

Development of a multiple-choice test to assess critical thinking in climate change

Magdalena Micoloi^{1,*}, Lana Ivanjek², Mieke De Cock³, Thomas Schubatzky⁴,
Sarah Wildbichler⁴, Rainer Wackermann⁵, and Gesche Pospiech¹

¹*TU Dresden, Faculty of Physics, Physics Education Research Group,
Haeckelstraße 3, 01069 Dresden, Germany*

²*JKU Linz, School of Education, Altenberger Straße 69, 4040 Linz, Austria*

³*KU Leuven, Department of Physics and Astronomy and LESEC,
Celestijnenlaan 200c—Box 2406, 3001 Leuven, Belgium*

⁴*University of Innsbruck, Physics Education Research Group, Fürstenweg 176, 6020 Innsbruck, Austria*

⁵*Ruhr University Bochum, Physics Education Research Group,
Universitätsstraße 150, 44801 Bochum, Germany*



(Received 14 May 2025; accepted 12 November 2025; published 15 December 2025)

Recognizing the complex environmental challenges that humanity faces, there is a growing need to develop critical thinking skills specifically oriented toward understanding and addressing climate change. To facilitate the assessment of these skills, a critical thinking in climate change (CTCC) test was developed. In the first step, 16 semistructured interviews were conducted with high school and university students. Drawing on insights from these interviews alongside previous research, a first version of the CTCC test comprising 41 items was administered to a sample of $N = 289$ high school and university students from Austria, Belgium, and Germany. The test development process and subsequent evaluation were performed using Rasch modeling to determine test functioning. Through an iterative process that included three rounds of testing involving a total number of $N = 809$ students and analyses, the test items were revised. The final version of the CTCC test consists of 24 items. Results from Rasch analysis suggest that the test has high item reliability (0.96) and person reliability (0.80), making it a useful tool for assessing critical thinking skills among high school graduates and first-semester university students.

DOI: [10.1103/jnbq-1qyk](https://doi.org/10.1103/jnbq-1qyk)

I. INTRODUCTION

As the world encounters the increasing frequency and complexity of climate change risks [1] and misinformation about climate change is spread on social media [2–5], humanity needs to cultivate specific skills to effectively cope with corresponding challenges. Critical thinking (CT) enables individuals to distinguish between facts and misinformation or “fake news,” also in the climate change debate [6,7]. The complexity of the ecological challenges facing humanity requires the ability to analyze, question claims, and make well-informed decisions within the discourse on climate change. Addressing climate change involves more than only adopting sustainable practices, such as using renewable energy; it requires an understanding of the complex system that contributes to the problem [8] and the capacity to make “reasoned and evidence-based investigations of a very wide range of personal and social issues and

concerns” [9]. Particularly when it comes to climate change, students should be supported in developing “problem-solving and collaboration skills; (...) critical thinking, media literacy skills, and systems thinking; (...) to confront and reduce the fear and disempowerment that learners might experience in the face of the planetary crises” [10].

Facione [9] noted that “CT is vitally important in the personal and civic life of all members of society.” CT is one of the four citizenship competences named by the European Commission and embedded in the curricula of many EU states, mostly at secondary school level [11] but also emphasized in higher education [12]. Similarly, climate change education is included in numerous curricula at different levels. According to a UNESCO report [13], 87% of the 80 countries surveyed incorporate climate change in primary or secondary education. “However, only 38% have a law, policy and/or strategy specifically focused on CCE” [13] (CCE = climate change communication and education).

Despite these curricula integrations and the well-documented importance of CT and climate education, research shows that high school and university students lack CT skills [14,15] and have difficulties understanding climate change [16,17]. To support the development of students’ CT skills, it is necessary to assess their current level and

*Contact author: magdalena.micoloi@tu-dresden.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

identify any gaps. Effective education on complex topics and curriculum development rely on recognizing where students struggle and on identifying their difficulties [18]. A well-designed assessment is the key to revealing those areas that require further instruction.

For this purpose, a multiple-choice test was developed in the project tEachiNG criticAl thinkiNG on cLIimate chaNGe (ENGAGING), which aims to evaluate and foster the critical thinking skills of high school and university students in the context of climate change. The test targets both groups and is designed to assess the following learning outcomes:

1. Use model experiments and experimental data to evaluate hypotheses related to climate events.
2. Differentiate between correlation and cause in the context of climate change.
3. Recognize the need for a sufficiently large sample size to draw conclusions and understand the importance of good experimental conditions related to climate change.
4. Evaluate conclusions and recognize false reasoning in the given argumentation on climate change.
5. Make informed assessments about future climate change scenarios based on known probabilities.
6. Effectively use diagrams and graphs as problem-solving tools and distinguish between relevant and irrelevant information to support problem-solving and decision-making related to climate change issues.

These learning outcomes were defined based on Halpern's critical thinking framework [19] and selected content areas related to climate change. They guided the development of the test items to ensure alignment with the targeted skill set and content. Halpern's framework and content areas of climate change are described in detail in Sec. II. This paper presents the development and evaluation process of the test. We present the research questions in Sec. III. Section IV gives an insight into the research design and methodology, especially into Rasch modeling. In the final sections, we discuss the results (Sec. V), the limitations (Sec. VI), and what can be concluded (Sec. VII) from the study.

II. LITERATURE OVERVIEW AND THEORETICAL FRAMEWORK

A. Critical thinking

The essence of CT and how to promote it has been discussed for quite some time. Since the early 1980s, CT has played a significant role in the U.S. education system, particularly following recommendations by the Rockefeller Commission on Humanities that urged the U.S. Department of Education to incorporate CT into its definition of basic skills, which rapidly increased interest in this topic [20]. This may also be one of the reasons why many American philosophers and educational researchers at the end of the

20th century, such as Robert Ennis, John McPeck, Peter and Noreen Facione, David Perkins, Matthew Lipman, and Diane Halpern, focused on this subject. However, there has been disagreement regarding the definition of CT and methodologies for its instruction. Definitions range from Ennis's emphasis on "reasonable reflective thinking" [21] to McPeck's focus on "reflective skepticism" [22], and Facione's view of CT as a process of judgment involving interpretation, analysis, and evaluation [9]. Halpern, on the other hand, sees CT as "the use of skills or strategies to increase the likelihood that a desired outcome will occur," explicitly pointing out that "desired" is defined by the individual itself [23]. Moseley *et al.* [24] compare CT skills with Bloom's taxonomy, but stress that CT typically involves higher-level cognitive skills rather than mere recall of information. When engaging in CT, it becomes clear that there is an increased focus on the "how" and "why," as compared to lower order knowledge levels, which focus simply on "what" [25]. Davies and Barnett [26] further articulate the multidimensional nature of CT by distinguishing between various cognitive levels: lower-level thinking skills (e.g., interpretation), higher-level thinking skills (e.g., analysis), complex thinking skills (e.g., evaluation, reasoning, and problem solving), and thinking about thinking (e.g., metacognition). This highlights the different levels of CT, from basic comprehension and understanding of information to more advanced thinking processes, emphasizing the importance of metacognition in reflecting and improving one's thinking. While various definitions of CT exist, this study adopts the view, in line with Halpern's definition, that CT is goal-directed reasoning encompassing problem-solving, formulation of inferences, calculation of likelihoods, and reasoned decision-making based on context-specific cognitive strategies [19].

The diversity in definitions of CT can be attributed to its interdisciplinary origins, with contributions emerging from psychology, educational science, and philosophy [27]. Different approaches have led to different definitions and emphases, making it more complex to analyze CT. However, even if there is some disagreement regarding the exact definitions, there is a consensus that CT involves a combination of certain skills and dispositions [9,19,27–29].

Dispositions, understood as affective states that include attitudes toward critical inquiry and a readiness to apply critical reasoning, encompass qualities such as curiosity, open-mindedness, understanding of other people's opinions, flexibility in considering alternative ideas, and self-confidence in one's own ability to reason [9,26,30]. A proficient critical thinker is characterized by the willingness to plan, be flexible, be persistent, correct errors, and work toward consensus when needed [19].

In addition to the dispositions mentioned, critical thinkers need specific skills. Skills that overlap across various definitions of CT are the ability to analyze arguments, to make inferences by using inductive or deductive reasoning,

the ability to make a judgment or evaluation, and problem solving and decision making [27]. It is precisely these skills that Halpern [19,23,31] explicitly defines as the five core CT skills: verbal reasoning, argument analysis, hypothesis testing, likelihood and uncertainty analysis, and problem solving and decision making. These are briefly elaborated below.

1. Verbal reasoning

This skill refers to an understanding of information in verbal or written form. Knowing reasoning patterns is essential, especially to defend against the manipulations common in everyday language. Language is often deliberately used falsely to deceive or lure people, as is the case with advertising [19]. Effective reasoning includes recognizing premises, conclusions, and key linguistic quantifiers (e.g., “all,” “none,” “some,” “only”) that shape the structure and clarity of information. Furthermore, the differentiation between deductive reasoning, concluding in a top-down fashion, and inductive reasoning, concluding in a bottom-up manner, is crucial for recognizing logical fallacies and ensuring sound reasoning [19].

2. Argument analysis

Argument analysis presents a considerable challenge, as arguments are often complex and multifaceted in real-world contexts. This skill requires deconstructing complex arguments by identifying premises and at least one conclusion, and to recognize justifications, counterarguments, underlying assumptions, and sometimes irrelevant information [23]. This is essential to make informed decisions. In addition, assessing the credibility of a source of information, recognizing the differences between opinions, reasoned judgments, and facts, and recognizing and refuting different fallacies in arguments are part of this skill set [19]. For instance, the part-whole fallacy occurs when it is incorrectly assumed that what holds true for the parts of an entity must also hold true for the whole or vice versa [19]. Consider the following example: *Birds have feathers. Feathers are very light. Therefore, birds must be very light.* This line of reasoning illustrates the fallacy by erroneously extrapolating from the properties of a part to the whole.

3. Hypothesis testing

Hypothesis testing is strongly associated with scientific inquiry. Nevertheless, its underlying principles are also relevant in everyday life. It involves the formulation of hypotheses based on observations, followed by the systematic evaluation of these hypotheses through the collection and analysis of relevant data. This process allows individuals to determine whether the available evidence supports or contradicts a given hypothesis. Effective hypothesis testing requires an understanding of the need for sufficiently large and representative samples, controlled

experimental conditions, and caution when drawing generalizations, particularly from small or biased datasets [19]. A key component of this skill is the ability to differentiate between correlation and cause. One common error in this context is the fallacy of false cause, in which a causal link is mistakenly inferred between two events simply because they occur simultaneously or sequentially [19].

4. Likelihood and uncertainty analysis

Uncertainty and probability play a vital role in our lives. This is precisely why this ability is important, as it involves estimating the probability of events and making decisions based on it. In the case of classical probability, probability is defined as the ratio of the number of certain outcomes to the total number of possible outcomes, assuming that all outcomes are equally likely. However, people tend to display overconfidence in favorable outcomes, underestimate the role of chance and frequently construct post hoc explanations for events that are, in fact, random in nature [19]. When making predictions about future events based on existing data, extrapolation is used to extend the observed trends. However, critical thinkers need to understand the limitations of extrapolation, which is connected to the variables being extrapolated, the quality of existing data, and underlying conditions [19].

5. Problem solving and decision making

The competency of problem solving and decision making incorporates all core CT skills to effectively address complex situations and make informed decisions. Central to this skill set is the ability to define a problem, generate potential solutions, and systematically develop and evaluate alternative courses of action [19]. Creative thinking is included in this skill and is essential in generating innovative solutions [23], while a systematic evaluation process, considering both positive and negative potential outcomes, is necessary to avoid confirmation bias and other common pitfalls in decision making [19]. To enhance decision making, it is useful to begin with a precise definition of the decision to be made, followed by the systematic listing and evaluation of possible alternatives based on relevant criteria. Problems encountered may be well-defined or ill-defined [19]. Well-defined problems have clear solution paths, goals, and often one correct answer. Ill-defined problems have a variety of interpretations and different objectives. In our lives, we usually face problems that cannot be solved with only one correct answer. Tools such as graphs, diagrams, hierarchical trees, matrices, and models can be helpful to solve a particular problem [19]. Importantly, critical thinkers must be able to distinguish between relevant and irrelevant information to maintain focus on the actual problem and avoid being misled by extraneous details.

The CT skills outlined above form a foundational framework for a skill-based approach to CT instruction. This framework emphasizes the development of teachable

and transferable skills that bridge the gap between the thinking skills cultivated in educational settings and those required in professional environments [23]. These CT skill categories are used in the assessment of educational outcomes [32]. They can be seen as specifications for assessments, similar to including different types of math problems in a computational test and should not be seen as independent dimensions [31]. Although Halpern's framework was originally developed with a focus on everyday situations, its five core skill categories are also fundamental to scientific thinking. In science education, and particularly in physics, these skills are needed, for example, to formulate hypotheses, interpret data, and draw conclusions based on empirical evidence. Accordingly, several studies in physics education have adopted Halpern's framework [33–35]. Given the increasing emphasis on connecting physics education to societally relevant issues, topics such as climate change provide a rich and meaningful context for fostering critical thinking.

Another ongoing debate in literature concerns the nature of CT, whether it is a general skill set that transfers across domains or whether it is inherently tied to specific disciplinary contexts [21–23]. One perspective argues that CT involves domain-general skills, i.e., skills that do not focus on a specific content area [23,36,37] and are applicable across a range of everyday scenarios or familiar contexts. Existing test instruments, such as Halpern's Critical Thinking Assessment (HCTA) [31], the Cornell Critical Thinking Test (CCTT) [38], the Ennis-Weir Critical Thinking Essay Test [39], and the Watson-Glaser Critical Thinking Appraisal [40], aim to measure these general CT skills. In contrast, another approach holds that CT is best understood and developed within specific disciplinary contexts [22], given that effective reasoning often relies on content knowledge and contextual understanding. This conceptual debate has direct implications for the development and selection of CT test instruments, whether they should assess general or domain-specific CT skills.

Many studies have demonstrated that integrating the skills into a specific context enhances learning outcomes (see, for example, Refs. [25,35,41,42]). These findings suggest that students are more likely to develop robust CT skills when they are applied to subject matter from their academic major or to a course-specific context [25,43]. As a result, varying instructional and assessment strategies for CT have emerged across disciplines, including nursing education [44], mathematics (e.g., [45]), biology (e.g., [46]), physics (e.g., [34,35,47]), and in intersection of STEM disciplines (e.g., [48]). In line with this previous research, we adopt a domain-specific view of CT. For example, Tiruneh *et al.* designed teaching-learning sequences for promoting domain-specific CT within the context of electricity and magnetism, and their findings indicated that learning gains occurred primarily in domain-specific CT, which was fostered through systematically designed, domain-specific instruction [49]. Further research is needed

to explore how such instruction can foster transfer from one domain to another or from domain-specific to domain-general CT.

Although there is research on domain-specific CT within different disciplines, a gap remains concerning climate change. Established general CT assessments, such as the HCTA, provide valuable measures of general CT skills, but they are not suited to the specific demands posed by domain-specific content, such as climate change. HCTA evaluates CT through scenarios set in various real-world contexts, including educational policy and healthcare settings [31]. These contexts are used to assess the examinee's ability to analyze assumptions, interpret data, evaluate argument strength, and identify reasoning flaws. Our instrument integrates CT skills within authentic tasks related to climate change, enabling more precise evaluation of students' skills in this context. To our knowledge, there is no established and reliable instrument for assessing CT skills in the context of climate change. Given that effective CT assessment should relate to authentic, real-world problems [23,50], the domain of climate change offers a particularly relevant and meaningful, and at the same time discipline-specific, context in which to evaluate domain-specific CT skills.

B. Climate change (CC)

Climate change is a pressing global problem of our time. It is addressed in science and at school as well as in the media. The wealth of information from the media on climate change, in particular, contributes to students developing different ideas [51]. However, according to the USGCRP, a solid understanding of the scientific principles of climate change is one of the four dimensions of becoming a climate literate person [8].

On the whole, climate change is a broad, interdisciplinary field of research in which there are many different elements and processes to consider. Narrowing down and defining specific concept areas is helpful to be able to focus on relevant aspects. Especially those aspects where students still have difficulties should be examined closely. Key concept areas that have been defined as important for dealing with climate change and should be addressed in climate change education are (1) the Earth's atmosphere, (2) climate as a system, (3) the carbon cycle, (4) weather vs climate, and (5) the greenhouse effect [52].

In the following, the important aspects of the different concept areas are briefly summarized, highlighting students' difficulties and specific ideas related to these aspects. For a more in-depth discussion about the concept areas and students' ideas, we refer to [17,53,54].

1. The Earth's atmosphere

This concept area primarily includes an understanding of the atmosphere's composition, structure, and size. Lower layers are particularly important for weather and climate

phenomena, as they hold most of the air and are influenced by the Earth's infrared (IR) radiation. When it comes to the composition of the atmosphere, a distinction must be made between nongreenhouse gases and greenhouse gases. Greenhouse gases, present only in traces, play a decisive role, as these gases interact with IR-radiation but not with visible light.

Previous studies show that students have difficulties in knowing which gases are greenhouse gases. A significant number of students focus solely on CO₂ as a greenhouse gas, unaware that water vapor also plays a critical role in the greenhouse effect [16,55–61]. Similarly, the concentration of greenhouse gases, especially of CO₂ in the atmosphere is estimated to be much higher than in reality [16,55]. There is also a misconception about an accumulation of greenhouse gases in upper layers of the atmosphere, comparable to a glass roof of a greenhouse, whereas they are evenly distributed [58,61–64].

2. Climate as a system

Earth's climate is a complex system in which various components, such as the atmosphere, oceans, and terrestrial environments, interact with each other and exchange energy and matter. Changes in the climate system are caused by feedback effects, natural influences, and human activities, involving the burning of fossil fuels. Students should understand the climate as a system to avoid seeking one-sided solutions and to be careful with interventions in this multiplex system.

The interplay of different components, sometimes involving only small quantities that are having an impact on the overall climate, is difficult for students to understand [16]. Concerning tipping points, a study among the British public shows that a quarter of those surveyed ($N = 1773$) had never heard of tipping points [65], and when looking at feedback effects in the climate system, students are usually not familiar with this concept [16].

3. The carbon cycle

Understanding the carbon cycle is about recognizing that for centuries the exchange of carbon has been undisturbed and consists of an interplay between the different reservoirs, such as atmosphere, oceans, and living organisms. Human intervention in this cycle, through the burning of fossil fuels and deforestation, destroys the natural balance. Particularly, the aspect of human influence should be emphasized here.

For students, it is difficult to connect their own understanding of the carbon cycle and the influence of humans on climate change [56,59,66,67]. Their focus regarding the carbon cycle is mainly on atmospheric CO₂ [16]. In addition, students do not have a good understanding of the exchange of carbon between reservoirs, for example, through the solubility of CO₂ in water and the resulting impact of increased atmospheric CO₂ on oceans [68].

4. Weather vs climate

The key aspects of this area are the definition of and distinction between weather and climate. Weather is directly perceptible as the atmospheric conditions at a specific place and at a specific time. Climate, on the other hand, is described in the long term by statistics such as average temperatures and their changes and can be specified on a global scale. Whereas the weather can only be predicted for a few days, climate models provide highly probable projections for decades.

Regarding the distinction between weather and climate, students tend to confuse the terms or mix them up [55,58,69,70]. In addition, students draw conclusions about the climate based on information about local weather [71]. They also claim to directly witness climate change by referring to individual weather events [72].

5. The greenhouse effect

The greenhouse effect is a complex phenomenon, elucidated below through a series of simplified steps: solar radiation of different wavelengths reaches the Earth, which is surrounded by the atmosphere. Some of the incoming solar radiation is reflected immediately (e.g., by clouds), some is absorbed in the atmosphere, and the rest reaches the Earth's surface directly. Depending on the surface properties, the incoming radiation is reflected (in bright areas such as ice and snow) or absorbed (in dark areas such as rock or water). By absorbing visible solar radiation, the Earth's surface heats up and emits long-wave IR-radiation because of its temperature. The present greenhouse gases in the atmosphere (such as CO₂, methane, water vapor, etc.) interact with the IR-radiation emitted by the Earth, causing the atmosphere to heat up and also emit IR-radiation in all directions, including back toward the Earth's surface. This process explains the natural greenhouse effect, which enables the vital conditions for humans. In the case of radiative equilibrium, the Earth emits the same amount of energy as it absorbs from the Sun. What is added by the anthropogenic greenhouse effect, however, is further warming due to an increasing concentration of CO₂ in the atmosphere. This leads to a temporary radiative imbalance, because with a higher CO₂ concentration in the atmosphere, more IR-radiation from the Earth's surface is absorbed by the atmosphere. Consequently, more energy remains on Earth.

Numerous studies have investigated students' conceptions of the greenhouse effect. For instance, students believe that global warming is caused by an increase in solar radiation [59,60,62–64]. Furthermore, the depletion of the ozone layer is believed to intensify the greenhouse effect by allowing more energy to reach the Earth's surface [16,56,58–61,66,69,70]. Moreover, only a few students distinguish between the natural and anthropogenic greenhouse effect [56–59,61,63].

As can be seen in the different concept areas, various student difficulties and ideas have already been examined in the context of climate change. Not all student ideas correspond with scientific explanations. However, these incorrect student ideas should not purely be seen as an obstacle, but as an opportunity to promote learning processes [73].

To assess students' difficulties on climate change, dedicated test instruments have been developed (see, for example, Refs. [52,53,57]). Research-based teaching materials targeting the identified difficulties can be used to improve students' conceptual understanding in the context of climate change [74].

III. RESEARCH QUESTIONS

Based on the existing research referred to above, we focus on still open aspects regarding a test instrument specifically assessing critical thinking skills in the context of climate change, and we seek an answer to the following research questions:

- (1) How can critical thinking be measured in the context of climate change among high school and university students?
- (2) To what extent are the identified critical thinking skills developed among students in the context of climate change?
- (3) What are the main difficulties of students when it comes to critical thinking in the context of climate change?

To answer these questions, both the test development and the findings of the assessments are presented.

IV. RESEARCH DESIGN AND METHODS

In this section, the necessary background on one-dimensional dichotomous Rasch modeling is provided, as it forms the foundation for the development of the test instrument. For a more complete description of Rasch modeling, we refer to [75–78]. Rasch modeling has already been used to develop and analyze test instruments for quite a number of studies in educational sciences, especially in physics education (see, for example, Refs. [52,79–84]). Below, the specific steps of test development, such as interviews and sampling, are presented.

A. Use of Rasch model for test construction and evaluation

The test instrument was developed following the 10 steps as proposed by Liu [76] and Planinic *et al.* [79]:

- (1) Defining the construct to be measured.
- (2) Identifying different levels of the construct corresponding to different cognitive levels.
- (3) Deriving of items from the construct (item pool).

- (4) Field testing with a representative sample of the target population.
- (5) Conducting Rasch modeling.
- (6) Evaluating fit statistics and revising the test items.
- (7) Reviewing the Wright map and adding or deleting of items.
- (8) Repeating the steps (4) to (7) until a set of items fit the Rasch model and defining a scale.
- (9) Determining the validity and reliability claims of the measurement instrument.
- (10) Developing the documentation for the measurement instrument.

In our study, the construct to be measured is defined as critical thinking in the context of climate change (step 1). Following the theoretical framework (see Sec. II) and the explanation by Halpern that the five CT skills should not be seen as independent dimensions [31], it is operationalized as a linear, unidimensional attribute. A clearly defined construct forms the foundation for the subsequent development of test items.

The items have been constructed in a way to ensure that different levels of a construct (critical thinking in the context of climate change) require different levels of cognitive reasoning (step 2). This has been done by organizing the construct around the five CT skills. In the next step, an initial item pool has been developed, consisting of 35 open-ended questions (step 3). These were evaluated through interviews with high school and university students. From that, the first version of the multiple-choice test critical thinking in climate change (CTCC) with 41 items was developed and administered in a pilot study to high school and university students (step 4). The data obtained were then analyzed using Rasch analysis [85] via Rasch modeling software (step 5).

This allows for a detailed examination of fit statistics and revisions of the test (step 6). Diagnostic measures provided by the Rasch model include the Infit and Outfit MNSQ values as well as person and item reliability [77,78], which are crucial for evaluating data-model fit. The Infit MNSQ statistic, as the weighted mean of the squared residuals, sensitive to responses near a participant's ability level, and the Outfit statistic, as the arithmetic mean of simple squared residuals, influenced by outliers, are used to evaluate whether each item conforms to the model. Fit values ideally range from 0.7 to 1.3, with broader acceptable limits for the first round extending from 0.5 to 1.5 [76]. Items that fail to meet these criteria need to be either revised or excluded. Poor fit values for items may stem, for example, from unclear wording or content issues of a specific item. However, misfit can also occur at participant level and needs to be checked. A person's poor fit can be caused, for example, by guessing or not reading and answering the items carefully. If necessary, participants with poor fit and low abilities need to be removed to avoid

TABLE I. Short description of the CT skills and overview of the distribution in the final version of the *CTCC* test with overall 24 items.

CT skills	Description	Items
Verbal reasoning	Students need to understand the given information and recognize misleading language, for example, when analogies are made or quantifiers are used.	9, 13, 21
Argument analysis	When students analyze the given arguments, they need to understand their structure and complexity, assess the credibility of sources, and be careful when conclusions are drawn that cannot be supported.	1, 3, 4, 10, 22
Hypothesis testing	Students should recognize the need for a sufficiently large sample and good experimental conditions before making generalizations. Students should also use given data or experiments to test hypotheses.	5, 6, 8, 11, 14, 17, 18
Likelihood and uncertainty analysis	Students estimate probability of events, use them to guide decisions and understand the limits of extrapolation when predicting future events.	7, 15, 19
Problem solving and decision making	When solving problems or making decisions, students should use all the skills mentioned as well as the graphs and information provided for their solving process or decision.	2, 12, 16, 20, 23, 24

distorting item fit values. To interpret the relationship between participant abilities and item difficulties on a linear logit scale, a Wright map is employed (step 7). This visualization tool not only shows the distribution of items across the ability range but also aids in identifying gaps where additional items might be needed or redundant items removed [79].

The testing as well as Rasch analysis procedures are repeated in iterative loops until an adequate test instrument is achieved. In our case, it was 3 rounds of testing and analysis (step 8). This iterative process ensures that the instrument exhibits high internal consistency. In accordance with this, Rasch analysis of the data from the pilot study was conducted and content considerations were discussed in the group. As a result, the items were revised and the adapted test was administered to a new sample. Again, the data were evaluated using Rasch analysis. After final adjustments, the test instrument entered the third survey phase, followed again by a Rasch analysis. After this analysis, 17 students have been removed from the sample because of their very high Outfit values connected with very low ability, which indicates guessing. In addition, 3 items have been removed from the test because the Outfit values were slightly higher than 1.3 and to further reduce the testing time for students. The final version of the *CTCC* test consists of 24 items and is structured according to combinations of CT skills and concept areas of climate change, meaning that in each item, we focus on one CT skill and one concept area of climate change. Table I shows the distribution of the items and operationalizes CT as defined in our research framework.

With regard to construct validity, it is evaluated through Rasch analysis, including the assessment of one-dimensionality, verified by Principal Component Analysis of Residuals (PCA-R), as well as fit statistics and item

correlations. In addition, the test instrument underwent expert review. Subject-matter experts scrutinized each item to determine its alignment with the defined construct, reinforcing both face and content validity [86]. Test reproducibility is assessed through the calculation of item and person reliability metrics (step 9). The closer the values are to 1, the better the test can capture differences between different participants (person reliability) and the more precisely the relative difficulty of the items is estimated (item reliability), suggesting good reproducibility for subsequent measurements. However, it is important to note that high reliability values alone do not provide information about the quality of the test. The entire process described must be considered for evaluating the test's quality. The last step of test development regarding Rasch, the documentation of the entire process, is dealt with in this paper (step 10).

B. Interviews

First, 16 semistructured think-aloud interviews were conducted. Eight preservice teachers at TU Dresden (two female, six male; from 4th to 9th semester) and eight high school students from Saxonian high schools (seven female, one male; around 15 and 16 years old) voluntarily participated in the interviews, employing a convenient sampling approach [86]. All participants have given their consent for the interviews to be recorded and were informed that they could stop the interview at any time. Each interview took on average 2 h to discuss the 35 developed items. The interviews primarily aimed to derive realistic distractors from the participants' responses, while also identifying students' argumentation patterns and challenges related to critical thinking in the context of climate change. The students at TU Dresden represent a good mix from preservice teachers from higher semesters to relatively beginners. They had not attended any specific

climate change or critical thinking courses before. The age of the participating high school students ranges from 15 to 16 years. In school, they had only dealt with climate change in geography lessons, not in physics class.

When conducting the interviews, the participants were asked to respond to the questions given and say everything that came to mind. They were also told that there were no wrong answers in order to motivate them to speak their thoughts on the text without hesitation. This method is referred to as think-aloud interviews [87,88]. This enabled the interviewer to recognize the reasoning behind the students' answers and, if necessary, to address any ambiguities and ask for more detailed explanations. The interviews were recorded, transcribed, categorized, and analyzed using qualitative content analysis [89].

C. Test development and sampling

After the interview analysis, out of the original 35 open-ended interview questions and corresponding student responses, a first version of the multiple-choice test *CTCC* was developed and administered in German. This version comprised 41 multiple-choice items with a single correct answer per item. We have decided to test a larger pool of items in order to find the best set of items after analysis. For pilot testing, the test was divided into two booklets, each containing 26 items, including 11 anchor items in both booklets. When allocating items to the two booklets, considerable care was taken to ensure that both groups were presented with comparable tasks. For instance, effort was made to balance the number of graphical analyses required by each group, or in some cases, two similar versions of one question were tested.

The test was conducted in all three rounds using an online survey program. Participants were informed about data protection measures, which all consented to. Participation was voluntary, with no disadvantages imposed on students who chose not to take part in the study. A convenience sampling approach was employed for data collection [86].

In the first study, $N = 289$ students participated. A total of 191 of the participants were university students (79 female, 103 male, 4 diverse, and 5 no response) from six universities across Austria, Belgium, and Germany spanning from the first year of studying physics to Ph.D. candidates. Students had no dedicated courses on critical thinking or climate change as such courses are not part of the physics curriculum. The other 98 were high school students (50 female, 40 male, 2 diverse, and 6 no response) from 10th to 12th grade (16 to 18 years old) from two Viennese high schools (Austria). As well, high school students had received no prior instruction on critical thinking or climate change. The average time needed to complete the test was approximately 35 min. Collected data were analyzed using the Winsteps software for Rasch data analysis [90] to examine item performance and overall test function. Based on the quantitative fit statistics and

qualitative item reviews, the test was modified, with 14 items being removed entirely and others slightly rewritten. In addition, five new items were developed to address gaps identified in the item-person map and to ensure balanced representation of all CT skills across the test. This resulted in a revised instrument containing 32 items.

For the second test run, the instrument was again divided into two booklets, each now comprising 22 items per booklet linked with 12 anchor items. A total of $N = 340$ participants took part, consisting of 244 high school students from 10th to 12th grade (88 female, 100 male, 3 diverse, and 53 no response) and 96 university students (30 female, 62 male, and 4 no response) from Austria and Germany in their first year of physics to Ph.D. candidates. In this round, the average completion time was reduced to approximately 30 min. A further Rasch analysis was performed, resulting in removing five items due to misfit and additional content considerations. In addition, we wanted to reduce testing time.

A third data collection of the *CTCC* test was implemented using a single booklet containing 27 items. A more coherent sample was targeted in terms of prior knowledge and age, which included 116 high school graduates (57 female, 57 male, and 2 no response) and 64 first-semester university students (15 female, 47 male, and 2 no response) from Austria and Germany. Again, participants had not received prior instruction on critical thinking or climate change. The allocated test time was approximately 35 min. A final Rasch analysis was conducted, resulting in the removal of three items and 17 misfitting students, leaving 163 students for the last analysis. The final version of the *CTCC* test consists of 24 multiple-choice items with one correct answer each.

Table II provides an overview of the evolution of the test development process, including details on the number of items administered, testing durations, participant demographics, and the number of items excluded or added following Rasch analysis. The English translation of the test instrument (which has not been validated) as well as the removed items—including a table listing their Infit and Outfit MNSQ values, correlations, and reasons for removal—are provided in Supplemental Material [91].

Additionally, to support the validation process, the authors and external professionals were asked to classify the items of the final test version into the predefined concept areas and CT categories. Two external professionals assigned the items into the five conceptual areas (the Earth's atmosphere, climate as a system, the carbon cycle, weather vs climate, and the greenhouse effect). Two other professionals categorized the items according to the CT skills (verbal reasoning, argument analysis, hypothesis testing, likelihood and uncertainty analysis, and problem solving and decision making). Criteria for each category were established based on the literature presented in Sec. II to ensure consistency and validity in the expert ratings.

TABLE II. Overview of the test development process.

Test version	Number of items	Time of testing	Number of participants	Type of participants	Items excluded after Rasch analysis	New items added after Rasch analysis
CTCC 1	41 in total; two booklets with 26 items each, with 11 anchor items	From June to December 2023	289	191 university students from Austria, Belgium, and Germany; 98 high school students from Austria	14	5
CTCC 2	32 in total; two booklets with 22 items each, with 12 anchor items	From February to July 2024	340	244 high school students and 96 university students from Germany and Austria	5	0
CTCC 3	27 items in one booklet	From November 2024 to March 2025	180	116 high school graduates and 64 first-semester university students from Austria and Germany	3	0

V. RESULTS AND DISCUSSION

A. Interviews

For the analysis of the interviews, specific categories were defined with the aim of making the students' ideas and thinking patterns visible. It became clear that both university and high school students have difficulties with CT in the context of climate change, especially when it comes to evaluating or formulating a conclusion based on the given text. Henle [92] already discusses this difficulty when respondents add their own opinions and beliefs about the task instead of deciding if the conclusion follows logically from given statements. Participants tend to fail at accepting the logical task and want to introduce their own personal logic. Given statements were changed in such a way that they could draw their own conclusion and integrate their individual prior knowledge, regardless of whether it is correct or incorrect. Consequently, distractors that emerged from the interviews are answers that contain phrases such as "because I know that," "because I am convinced that," or "because I believe that."

Another finding from the interviews is that students face difficulties with hypothesis testing, especially when it involves controlling variables, which has been documented in the literature [93]. This issue became evident when looking at the experiment on sea level rise, where different jars with different ice cubes are provided. Both university and high school students preferred to use different amounts of water with different volumes of ice cubes "because you have to make all kinds of comparisons." However, the fact that comparability is only established through controlling variables was not noticed.

In addition, common student difficulties with climate change already discussed in Sec. II were found in the interviews. For example, most of the interviewed students

were not aware that water vapor is a greenhouse gas as well [16,55–61]. Students also had difficulties describing the greenhouse effect and considered the IR-radiation from the Sun as the cause of global warming [16,56–59,61,63]. Furthermore, those students also expressed the idea of reflection of the incoming IR-radiation on the Earth's surface that can no longer pass through to space because of the blocking greenhouse gases. That is why water vapor is also used as a greenhouse gas in test items instead of the more familiar CO₂ or why the idea of the reflection of the IR-radiation is tested as a hypothesis in one item, to provide realistic distractors.

As the main aim of the interviews for the test was merely to collect realistic distractors, the analysis of the interviews will not be discussed further at this point. An in-depth analysis of the interviews and a detailed description of the developed category system is currently in preparation and will be published in a subsequent paper.

B. Test analysis

In this section, we present the results of Rasch analysis of the final version of the German CTCC test administered to 180 students. The principles of Rasch analysis are described in more detail in Sec. III. After Rasch analysis, 17 misfitting persons were removed from the sample as described in Sec. IV. For clarity, the results are expressed in logits, with each correct answer scored as 1 and each incorrect answer as 0. Main test statistics are presented in Table III, while the whole item statistics of the German version of the CTCC test and the English translation of the test instrument are accessible in Supplemental Material [91]. The test functioning was good, with an item reliability of 0.96, a person reliability of 0.80, and a Cronbach's α at 0.82 [77].

TABLE III. Overview of the key metrics of Rasch modeling for the final *CTCC* test version with 24 items.

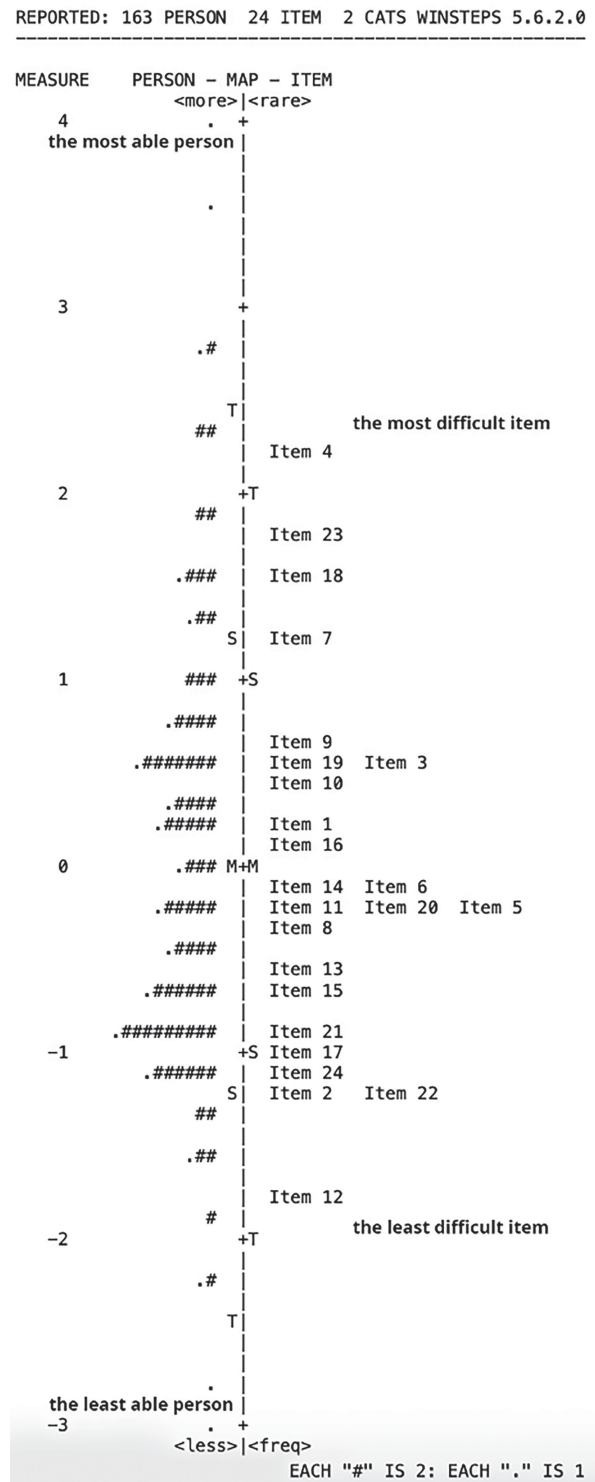
Metric	Value
Item reliability	0.96
Person reliability	0.80
Cronbach's α	0.82
Outfit MNSQ	0.75–1.18
Infit MNSQ	0.81–1.16

The distribution of item difficulties and person abilities is shown in the Wright map (item-person map) in Fig. 1. The Wright map shows the visual distribution of items and the distribution of students' abilities on the same scale. The vertical line in the middle represents the underlying variable in the test, which is students' critical thinking skills in the context of climate change. The item-person map shows good targeting of the test on the sample. The width of the test is appropriate, covering almost all respondents with a good distribution of items according to difficulty. There are five respondents with lower ability than the difficulty of the easiest item, and seven respondents with high ability with no items matching their ability. However, only two of those students have no items in ± 1 logit range of their ability estimate. Since the test is already quite long and has good targeting on the sample, it was decided not to add more difficult items to the test.

The fit of the items to the model is very good, with Outfit values ranging from 0.75 to 1.18, and Infit values ranging from 0.81 to 1.16. Infit values are also shown in the bubble chart in Fig. 2. The bubble chart visually represents the fit of the items, where the items that are closer to the central line in the figure are closer to the expected fit value of 1. The uncertainty of the items is represented with the size of the bubbles. All point-measure correlations were also inspected and are in acceptable range with values 0.34 or higher.

The empirical one-dimensionality of the construct was inspected through the analysis of point-measure correlations and item fit, where all values indicate that the test can be considered as unidimensional. In addition, the principal component analysis of residuals (PCA-R) was conducted in order to investigate whether there are secondary dimensions in the data. For data not having any secondary dimensions, the random size of the largest eigenvalue is almost always < 2.0 [90,94]. In our case, it was 2.1. For this reason, disattenuated correlations were also calculated, in a way that for each contrast, each person was measured on each cluster of items. These measures were then correlated for each pair of clusters. All correlations were above 0.82, indicating that there is no second dimension in the test [90].

Based on the presented analysis, the first research questions can be answered. The results of Rasch analysis suggest that the critical thinking in the context of climate change can be measured with the chosen item set in the

FIG. 1. Wright map of the final *CTCC* test version with 24 items, to show the distribution of person abilities (on the left) and item difficulties (on the right).

German version of the *CTCC* test, as it has succeeded in creating a sufficiently coherent and unidimensional construct, supported by evidence of high reliability, good model fit, and correlations.

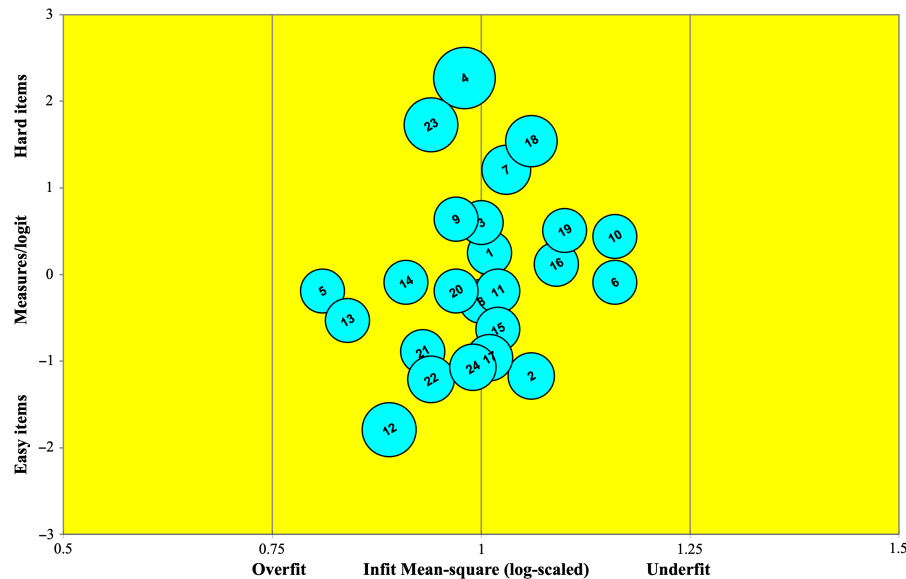


FIG. 2. Bubble chart of the *CTCC* item difficulties vs Infit MNSQ values in logits. Each item is represented by a circle whose size is proportional to its calibration uncertainty (standard error). The vertical axis shows the item difficulty scale in logits. The horizontal axis shows the Infit mean square (MNSQ) values; the central line represents the Infit model value of 1.

The test instrument was administered only in German in the second and third test runs, allowing only the functionality of the German version of the *CTCC* test to be examined.

C. Difficulties of different critical thinking skills and climate change concept areas

As Halpern has explained, the five critical thinking skills should not be seen as independent dimensions [31]. Especially in everyday situations, different or all skill sets are required at the same time depending on the task. However, in the following analysis, we examine whether the five different CT skills differ in difficulty, as expected from the theoretical construct. To do this, the *CTCC* items were assigned to the five skill categories. Additionally, the items were grouped according to the five different concept areas of climate change.

To evaluate the alignment of the items to different CT skills and concept areas, two external professionals assigned the 24 items from the last version of the test to the defined CT skills, while two other external professionals assigned them to defined concept areas of climate change. The results show that, on average, 77% of the 24 items were correctly assigned to the concept areas of climate change, while 71% were correctly assigned to the CT skills. Specifically, 15 items had 100% alignment with the concept areas of climate change, and 13 items had 100% alignment with the CT skills for both professionals. Two items were reclassified to different topic areas or CT skills based on the experts' consensus on alternative categorization. After the review and adjustments, the experts reached full agreement of 78% for the concept

areas of climate change and 77% for the CT skills. The discrepancies were discussed in the research group and additional ratings for those items were performed by experts within the research group. As stated above, multiple skills are often required simultaneously, presenting a challenge to assign the items to a dedicated CT skill or to consider specific skills in isolation. Therefore, we propose not to speak of testing the CT skills directly, but to identify item groups, each group predominantly testing one skill. The same is true for the topics of climate change, which are often interrelated to each other.

To inspect the relative difficulties of different CT skills, the average difficulties for each of five item groups each predominantly testing one of the CT skills were calculated and are presented in Fig. 3. It is important to note that the distribution of difficulties relates to the current level of skills and knowledge of the last sample. The results illustrate differences in the relative difficulty of the item groups. As Winsteps sets the average item difficulty to zero, the skills with negative values are easier than the test average, while the groups with positive values are harder than the test average. The level of dispersion of item difficulties from the average difficulty is represented with error bars. The graph shows the tendency that different item groups corresponding to the respective skill have different difficulties, although a clear separation of difficulties is not possible through high standard errors.

As shown in Fig. 3, the item group clustered as predominantly testing problem solving and decision making skills seems the easiest for participants, which is quite surprising at first. Normally, problem solving and decision making are expected to be the most difficult skill because the integration of all CT skills is necessary to effectively

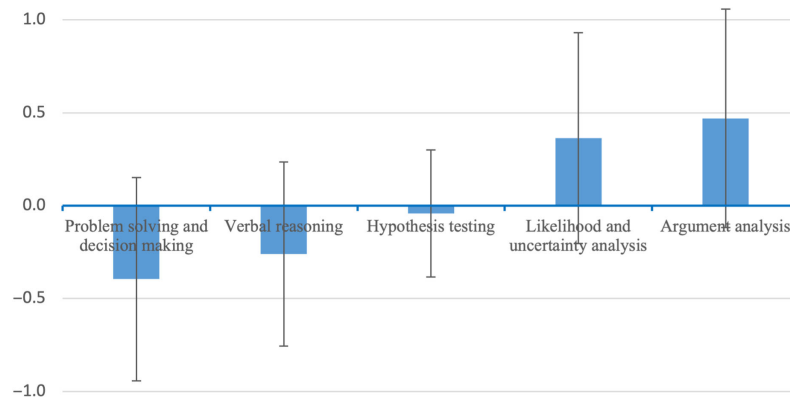


FIG. 3. Average difficulties of five item groups testing five CT skills from easiest to most difficult.

address complex problems, generate solutions, and use creative thinking. In multiple-choice format, most of these steps were taken away from the students, as they were offered different solutions they needed to choose from. In this form, the items likely assessed the evaluation of given solutions rather than the generation and processing of problem solving. This is a limitation of the multiple-choice test when measuring that skill. Since students did not need to define the problem and generate their own solutions but only choose from a range of alternatives given, their thought process is simplified, which makes this ability appear easier than it actually is. For that reason, the *CTCC* test addresses the evaluation of solutions as a subskill of problem solving.

The second easiest item group appears to be verbal reasoning. Here students only have to carefully read the information given and pay attention to misleading language. In contrast, the group of items that predominantly tests argument analysis seems to be the most difficult for the students, which depends on the fact that the students have to analyze the given arguments and evaluate the conclusions drawn. This involves content knowledge and is exactly the skill that is necessary when analyzing social

media posts and statements in newspapers. Furthermore, it is understandable that items testing mainly likelihood and uncertainty analysis skills are difficult because these skills involve statistical understanding, which high school students struggle with and which needs to be fostered [95,96]. The group of items that refer mainly to the hypothesis testing skill ranks in the middle, which could indicate that students are used to dealing with experiments, especially in science courses. Nevertheless, the difficulties of variable control known from literature are apparent here [93].

In addition, the average difficulty of the five item groups each dedicated to one climate change concept area was calculated and is presented in Fig. 4. Again, the standard deviation of the average difficulties is shown by the error bars.

Figure 4 suggests that carbon cycle and weather vs climate are the easiest areas. That weather vs climate appears to be one of the easiest for the students may stem from the exposure students have to it in everyday life. Weather is a constant part of our day-to-day experience indicating some familiarity. Still, students confuse the terms “weather” and “climate” or use the terms synonymously. The greenhouse effect and facts about the atmosphere are the most difficult areas for the

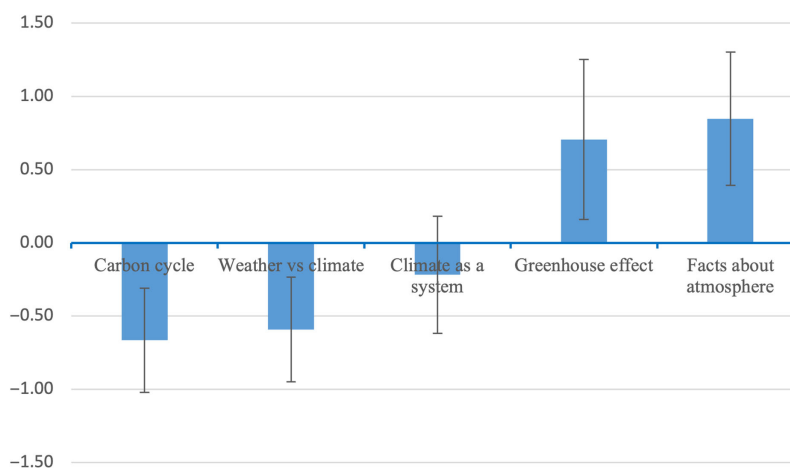


FIG. 4. Average difficulties of five item groups testing five climate change concept areas from easiest to most difficult.

participants. Especially in this area, typical students' difficulties are observed as outlined in Sec. II, such as the difficulty to distinguish between the natural and anthropogenic greenhouse effect. Climate as a system lies in the middle but is still perceived as slightly easier on average. A similar distribution of average difficulties can be seen using the CCCI-422 test instrument [52], with weather vs climate as the easiest and greenhouse effect and facts about the atmosphere as the most difficult area. As a possible explanation, limited integration of the five concept areas and their implementation in educational settings is mentioned. It is likely that improved instruction of the areas might result in a realignment of the average difficulties.

In addition, an analysis of students' answers reveals specific difficulties. The findings coincide with the results from the interviews. For example, students find it difficult when asked to evaluate statements or conclusions based solely on the information provided in the text, instead they use their own knowledge of beliefs for making conclusions (e.g., for item 4 and item 13). These items asked students to evaluate the conclusion based only on the text or to draw a conclusion from the text and these words were both bold and underlined. This is particularly evident when choosing distractors that include "because I know" in the reasoning (e.g., item 9, item 21). For item 9, 22% of the students of the sample chose the distractor "because I know," which is the second most frequently chosen answer after the correct answer (selected by 37%). For item 21, 15% of students chose the answer with the reasoning "because I know," while 67% chose the correct answer.

Additionally, students face problems with text comprehension or with understanding a written hypothesis and selecting an experiment to test it (item 6 and item 9). As known from literature [93], difficulties with variable control or when data must be compared to draw generalizations occur. Students are not familiar with the difference between correlation and causality (item 18) and have difficulties when reading given graphs, which is also reported in the literature [97].

When looking at content knowledge, participants struggle to differentiate between CO₂ emissions and CO₂ concentration in the atmosphere (item 7), they are unaware of what the percentages indicate in weather forecasts (item 15) or do not identify water vapor as a greenhouse gas (item 10 and item 11). Regarding item 10, 41% of the participants chose the correct answer, while 32% selected the distractor claiming that water vapor is not a greenhouse gas. Similarly, for item 11, 23% answered that water vapor is not a greenhouse gas compared to 53% choosing the correct answer.

Students' difficulties were particularly reflected in item 6 when the students from the sample chose the distractor that refers to the reflection of incoming infrared (IR) radiation from the Sun by the greenhouse gases in the atmosphere. Again, similarities to the findings from the CCCI-422 [53]

are evident, with 32% of the participants of this study selecting the distractor that pertains to the concept of reflection of solar radiation on the Earth's surface back to the atmosphere. It is the most chosen answer. Notably, no specific reference is made to IR-radiation from the Sun, only to solar radiation in general. The deliberate emphasis on IR-radiation in our test instrument arises from repeated statements throughout the interviews conducted with high school and university students.

Regarding the second and third research questions, the results suggest that students in the third survey struggle most with item groups predominantly testing argument analysis, followed by likelihood and uncertainty analysis. Students' difficulties with CT include not referring to information in the text and using their own beliefs or knowledge to make conclusions, struggling with testing a hypothesis and control of variables as well as not differentiating between correlation and cause.

VI. LIMITATIONS

Although we carefully followed all different steps for test design, we should remark that the process and result have some limitations.

Regarding the test sampling, all surveys used convenience sampling, and in the first and second test runs, students from various semesters and high school grades were mixed. While it is important to have a large sample with diverse cognitive levels when developing a test, meaningful results in further research such as studying student difficulties depend on a coherent sample. In the third survey, the focus was on high school graduates and first-semester university students taking the *CTCC* test during class to include both motivated and less interested students within a similar age group. The surveys indicate that this target group is particularly appropriate for the test, considering its length and text-heavy nature.

This brings us to another consideration: the requirement for participants to engage with longer items. Unlike content knowledge tests, which often use shorter formats to assess facts, evaluating CT skills in climate change necessitates more comprehensive explanations. As Ramadhani *et al.* [98] suggest, there is likely a correlation between reading comprehension and CT skills, which educators at both the high school and university levels should consider when administering the test.

Moreover, the choice of a closed test format has faced criticism for assessing CT skills [99]. Particularly in evaluating problem-solving skills, multiple-choice formats can be limiting. In our case, we did not manage to evaluate the problem-solving skill as described by Halpern [19,23], but as stated in the discussion section, we have addressed the skill of evaluation of solutions as a subskill of problem solving. This directly challenged our decision to use a multiple-choice format. However, that format was initially selected for its practical advantages in administering tests to

large samples, saving time on evaluation, and enhancing result comparability [25].

VII. CONCLUSION

The CTCC test is a new multiple-choice diagnostic instrument for assessing critical thinking skills in the context of climate change of high school graduates and first-semester university students. The test was constructed and the German version of the test was evaluated using the Rasch model. The results show good functionality and structure with high item and person reliability and good item fit and correlations. The German version of the CTCC test can be used to measure domain-specific CT skills in climate change. In addition, it can help educators assess with which group of items, each testing predominantly one of the CT skills, students have the most difficulties, providing them with information about where students may need further support or development. The results suggest that students have difficulties with argument analysis, which is a crucial skill when assessing social media posts and media reports. This skill therefore should be further targeted in high school and university courses, in physics instruction as well as in other disciplines. The English version of the CTCC test has not yet been evaluated, and further validation with English-speaking students is required.

In future research, the CTCC test can serve as a pre-post assessment tool for evaluating the impact of instructional interventions designed to enhance CT skills in the context

of climate change. By using the test before and after such interventions, researchers can measure the growth in students' CT skills, offering evidence on the effectiveness of their teaching strategies.

Future research should explore its application across diverse educational settings and populations, as well as investigate a further reduction of items to decrease administration time and enhance its practicality for implementation in school. In addition, open-ended questions for assessing problem solving should be developed and implemented in order to test the full set of problem-solving skills.

ACKNOWLEDGMENTS

This study was co-funded by the European Union under Project No. 2022-1-DE01-KA220-HED-0089246 within the Erasmus+ Program KA2, Cooperation Partnerships in Higher Education.

The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union or the National Agency DAAD. Neither the European Union nor the granting authority can be held responsible for them.

DATA AVAILABILITY

The data are not publicly available. The data are available from the authors upon reasonable request.

-
- [1] Intergovernmental Panel on Climate Change (IPCC), Climate change 2022—Impacts, in *Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, 1st ed. (Cambridge University Press, Cambridge, England, 2023), [10.1017/9781009325844](https://doi.org/10.1017/9781009325844).
 - [2] K. M. d'I. Treen, H. T. P. Williams, and S. J. O'Neill, Online misinformation about climate change, *WIREs Clim. Change* **11**, e665 (2020).
 - [3] D. Ding, E. W. Maibach, X. Zhao, C. Roser-Renouf, and A. Leiserowitz, Support for climate policy and societal action are linked to perceptions about scientific agreement, *Nat. Clim. Change* **1**, 462 (2011).
 - [4] J. A. Harvey *et al.*, Internet blogs, polar bears, and climate-change denial by proxy, *BioScience* **68**, 281 (2018).
 - [5] H. T. P. Williams, J. R. McMurray, T. Kurz, and F. Hugo Lambert, Network analysis reveals open forums and echo chambers in social media discussions of climate change, *Global Environ. Change Rep.* **32**, 126 (2015).
 - [6] T. Dingler, B. Tag, and A. Vargo, *Technologies to Support Critical Thinking in an Age of Misinformation* (Dagstuhl Seminar 22172) (Leibniz-Zentrum für Informatik, Schloss Dagstuhl, 2022), [10.4230/DAGREP.12.4.72](https://doi.org/10.4230/DAGREP.12.4.72).
 - [7] J. Cook, P. Ellerton, and D. Kinkead, Deconstructing climate misinformation to identify reasoning errors, *Environ. Res. Lett.* **13**, 024018 (2018).
 - [8] USGCRP, *Climate Literacy: The Essential Principles of Climate Science* (United States Global Change Research Program, Washington, DC, 2009).
 - [9] P. A. Facione, *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction (The Delphi Report)*, American Psychological Association (California Academic Press, Millbrae, CA, 1990).
 - [10] Council of the European Union, *Council Recommendation on Learning for the Green Transition and Sustainable Development*, Official Journal of the European Union (Council of the European Union, Brussels, 2022).
 - [11] Education, Audiovisual, and Culture Executive Agency, *Citizenship Education at School in Europe* (Publications Office, Luxembourg, 2017), [10.2797/612559](https://doi.org/10.2797/612559).

- [12] R. F. Yuretich, Encouraging critical thinking: Measuring skills in large introductory science classes, *J. Coll. Sci. Teach.* **33**, 40 (2003), <https://www.jstor.org/stable/26491263>.
- [13] Global Education Monitoring Report Team, Climate change communication and education country profiles: Approaches to greening education around the world, GEM Report No. 0000387867, UNESCO, 2023, [10.54676/XBVG6945](https://doi.org/10.54676/XBVG6945).
- [14] H. L. Bouygues, Teaching critical thinking in K-12: When there's a will but not always a way, 2022, https://reboot-foundation.org/wp-content/uploads/2022/07/Reboot-White-Paper_NAEP-5.pdf [accessed March 31, 2025].
- [15] K. L. Flores, G. S. Matkin, M. E. Burbach, C. E. Quinn, and H. Harding, Deficient critical thinking skills among college graduates: Implications for leadership, *Educ. Philos. Theory* **44**, 212 (2012).
- [16] L. Jarrett and G. Takacs, Secondary students' ideas about scientific concepts underlying climate change, *Environ. Educ. Res.* **26**, 400 (2020).
- [17] D. Bhattacharya, K. Carroll Steward, and C. T. Forbes, Empirical research on K-16 climate education: A systematic review of the literature, *J. Geosci. Educ.* **69**, 223 (2021).
- [18] L. C. McDermott and P. S. Shaffer, Research as a guide for curriculum development: An example from introductory electricity. Part I: Investigation of student understanding, *Am. J. Phys.* **60**, 994 (1992).
- [19] D. F. Halpern, *Thought and Knowledge: An Introduction to Critical Thinking* (Psychology Press, New York, 2014).
- [20] R. H. Ennis, Critical thinking and the curriculum, in *Thinking Skills Instruction: Concepts and Techniques. Building Students' Thinking Skills Series* (National Education Association, Washington, DC, 1987).
- [21] R. H. Ennis, Critical thinking assessment, *Theory Pract.* **32**, 179 (1993).
- [22] J. E. McPeck, *Critical Thinking and Education*, Issues and Ideas in Education Series (Robertson, Oxford 1981).
- [23] D. F. Halpern, Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and meta-cognitive monitoring, *Am. Psychol.* **53**, 449 (1998).
- [24] D. Moseley *et al.*, *Frameworks for Thinking: A Handbook for Teaching and Learning* (Cambridge University Press, Cambridge, England, 2005).
- [25] R. D. Renaud and H. G. Murray, A comparison of a subject-specific and a general measure of critical thinking, *Thinking Skill Creativity* **3**, 85 (2008).
- [26] The Palgrave handbook of critical thinking in higher education, in *Palgrave Handbooks*, edited by M. Davies and R. Barnett (Palgrave Macmillan, New York, NY, 2015).
- [27] E. R. Lai, *Critical Thinking: A Literature Review* (Pearson, 2011), https://www.researchgate.net/publication/297782058_Critical_thinking_A_literature_review.
- [28] P. Hager and M. Kaye, Critical thinking in teacher education: A process-oriented research agenda, *Aust. J. Teach. Educ.* **17**, 26 (1992).
- [29] C. P. Dwyer, An evaluative review of barriers to critical thinking in educational and real-world settings, *J. Intell.* **11**, 105 (2023).
- [30] M. Lipman, *Thinking in Education*, 2nd ed. Reprinted (Cambridge University Press, Cambridge, England, 2003), [10.1017/CBO9780511840272](https://doi.org/10.1017/CBO9780511840272).
- [31] D. F. Halpern, *Halpern Critical Thinking Assessment Items and Scoring Procedures* (Schuhfried GmbH, Mödling, 2016).
- [32] American Psychological Association, The assessment cyberguide for learning goals and outcomes, 2009, <https://www.apa.org/ed/governance/bea/assessment-cyberguide-v2.pdf> [accessed February 8, 2024].
- [33] F. Sadidi, *Development and Evaluation of a Teaching Unit in Particle Physics to Promote Students' Critical Thinking* (TU Dresden, Dresden, 2022), <https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-849186>.
- [34] D. T. Tiruneh, M. De Cock, A. G. Weldelessie, J. Elen, and R. Janssen, Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism, *Int. J. Sci. Math. Educ.* **15**, 663 (2017).
- [35] J. Sermeus, M. De Cock, and J. Elen, Critical thinking in electricity and magnetism: Assessing and stimulating secondary school students, *Int. J. Sci. Educ.* **43**, 1 (2021).
- [36] P. C. Abrami, R. M. Bernard, E. Borokhovski, D. I. Waddington, C. A. Wade, and T. Persson, Strategies for teaching students to think critically: A meta-analysis, *Rev. Educ. Res.* **85**, 275 (2015).
- [37] R. H. Ennis, Critical thinking and subject specificity: Clarification and needed research, *Educ. Res.* **18**, 4 (1989).
- [38] R. H. Ennis, J. Millman, and T. Tomko, *Cornell Critical Thinking Tests level X & level Z—Manual* (Midwest Publications, Pacific Grove, CA, 1985).
- [39] R. H. Ennis and E. Weir, *The Ennis-Wier Critical Thinking Essay Test: An Instrument for Testing and Teaching* (Pacific Grove, CA, 1985).
- [40] G. Watson and E. Glaser, *Watson-Glaser Critical Thinking Appraisal* (Pearson Assessment, London, England, 2002).
- [41] L. Niu, L. S. Behar-Horenstein, and C. W. Garvan, Do instructional interventions influence college students' critical thinking skills? A meta-analysis, *Educ. Res. Rev.* **9**, 114 (2013).
- [42] D. T. Tiruneh, A. Verburgh, and J. Elen, Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies, *High. Educ. Stud.* **4**, 1 (2014).
- [43] J. H. McMillan, Enhancing college students' critical thinking: A review of studies, *Res. High. Educ.* **26**, 3 (1987).
- [44] M. M.-Y. Yeung, J. W.-M. Yuen, J. M.-T. Chen, and K. K.-L. Lam, The efficacy of team-based learning in developing the generic capability of problem-solving ability and critical thinking skills in nursing education: A systematic review, *Nurse Educ. Today* **122**, 105704 (2023).
- [45] B. Harjo, B. Kartowagiran, and A. Mahmudi, Development of critical thinking skill instruments on mathematical learning high school, *Int. J. Instr.* **12**, 149 (2019).
- [46] I. J. Quitadamo, C. L. Faiola, J. E. Johnson, and M. J. Kurtz, Community-based inquiry improves critical thinking in general education biology, *CBE Life Sci. Educ.* **7**, 327 (2008).
- [47] C. Walsh, K. N. Quinn, C. Wieman, and N. G. Holmes, Quantifying critical thinking: Development and validation

- of the physics lab inventory of critical thinking, *Phys. Rev. Phys. Educ. Res.* **15**, 010135 (2019).
- [48] A. B. Heim, G. Lawrence, R. Agarwal, M. K. Smith, and N. G. Holmes, Perceptions of interdisciplinary critical thinking among biology and physics undergraduates, *Phys. Rev. Phys. Educ. Res.* **21**, 010138 (2025).
- [49] D. T. Tiruneh, X. Gu, M. De Cock, and J. Elen, Systematic design of domain-specific instruction on near and far transfer of critical thinking skills, *Int. J. Educ. Res.* **87**, 1 (2018).
- [50] J. C. Bonk and S. G. Smith, Alternative instructional strategies for creative and critical thinking in the accounting curriculum, *J. Acc. Educ.* **16**, 261 (1998).
- [51] S. Liu, G. Roehrig, D. Bhattacharya, and K. Varma, In-service teachers' attitudes, knowledge and classroom teaching of global climate change, *Sci. Educ.* **24**, 12 (2015), <https://files.eric.ed.gov/fulltext/EJ1069990.pdf>.
- [52] T. Schubatzky *et al.*, Entwicklung des concept-inventory CCCI-422 zu den naturwissenschaftlichen Grundlagen des Klimawandels, *Z. Didakt. Naturwiss.* **29**, 10 (2023).
- [53] T. Schubatzky, R. Wackermann, C. Haagen-Schützenhöfer, and C. Wöhlke, How well do German a-level students understand the scientific underpinnings of climate change?, *Sustainability* **16**, 7264 (2024).
- [54] S. Wildbichler, C. Haagen-Schützenhöfer, and T. Schubatzky, Students' ideas about the scientific underpinnings of climate change: A systematic review of the literature, *Stud. Sci. Educ.* **61**, 117 (2025).
- [55] A. M. Bodzin, D. Anastasio, D. Sahagian, T. Peffer, C. Dempsey, and R. Steelman, Investigating climate change understandings of urban middle-level students, *J. Geosci. Educ.* **62**, 417 (2014).
- [56] A. N. Versprille and M. H. Towns, General chemistry students' understanding of climate change and the chemistry related to climate change, *J. Chem. Educ.* **92**, 603 (2015).
- [57] A. Versprille, A. Zabih, T. A. Holme, L. McKenzie, P. Mahaffy, B. Martin, and M. Towns, Assessing student knowledge of chemistry and climate science concepts associated with climate change: Resources to inform teaching and learning, *J. Chem. Educ.* **94**, 407 (2017).
- [58] V. Dawson and K. Carson, Australian secondary school students' understanding of climate change, *Teach. Sci.* **59**, 9 (2013).
- [59] D. P. Shepardson, D. Niyogi, A. Roychoudhury, and A. Hirsch, Conceptualizing climate change in the context of a climate system: Implications for climate and environmental education, *Environ. Educ. Res.* **18**, 323 (2012).
- [60] E. Boyes and M. Stanisstreet, The "greenhouse effect": Children's perceptions of causes, consequences and cures, *Int. J. Sci. Educ.* **15**, 531 (1993).
- [61] J. L. McCuin, K. Hayhoe, and D. Hayhoe, Comparing the effects of traditional vs. misconceptions-based instruction on student understanding of the greenhouse effect, *J. Geosci. Educ.* **62**, 445 (2014).
- [62] D. Pruneau, H. Gravel, W. Bourque, and J. Langis, Experimentation with a socio-constructivist process for climate change education, *Environ. Educ. Res.* **9**, 429 (2003).
- [63] D. P. Shepardson, D. Niyogi, S. Choi, and U. Charusombat, Students' conceptions about the greenhouse effect, global warming, and climate change, *Clim. Change* **104**, 481 (2011).
- [64] D. Lombardi, G. M. Sinatra, and E. M. Nussbaum, Plausibility reappraisals and shifts in middle school students' climate change conceptions, *Learn. Instr.* **27**, 50 (2013).
- [65] R. Bellamy, Public perceptions of climate tipping points, *Publ. Understand. Sci.* **32**, 1033 (2023).
- [66] M. S. McCaffrey and S. M. Buhr, Clarifying climate confusion: Addressing systemic holes, cognitive gaps, and misconceptions through climate literacy, *Phys. Geogr.* **29**, 512 (2008).
- [67] L. Zangori, A. Peel, A. Kinslow, P. Friedrichsen, and T. D. Sadler, Student development of model-based reasoning about carbon cycling and climate change in a socio-scientific issues unit, *J. Res. Sci. Teach.* **54**, 1249 (2017).
- [68] K. I. Danielson and K. D. Tanner, Investigating undergraduate science students' conceptions and misconceptions of ocean acidification, *CBE Life Sci. Educ.* **14**, ar29 (2015).
- [69] V. Papadimitriou, Prospective primary teachers' understanding of climate change, greenhouse effect, and ozone layer depletion, *J. Sci. Educ. Technol.* **13**, 299 (2004).
- [70] D. Lombardi and G. M. Sinatra, College students' perceptions about the plausibility of human-induced climate change, *Res. Sci. Educ.* **42**, 201 (2012).
- [71] D. Read, A. Bostrom, M. G. Morgan, B. Fischhoff, and T. Smuts, What do people know about global climate change? 2. Survey studies of educated lay people, *Risk Anal.* **14**, 971 (1994).
- [72] R. M. V. Gowda, J. C. Fox, and R. D. Magelky, Students' understanding of climate change: Insights for scientists and educators, *Bull. Am. Meteorol. Soc.* **78**, 2232 (1997).
- [73] *Schülvorstellungen und Physikunterricht: Ein Lehrbuch für Studium, Referendariat und Unterrichtspraxis*, edited by H. Schecker, T. Wilhelm, M. Hopf, and R. Duit (Springer, Berlin, Heidelberg, 2018), 10.1007/978-3-662-57270-2.
- [74] S. Wildbichler *et al.*, Tutorials in climate change, *Phys. Teach.* **63**, 414 (2025).
- [75] X. Liu and W. J. Boone, *Applications of Rasch Measurement in Science Education* (JAM Press, Maple Grove, MN, 2006).
- [76] X. Liu, *Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach* (Information Age Publishing, Charlotte, NC, 2010).
- [77] W. J. Boone, J. R. Staver, and M. S. Yale, *Rasch Analysis in the Human Sciences* (Springer, Dordrecht, 2014), 10.1007/978-94-007-6857-4.
- [78] T. G. Bond and C. M. Fox, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3rd ed. (Routledge, Taylor & Francis Group, New York, London, 2015).
- [79] M. Paninic, W. J. Boone, A. Susac, and L. Ivanjek, Rasch analysis in physics education research: Why measurement matters, *Phys. Rev. Phys. Educ. Res.* **15**, 020111 (2019).
- [80] L. Ivanjek, L. Morris, T. Schubatzky, M. Hopf, Jan-Philipp Burde, C. Haagen-Schützenhöfer, L. Dopatka, V. Spatz, and T. Wilhelm, Development of a two-tier instrument on simple electric circuits, *Phys. Rev. Phys. Educ. Res.* **17**, 020123 (2021).

- [81] K. Matejak Cvenic, M. Planinic, A. Susac, L. Ivanjek, K. Jelacic, and M. Hopf, Development and validation of the conceptual survey on wave optics, *Phys. Rev. Phys. Educ. Res.* **18**, 010103 (2022).
- [82] V. Mešić *et al.*, Measuring students' conceptual understanding of wave optics: A Rasch modeling approach, *Phys. Rev. Phys. Educ. Res.* **15**, 010115 (2019).
- [83] I. Neumann, K. Neumann, and R. Nehm, Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test, *Int. J. Sci. Educ.* **33**, 1373 (2011).
- [84] R. Berger, C. Kulgemeyer, and P. Lensing, Ein multiple-choice-test zum konzeptuellen verständnis der kraftwirkung auf ladungsträger in statischen elektrischen und magnetischen Feldern, *Z. Didakt. Naturwiss.* **25**, 197 (2019).
- [85] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (Danmarks Paedagogiske Institut, Copenhagen, 1960).
- [86] A. E. Kazdin, *Research Design in Clinical Psychology*, 5th ed. (Cambridge University Press, Cambridge, England, 2021), [10.1017/9781108993647](https://doi.org/10.1017/9781108993647).
- [87] K. A. Ericsson and H. A. Simon, How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking, *Mind Cult. Act.* **5**, 178 (1998).
- [88] A. Reinhart *et al.*, Think-aloud interviews: A tool for exploring student statistical reasoning, *J. Stat. Data Sci. Educ.* **30**, 100 (2022).
- [89] U. Kuckartz and S. Rädiker, *Qualitative Inhaltsanalyse. Methoden, Praxis, Umsetzung mit Software und künstlicher Intelligenz*, 6th ed. (Juventa Verlag, Weinheim, 2024).
- [90] J. M. Linacre, A user's guide to Winsteps ministeps Rasch-model computer programs, 2025, <https://www.winsteps.com/a/Winsteps-Manual.pdf> [accessed April 11, 2025].
- [91] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/jnbq-1qyk> for the English translation of the CTCC test (with the correct answers marked) and also the removed items throughout the development process—including a table listing their Infit and Outfit MNSQ values, correlations, and reasons for removal.
- [92] M. Henle, On the relation between logic and thinking, *Psychol. Rev.* **69**, 366 (1962).
- [93] M. Schwichow, S. Croker, C. Zimmerman, T. Höffler, and H. Härtig, Teaching the control-of-variables strategy: A meta-analysis, *Dev. Policy Rev.* **39**, 37 (2016).
- [94] J. M. Linacre, Critical eigenvalue sizes (variances) in standardized residual principal components analysis (PCA), <https://www.rasch.org/rmt/rmt191h.htm>.
- [95] S. Mohamad Hasim, R. Rosli, and L. Halim, A quantitative case study of secondary school students' level of statistical thinking, *Eurasia J. Math. Sci. Technol. Educ.* **20**, em2421 (2024).
- [96] M. Marshman and P. K. Dunn, Improving statistical thinking, *Math. Educ. Res. J.* **36**, 1 (2024).
- [97] L. Ivanjek, A. Susac, M. Planinic, A. Andrasevic, and Z. Milin-Sipus, Student reasoning about graphs in different contexts, *Phys. Rev. Phys. Educ. Res.* **12**, 010106 (2016).
- [98] N. Ramadhani, K. Salija, and S. Baa, The correlation between students' critical thinking skills and their reading comprehension at undergraduate study program English Department Faculty of Languages and Literature Universitas Negeri Makassar, *Klasikal J. Educ. Lang. Teach. Sci.* **5**, 25 (2023).
- [99] K. Y. L. Ku, Assessing students' critical thinking performance: Urging for measurements using multi-response format, *Thinking Skills Creating* **4**, 70 (2009).