://doi.org/10.1111/1467-9752.00270
Snyder, L., & Snyder, M. J. (2008). Teaching critical thinking and problem solving skills. *The Delta Pi Epsilon Journal, 50*, 90–99.
Thayer-Bacon, B. J. (2000). *Transforming critical thinking: Thinking constructively*. New York: Teachers College Press.
Tiruneh, D. T., Weldeslassie, A. G., Kassa, A., Tefera, Z., De Cock, M., & Elen, J. (2016). Systematic design of a learning environment for domain-specific and domain-general critical thinking skills. *Educational Technology Research and Development, 64*(3), 481–505. https://doi.org/10.1007/s11423-015-9417-2
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293
Utdanningsdirektoratet. (2019). Nasjonale prøver 8. og 9. trinn - resultater. Retrieved from https://www.udir.no/tall-og-forskning/statistikk/statistikk-grunnskole/nasjonale-prover-8.-og-9.-trinn/.
Ventura, M., Lai, E., & DiCerbo, K. (2017). *Skills for today: What we know about teaching and assessing critical thinking*. London: Pearson.
Walton, D. (2008). *Informal logic: A pragmatic approach*. Cambridge: University Press.

Online supplementary material

Calculating interrater reliability

Because the themes were mutually non-exclusive (i.e., there were several possibilities of theme-combinations for each answer) using Cohen's (1960) original Kappa formula might have been problematic. Thus, we calculated an interrater reliability score based on a modified version of Cohen's Kappa, *proportional overlap*, proposed by Mezzich, Kraemer, Worthington, and Coffman (1981). For each student answer, we calculated a proportion of agreement for each answer between the two raters (e.g., two overlapping themes of four unique themes gives an overlap of 0.5 for that answer) and averaged across all answers to find an overall observed proportion of agreement (Po) for the sample. To obtain a reliability measure corrected for chance, we determined a proportion of chance agreement (Pe). Based on the rate each rater used each theme we computed the probability of chance agreement between both raters for each theme and summed across all themes. We calculated a value of Po and Pe for two hierarchical levels of themes: the lower-level themes (i.e., the semantic codes) and the higher-level themes. Po and Pe can range from 0 to 1, no agreement to complete agreement. In order to generate the measure of interrater reliability, Po and Pe were applied to Cohen's original Kappa formula:

K=Po-Pe/1-Pe

Mezzich originally used this method to determine interrater reliability between four raters of psychiatric diagnoses. One other known study has applied this method (Eccleston, Werneke, Armon, Stephenson, & MacFaul, 2001). In that case, it was applied to qualitative interview data. There are no proposed criteria for reliability values using Mezzich's extension. However, for Cohen's Kappa there are several. For example, Cicchetti (1994) proposed this convention: 0.75–1.00 = *excellent*; 0.60–0.74 = *good*; 0.40–0.59 = *fair*; and < 0.40 = *poor*. Due to the proportional overlap, and the possibility of several themes per segment, the method used in this study might yield more conservative Kappa values than Cohen's Kappa. Thus, a number higher than 0.60 was deemed acceptable.

Coding criteria for right reason

Codes

1: Right answer and right explanation.

2: Wrong answer and right explanation. Notated as 2A, 2B, or 2C, depending on the chosen answer.

Observation vs inference:
28, 43, 44, 46

A right explanation should point out that there is a difference between the observation statement and the inference statement. This could be done by explaining the difference between the two statements,

by explaining that one of the statements is simply an observation, or by explaining that the information in the inference statement is more uncertain.

Sometimes students will be able to point out a difference but choose a wrong answer. This might be because they have a false belief that the difference is favoring the believability of the opposite of what is correct, that is, that an inference is more certain than an observation. In other cases, they might choose the wrong answer because they perceive other aspects to be more or equally important than the correct explanation (e.g., authority). Because these students are touching upon the correct explanation, such instances should get the code 2.

Method of observation:
41, 42, 45, (47, 48, 50)

A right explanation should point out that there is a difference between the methods used. This could be done by explaining the difference between the two methods, by explaining why one method is better, or by explaining why one method is worse.

Sometimes students will be able to point out a difference but choose a wrong answer. This might be because they have a false belief that the difference is favoring the believability of the opposite of what is correct, for example, that a statement based on written notes is LESS believable than a statement based on memory (e.g., because writing notes means not keeping eyes on the relevant situation). In other cases, they might choose the wrong answer because they perceive other aspects to be more or equally important than the correct explanation (e.g., authority). Because these students are touching upon the correct explanation, such instances should get the code 2.

Conflict of interest:
36, 38

A right explanation should point out that there is a conflict of interest, or that people are more likely to lie if it can benefit them.

Sometimes students will be able to point out a difference but choose a wrong answer. This might be because they have a false belief that the difference is favoring the believability of the opposite of what is correct, for example, that they perceive the conflict of interest to be in a different person than the correct explanation. In other cases, they might choose the wrong answer because they perceive other aspects to be more or equally important than the correct explanation (e.g., authority). Because these students are touching upon the correct explanation, such instances should get the code 2.

Examples of other logical fallacies (theme 4b)

One response to item 36:

The student answered A, and reasoned: “Because the man joined the group and another person sat down which means that it was not [Captain James].”

Two responses to item 38:

The first example comes from a student who answered A, and reasoned: “If they both say that they have paid attention, they have paid attention and it seems that they have kind of the same opinion.” The second example comes from a student who answered B, and reasoned: “Because a little above in the text we get to know that another person came and sat where the previous one was.”

References:

Cicchetti, D. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instrument in Psychology. *Psychological Assessment, 6*, 284-290. doi:10.1037/1040-3590.6.4.284

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*(1), 37-46. doi:10.1177/001316446002000104

Eccleston, P., Werneke, U., Armon, K., Stephenson, T., & MacFaul, R. (2001). Accounting for overlap? An application of Mezzich's kappa statistic to test interrater reliability of interview data on parental accident and emergency attendance. *J Adv Nurs, 33*(6), 784-790. doi:10.1046/j.1365-2648.2001.01718.x

Mezzich, J. E., Kraemer, H. C., Worthington, D. R. L., & Coffman, G. A. (1981). Assessment of agreement among several raters formulating multiple diagnoses. *Journal of Psychiatric Research, 16*(1), 29-39. doi:https://doi.org/10.1016/0022-3956(81)90011-X

ARTICLE III

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230844038 (print)
9788230844991 (PDF)