Social Sciences & Humanities Open 13 (2026) 102750

Contents lists available at ScienceDirect

Social Sciences & Humanities Open

journal homepage: www.sciencedirect.com/journal/social-sciences-and-humanities-open

Regular Article

Assessing scientific thinking: Development and validation of a classroom test battery

Langat Gilbert Cheruiyot *, Gyöngyvér Molnár

Institute of Education, MTA-SZTE Digital Learning Technologies Research Group, University of Szeged, Hungary

ARTICLE INFO

Keywords:
Assessment
Control of variables
Deductive reasoning
Science education
Scientific reasoning
Scientific thinking

ABSTRACT

Scientific thinking is a fundamental pillar of modern science education. However, ensuring the reliability and validity of various assessment tools is a challenging task. Therefore, this study developed, validated, and optimized a test battery for classroom use that evaluates three key dimensions of scientific thinking, namely scientific reasoning (SR), deductive reasoning (DR), and control of variables (CoV). Reliability and construct validity were determined using Cronbach's alpha and confirmatory factor analysis (CFA), respectively. Excellent internal consistency was determined for the revised and optimized versions of the test battery ($\alpha = .873$ for SR, $\alpha = .912$ for DR, and $\alpha = .767$ for CoV). Furthermore, the CFA revealed a good model fit for the three optimized instruments, which was evidence of unidimensionality. This study offers educators, researchers, and policymakers a valuable resource to evaluate and improve scientific thinking in STEM.

1. Introduction

Today, learners face difficulties due to the rapid pace of technological progress and global issues such as climate change and health emergencies. It is no longer sufficient to possess mere knowledge of the topic; instead, students must develop scientific thinking skills to navigate uncertainties, tackle pressing challenges, and drive evidence-based change. This involves developing the ability to ask questions, conduct investigations, analyze information, and reason logically based on available evidence. This set of cognitive skills will enable learners to distinguish relevant information from distractions, turn their questions into significant issues, and formulate responses based on scientific evidence rather than assumptions. However, while educators recognize scientific thinking as a fundamental part of science education (Bao et al., 2022) and an essential skill for the 21st century, there are no reliable and validated instruments to assess it effectively. This represents a critical gap, as precise assessments are necessary to cultivate the vital skills mentioned above.

Although the related terms 'scientific thinking' and 'scientific reasoning' (SR) are often used interchangeably, they are separate cognitive constructs. While both involve hypothesis generation, evidence evaluation, and experiments, among other similar elements, their scope and uses differ significantly between educational and psychological research (Kuhn, 2010). SR is usually narrower in scope and emphasizes logical and systematic processes. It involves formulating and testing hypotheses, controlling variables, and drawing conclusions based on empirical evidence (Pedaste et al., 2015; Klahr, 2000).

Scientific reasoning is typically assessed through tasks that require learners to isolate variables, design fair tests, and apply deductive or inductive logic to experimental results. In addition, SR includes the control of variables (CoV) and deductive reasoning (DR). First, in the CoV strategy, the key idea is that only one variable should be changed at a time to establish causality (Chen & Klahr, 1999). This reasoning is rooted in the hypothetical-deductive model, which sees students intentionally connect theory and evidence to improve their understanding (Kuhn, 2005; Zimmerman, 2007). Second, in DR, students apply general principles to logically derive valid conclusions. On the contrary, CoV involves designing experiments that isolate causal relationships by manipulating one variable at a time, which is considered an essential skill in fair testing and evidence evaluation (Chen & Klahr, 1999; Schwichow et al., 2016). Various methods have been used to measure aspects of SR skills, with Lawson's (1978) classroom test of SR being the most widely used to evaluate reasoning, CoV, and DR.

Compared to SR, scientific thinking is considered a broader construct that engages both scientific content and the use of reasoning processes, such as induction, deduction, and hypothesis testing. These processes reflect general cognitive processes applied in an evidence-based way (Dunbar & Klahr, 2012). Furthermore, Yaşar (2022) defines it as a

* Corresponding author.
E-mail addresses: gilbert.cheruiyot.langat@edu.u-szeged.hu (L.G. Cheruiyot), gymolnar@edpsy.u-szeged.hu (G. Molnár).

https://doi.org/10.1016/j.ssaho.2026.102750
Received 18 October 2025; Received in revised form 28 March 2026; Accepted 30 March 2026
Available online 3 April 2026

mindset grounded in fundamental cognitive processes that are shared with daily thinking and applied more iteratively and cyclically. Therefore, scientific thinking involves problem solving, model building, and coordinating the use of inductive and deductive reasoning, generating tests, and refining explanations of phenomena. It is also referred to as cognitive orientation, characterized by curiosity, a tendency to ask meaningful questions, openness to new evidence, and an overall scientific worldview (Kuhn, 2010). It is often developed in students by encouraging them to observe, ask questions, interpret data, and discuss findings (Zimmerman & Klahr, 2018). Although SR is part of scientific thinking, the latter also includes emotional and motivational elements, such as interest in exploration and persistence in solving problems (Koerber et al., 2015). Therefore, scientific thinking combines cognitive strategies like reasoning, experimentation, and modeling with epistemic beliefs that shape how individuals view scientific knowledge (Zimmerman, 2000) and highlights that SR mainly emphasizes procedural skills, such as forming hypotheses, designing fair tests, interpreting data, and drawing conclusions (Zhou et al., 2016). Furthermore, Hogan and Fisherkeller (2005) noted that SR is not just a scientific ability but also a means of understanding through questioning, investigation, and analysis. This method has become increasingly crucial in an era characterized by complex and often conflicting information. Furthermore, scientific thinking is crucial for lifelong learning, especially for making informed decisions based on evidence in real-world situations. In terms of the differences between the two concepts, Echevarria (2003) described SR as involving hypothesis creation and experimental setup, while Kuhn (2005) described scientific thinking as being based on observation and empirical inquiry. In terms of similarities, both concepts stress the importance of logic, critical analysis, and evidence-based decision-making. However, inconsistent terminology has caused some overlap between the two.

Various studies have explored methods for assessing scientific thinking through performance-based tests (Lazonder & Janssen, 2018; Lazonder, Janssen, Gijlers, & Walraven, 2021) and scale-based assessments (Koerber et al., 2015). However, the most relevant tools tend to have a narrow focus on isolated cognitive skills or be designed for use with specific age groups. (Molnár & Csapó, 2018) introduced a promising computer-based assessment of SR in simulated inquiry tasks; however, their tool does not fully encompass the broader conceptual scope of scientific thinking, which includes components like DR and CoV. Furthermore, comprehensive, psychometrically validated instruments capable of assessing scientific thinking as a unidimensional construct are currently lacking. Many relevant tools fail to integrate different types of reasoning into a coherent framework or lack evidence of reliability and validity in their specific context. This fragmentation highlights the need for a robust and unified assessment tool that captures the full range of reasoning skills essential for scientific thinking. Scientific thinking is conceptualized as a multidimensional construct that encompasses reasoning, hypothesis testing, and evidence-based decision making (Kuhn, 2010; Zimmerman, 2007). Within this framework, scientific reasoning, deductive reasoning, and control of variables represent core cognitive processes that underlie scientific thinking (see Fig. 1). The present test battery was designed to operationalize these theoretical components of scientific thinking. Items assessing scientific reasoning targeted students' ability to interpret data and justify conclusions, deductive reasoning items focused on rule-based inference and logical consistency, and control of variables items were designed based on Chen and Klahr's (1999) framework.

Prior studies have rarely developed classroom assessment instruments that assess scientific thinking. The present study aimed to address this gap by designing and validating a test battery for classroom use that measures SR, DR, and CoV, providing a reliable and versatile tool for educational research and classroom practice. Consequently, the specific research objective of this study was to:

1. Develop a test battery measuring the key components of scientific thinking, including scientific reasoning, deductive reasoning, and control of variables;
2. Examine the psychometric properties of the test battery that include reliability, validity, and objectively assess the scientific thinking of junior high school students in a classroom environment;
3. Evaluate the suitability of the test battery for classroom use and provide immediate feedback to students.

Furthermore, this study addressed the following research questions (RQs):

RQ1. What are the psychometric properties, reliability, and construct validity of three instruments for assessing SR, DR, and CoV? Are the tests sufficiently unidimensional for their final scores to be used as achievement indicators that describe the skills of SR, DR, and CoV of learners?

RQ2. To what extent do SR, DR, and CoV function as latent variables that represent key dimensions of scientific thinking? Is the test battery appropriate to assess the scientific thinking of learners in the classroom with immediate feedback?

2. Review of the literature

2.1. Scientific reasoning and its assessment

SR plays a crucial role in the field of science education. Emphasis is increasingly placed on developing students' reasoning skills, which are essential for acquiring a deep understanding of scientific concepts through inquiry. Such skills not only help learners acquire scientific knowledge but also enable them to actively participate in scientific investigations (Lawson, 2005; Tytler & Peterson, 2003). SR encompasses various cognitive processes, primarily of a hypothetical-deductive nature, including CoV, probabilistic reasoning, and correlational thinking (Lawson, 2005). Therefore, DR and CoV are fundamental components of SR. Defining SR as thinking with and about scientific knowledge, Hogan and Fisherkeller (2005) emphasized its reflective and interpretative nature. The Lawson Classroom Test of SR (1978), one of the most widely used tools for assessing SR, evaluates key reasoning skills, including CoV, proportional reasoning, and probabilistic reasoning (Bao et al., 2022). In light of this, this study adopted the principles underlying the

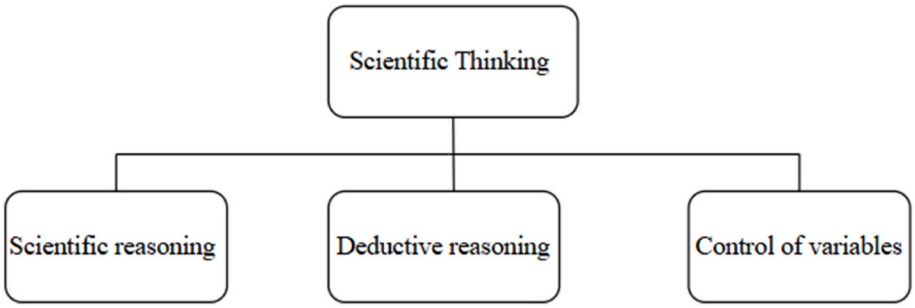

Fig. 1. Conceptual model of scientific thinking and its core components.

Lawson test to develop a context-specific SR assessment tailored to the Kenyan educational environment to accurately assess students' SR levels.

2.2. Deductive reasoning and its assessment

DR is a fundamental part of logical thinking that involves the ability to draw valid conclusions from established premises; if the premises are true, then the conclusion must also be true (Evans & St, 2005). Rooted in formal logic, DR is widely regarded as a key indicator of rational thought and a fundamental component of SR. DR can generally be divided into the following two main types: conditional reasoning and syllogistic reasoning (Sternberg, 2012), while assessing DR often involves tasks such as syllogistic reasoning, propositional logic problems, and the Wason selection task. Syllogistic reasoning, for example, requires individuals to judge the validity of arguments based on categorical premises (e.g., all A are B, and all B are C; therefore, all A are C; Evans & St, 2005). DR works by applying formal logical rules to known premises, thereby enabling individuals to reach conclusions without further empirical testing (Ayalon & Even, 2010). While conditional reasoning involves using 'if-then' statements to solve problems logically, syllogistic reasoning uses two or more premises to conclude. DR draws specific conclusions from general principles or known facts, with each premise serving as direct evidence to support the conclusion (Radulović & Stojanović, 2018).

In addition, various methods have been developed to assess DR in different age groups and contexts. For example, Carreira et al. (2020) used analytical reasoning problems to assess deductive abilities, while De Chantal and Markovits (2017) assessed the syllogistic reasoning of preschoolers through an interactive, game-based format. Furthermore, Csapó, Molnár, and Nagy (2014) investigated early DR and school readiness in first-grade learners through a computer-based assessment that involved logical reasoning tasks that required children to draw appropriate conclusions. The study developed deductive reasoning tests based on syllogistic reasoning and conditioning principles while focusing on the Kenyan education system.

2.3. Control of variables and their assessment

CoV essentially means varying one thing at a time (Chen & Klahr, 1999). It extends beyond the design of experiments with controlled conditions to include the ability to distinguish between confounded and unconfounded experiments. This involves drawing valid conclusions from unconfounded experiments and recognizing the uncertainty inherent in conclusions drawn from confounded ones (Chen & Klahr, 1999) (Fig. 2).

Despite its importance, the CoV strategy lacks a clearly defined set of sub-skills or performance standards. Lawson (1978) described it as the isolation of variables, while Millar and Driver (1987) referred to it as eliminating alternative interpretations of a situation. Ross (1988) identified the following four subskills of CoV: (1) recognizing controlled and uncontrolled experiments; (2) fixing uncontrolled experiments; (3) planning controlled experiments; and (4) justifying experimental designs by referencing a general rule. However, these subskills offer an incomplete definition of CoV as they do not cover skills related to interpreting experimental results or understanding the limitations of confounded experiments.

As the CoV strategy allows scientists to analyze cause-and-effect relationships in their experiments, it is a fundamental component of inquiry-based learning in which students conduct their research. In general, CoV is essential for fostering scientific literacy and is closely linked to other educational goals, such as teaching students how to formulate questions and develop scientific arguments (Kuhn, 2005).

3. Methods

This section presents the methods used to conduct the study. First, the participants and study procedures are described in Section 3.1, followed by the concept of operationalizing scientific thinking in Section 3.2. Next, the procedures followed in the development of the instrument are presented in Section 3.3. Lastly, Section 3.4 describes how the data analysis was conducted.

3.1. Participants and study procedures

The target population for this study consisted of junior high school students (Grades 7 to 9) enrolled in a public school that implemented Kenya's competency-based curriculum (CBC). This population was selected because the junior high school represents the first cohort under the CBC framework. The study sample consisted of 200 high school students drawn from public schools implementing a competency-based curriculum (aged 13–14 years; M = 13.50; 98 boys, 102 girls). The students were randomly selected to participate in the study during school hours in the school's computer laboratories. Three computer-based tests were developed to assess scientific reasoning (SR), deductive reasoning (DR), and control of variables (CoV). The tests were administered through the eDia assessment platform (Csapó & Molnár, 2019), and students access the system by logging in with personal codes.

3.2. Construct operationalization of scientific thinking

Although scientific thinking encompasses a wide range of cognitive skills, such as creativity and problem-solving, they were excluded. In the present study, scientific thinking was operationalized through scientific

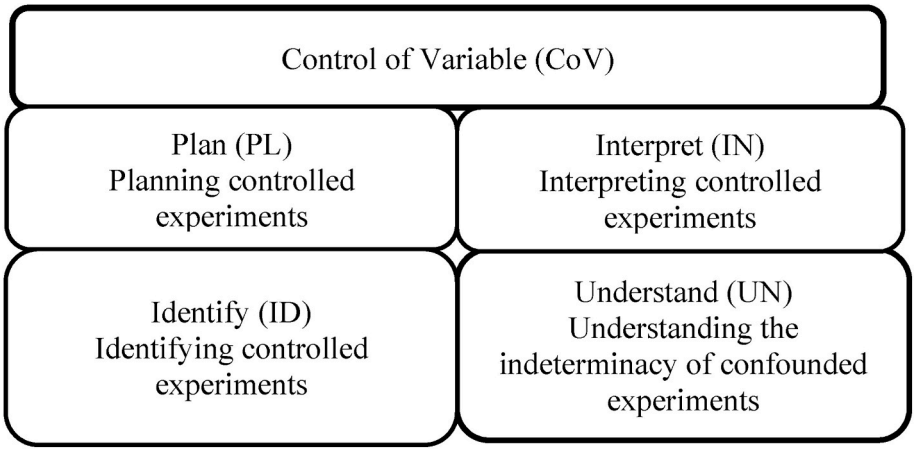

Fig. 2. CoV subskills adapted from Chen and Klahr (1999).

reasoning (SR), deductive reasoning (DR), and control of variables (CoV), which represent the foundational mechanisms that underlie hypothesis testing, evidence evaluation, and causal inference.

Scientific reasoning was operationalized as the ability of the student to interpret evidence, evaluate explanations, and draw conclusions. Deductive reasoning was defined as the ability to apply logical rules and derive a conclusion from a given premise, while control of variables was defined as the ability to plan, identify, interpret, and understand the variables in an experimental context. The test was divided into the following three separate sessions: SR, DR, and CoV.

3.2.1. Instrument development and assessment design

This study was based on evidence-based principles of assessment design to ensure content validity and cognitive alignment when developing and validating the instruments, drawing from the 10-step framework outlined by Althouse (2000). Instruments were developed through a four-step procedure of framework development, item development, content validation, and construct validation. A structured process was followed from the outset, guided by a clear test blueprint derived from curriculum standards. A pool of items was generated that targeted SR, DR, and CoV, and each item was linked to its respective objective and designed to elicit reasoning processes beyond mere recall, emphasizing application, analysis, and logical inference (Krathwohl, 2002).

3.2.2. Validation and test optimization procedures

The validation process began with content validation by external experts using the frameworks (later called 'original tests'). Next, validated SR, DR and CoV tests were administered, and items that affected their homogeneity (e.g., negative or near-zero item-total correlations) were removed (first revision). Using the revised tests, a unidimensional construct validity analysis was performed with first-order confirmatory factor analysis (CFA) to determine whether SR, DR, and CoV were sufficiently unidimensional, that is, whether it would be appropriate to derive fundamentally unidimensional total scores for each. In which the items were analyzed as categorical indicators to determine whether SR, DR, and CoV could be treated as a unidimensional context.

Third, a second-order CFA was conducted to test the suitability of SR, DR, and CoV to measure student scientific thinking, and the tests were further optimized (optimized tests) by excluding items that also measured other constructs. Lastly, external experts again evaluated the final tests to monitor their fit to the original theory.

3.3. Instruments

The instruments developed in this study were designed to assess the SR and DR skills of junior high school students, as well as their ability to control variables. The tests were developed and administered through the eDia platform (Csapó & Molnár, 2019), which facilitates automatic scoring and offers instant feedback after completion.

3.3.1. SR test

The SR test consisted of 20 items that were specifically designed to evaluate SR in the Kenyan educational context. The topics covered included conservation, classification, proportional reasoning, and correlational reasoning. The content was based on fundamental integrated science concepts for secondary school students and was consistent with the Kenyan competency-based curriculum. Some of the test elements were adapted from the assessment of Korom et al. (2017), while others were derived from the principles outlined in the Lawson test (2000). All items were in the format of multiple-choice questions, each with one correct answer and three distractors. Fig. 3 presents a sample item from the SR test developed using the eDia platform.

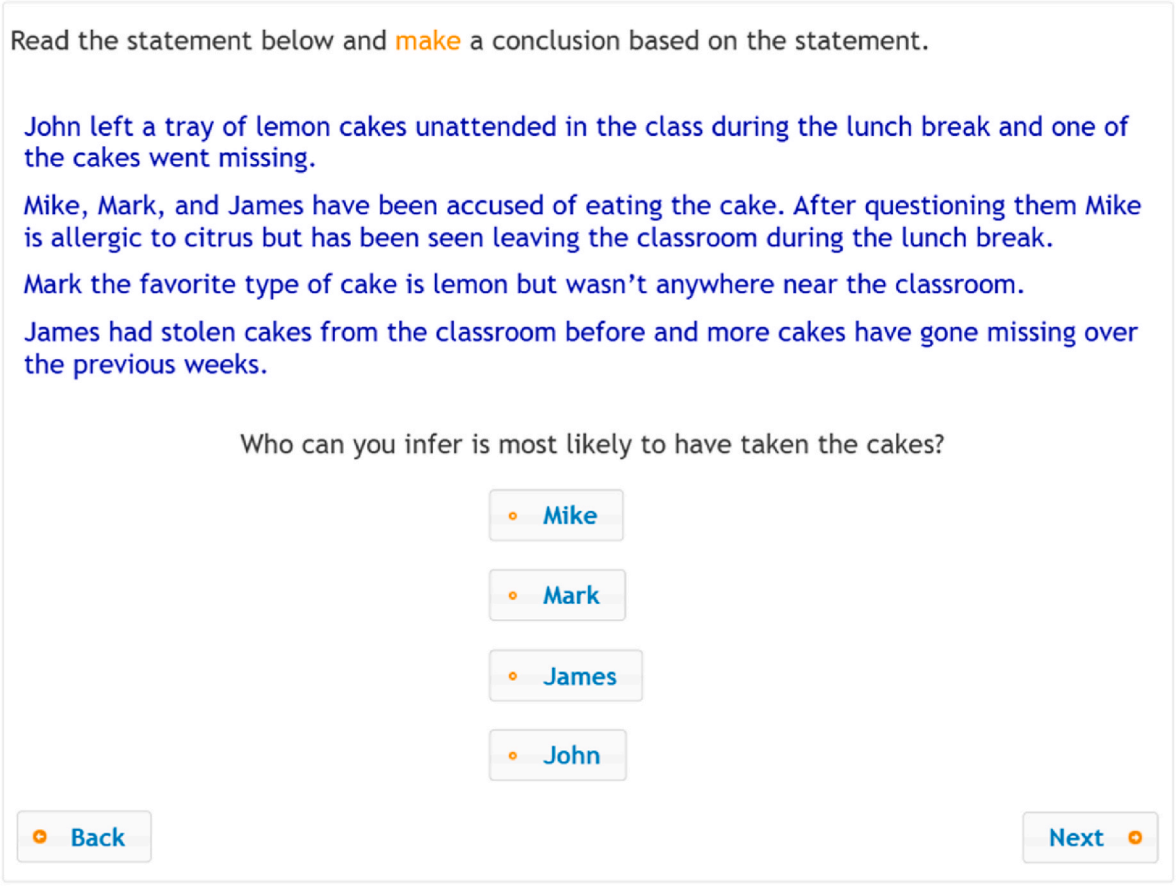

Fig. 3. Sample item from the self-developed scientific reasoning test.

3.3.2. DR test

The DR test comprised 35 items that required test-takers to draw specific conclusions from general premises or statements (Walton, 2008). Elements were designed to target reasoning processes and were based on a syllogistic structure that adhered to specific assessment principles. Each test item consisted of a major premise, a minor premise, and a conclusion derived from them (Khemlani & Johnson-Laird, 2012). Fig. 4 presents the DR test items.

3.3.3. CoV test

The CoV test was designed to be easily administered, scored and analyzed. The elements reflected the real-world context of the Kenyan integrated science curriculum, particularly topics in biology, physics, and chemistry that are appropriate for junior high school. The test consisted of 16 multiple-choice items embedded in an integrated science context at the said level (Fig. 5).

The test assessed the following four core sub-skills (Chen & Klahr, 1999): identifying controlled experiments (ID); interpreting experimental outcomes (IN); understanding the indeterminacy of confounded experiments (UN); and planning controlled experiments (PL). ID items presented a scenario that ended with a hypothesis involving a causal relationship, where students had to select the correct answer from four options. The IN and UN items had a similar structure, as both began with a diagram illustrating an experimental outcome that the students had to interpret; however, the IN items required students to evaluate whether the experimental setup was controlled, while the UN items asked them to determine whether the experiment was confounded, without making causal inferences. Lastly, PL items required students to design an experiment to test a hypothesis by identifying the independent variables.

3.4. Validity of the content

To ensure content validity, a panel of experienced junior high school science teachers with extensive backgrounds in teaching integrated science conducted an expert review of the test items. The review process was guided by the principles underlying the Lawson Classroom Test of Scientific Reasoning, which operationalizes the core components of scientific reasoning, deductive reasoning, and control of variables. The reviewers evaluated the alignment of each item with the targeted cognitive domain, with particular attention to the level of reasoning required, the representation of the intended construct, and the curricular relevance within the integrated science framework. Feedback from the expert panel informed item refinement, including revisions to item wording, cognitive demand, and construct coverage.

For the SR test, the reviewers classified the items according to the following four dimensions: classification, correlation, conservation, and probability CoV test. Similarly, for the CoV test, items were classified according to the following four core skills: identifying variables, manipulating variables, interpreting results, and drawing conclusions. For the three instruments, experts evaluated each item for clarity, relevance, and alignment with the intended construct to minimize the variance relevant to the construct and ensure conceptual precision. Following this validation process, a finalized set of 20 SR items, 35 DR items, and 16 CoV items was prepared for field testing.

3.5. Data analysis and statistical procedures

SPSS 27, Mplus 8.4 and R software application version 4.3.3 were used for data analysis. Cronbach's alpha coefficients were calculated to document internal consistency in all stages of test development. However, Cronbach's alpha is widely not considered a primary measure of reliability for the optimized test in the context of confirmatory factor

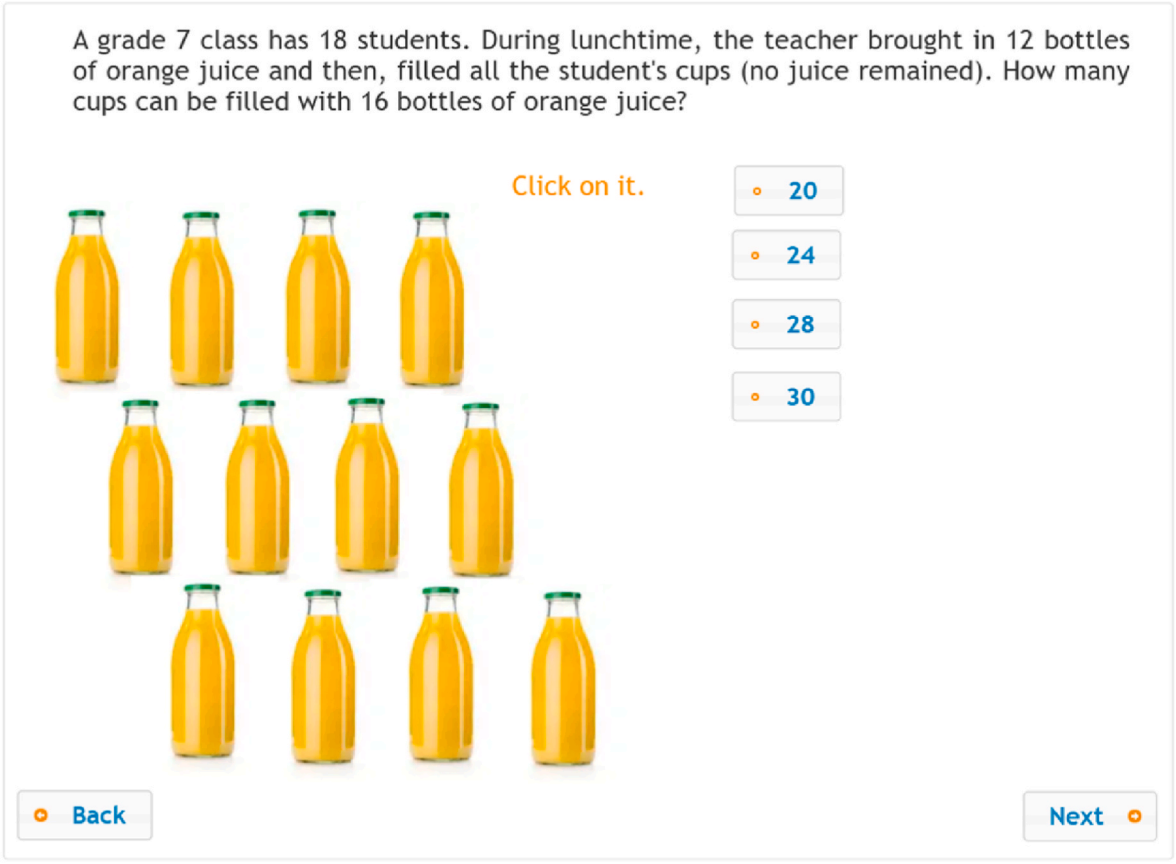

Fig. 4. Sample deductive reasoning item.

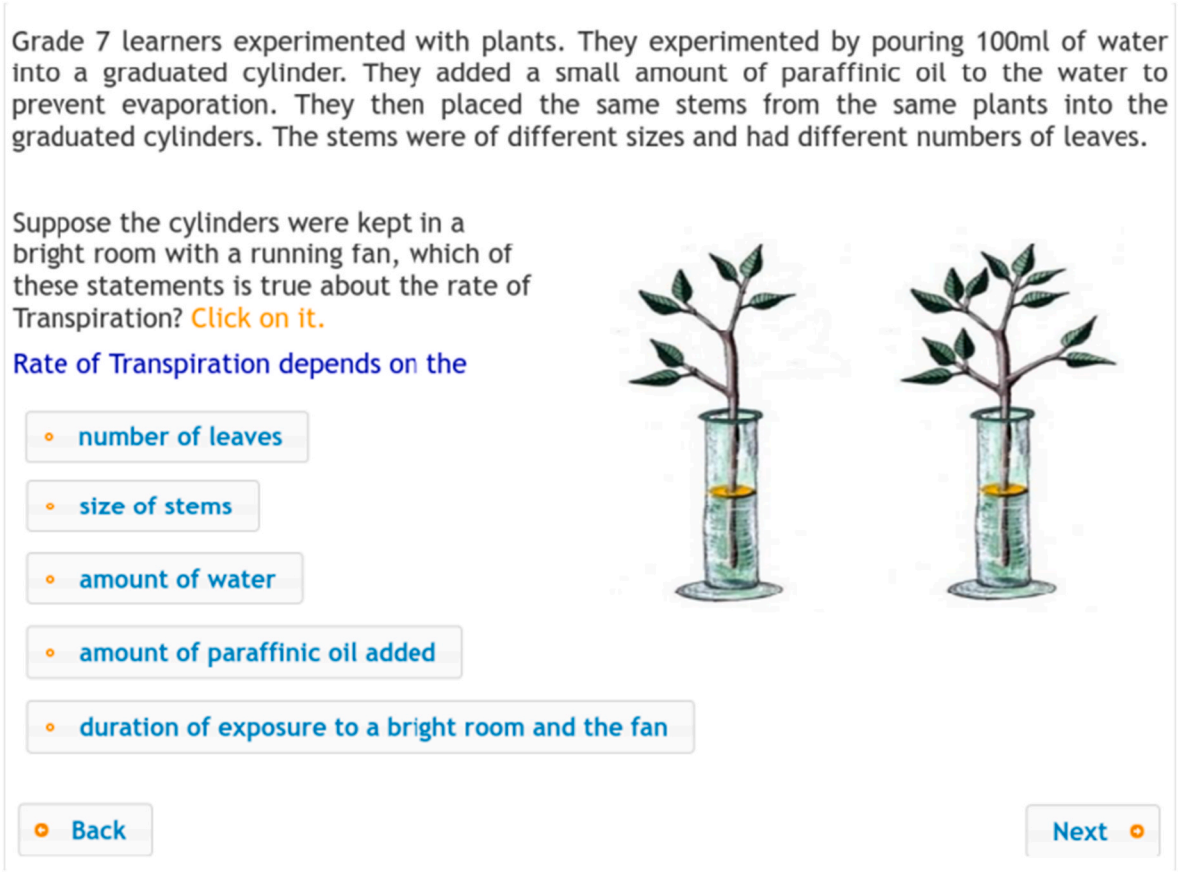

Fig. 5. Sample of control of variable items.

analysis (CFA) (Dunn et al., 2014; Flora, 2020; McNeish, 2018). Since alpha assumes tau equivalence, that is, equal factor loading across items in CFA, which is rarely met (Kline, 2023). Therefore, McDonald's omega (ω) coefficient in the R psych package was used to determine reliability estimates for optimized tests derived from the fitted CFA since it is less biased than Cronbach's alpha in this case (Dunn et al., 2014). Cronbach's alpha and McDonald's omega were interpreted as internal consistency effect size indicators, reflecting the magnitude of reliability in the test versions. Further, omega does not assume tau equivalence (equal factor loadings between items), which is rarely met and not tested by the authors (Rogers, 2024).

The CFA was used to determine the internal structure and construct the validity of the test battery during an initial validation phase. A first-order CFA (Cohen et al., 2017) was performed to test the validity of the construct, specifically (1) whether the data confirmed the theoretical model and structure of the test and (2) whether SR, DR, and CoV were sufficiently unidimensional (i.e., whether the final scores, as unidimensional results, could be taken as indicators of SR, DR, and CoV in subsequent analyzes). Then a second-order CFA was performed to examine the suitability of the SR, DR, and CoV tests to describe the scientific thinking of the students. Several fit indices were used to assess model adequacy: the comparative fit index (CFI) and the Tucker-Lewis index (TLI), for which the' satisfactory' values are those above .90, and the root mean square error of approximation (RMSEA), for which values below .06 indicate a good fit (Hu & Bentler, 1999). The standardized factor loadings obtained from the CFA were interpreted as effect size, indicating the strength of the relationship between items and their underlying constructs (see Appendix B). The model fit considered all fit indices rather than focusing on the cutoff values. The commonly used thresholds do not equally apply to all models, especially when the models differ in their number of dimensions (McNeish & Wolf, 2023; Rogers, 2024). A post hoc power analysis based on RMSEA conducted using R software showed that the sample size (N = 200), with the observed degree of freedom, the study had high statistical power (>.99) to detect deviations from the close fit of the model, indicating that the sample size is adequate for CFA models.

4. Results

4.1. Reliability analysis

This subsection addresses the results relevant to RQ1, which were as follows:

What are the psychometric properties, reliability, and construct validity of the three instruments to assess SR, DR, and CoV? Are the tests sufficiently unidimensional for their final scores to be used as achievement indicators that describe learners' SR, DR and CoV skills?

First, reliability analyzes were performed on all three tests, and items that exhibited low item-total score correlations were excluded. Table 1 presents the reliability indices and the number of items in the original tests, as well as those after the first round of revision and optimization.

As the table shows, all tests exhibited good to excellent reliability in the various stages of development. The results suggested that the

Table 1
Reliability coefficients for the original, revised, and optimized instruments.

Skill	Cronbach's α – Original Test (No. of items)	Cronbach's α – First Revision (No. of items)	Cronbach's α – Optimized Test (No. of items)	McDonald's ω – Optimized Test
SR	.849 (23)	.854 (21)	.873 (14)	.878
DR	.880 (36)	.908 (27)	.912 (25)	.948
CoV	.738 (16)	.755 (13)	.767 (8)	.772

development process improved its internal consistency, producing more reliable measures to assess SR, DR, and CoV. Therefore, the findings strongly suggest that the final scores can be interpreted sufficiently as indicators of the SR, DR, and CV skills of the learners.

Next, a first-order CFA was performed to evaluate the construct validity of the instruments and their unidimensionality. Table 2 presents the CFA fit indices for the original and revised tests.

In all three cases, the revised tests exhibited consistently improved fit indices compared to the original versions, indicating that the revisions further improved their construct validity. These findings suggest that the modified instruments align well with the unidimensional model and are appropriate for assessing the SR, DR and CoV of the students as a single construct. Detailed visualizations of the CFA models for the SR, DR and control variables are provided separately in Appendix A.

4.2. Dimensionality assessment

This subsection addresses the results relevant to RQ2, which were as follows:

To what extent do SR, DR, and CoV function as latent variables that represent key dimensions of scientific thinking? Is the test battery appropriate to assess scientific thinking in the classroom with immediate feedback?

A second-order CFA was conducted to test the theoretical three-dimensional model of scientific thinking and to optimize the length of the tests, building on their construct validity. Before optimization, the fit indices indicated that the model was properly fit (χ^2 [1626] = 2304.727, $p < .001$, CFI = .893, TLI = .887, RMSEA = .047 [.042,.051]). However, the comparative fit indices were slightly below the accepted values, as several items were found to have stronger correlation coefficients between SR and DR or CoV. This indicated that they measure more than just SR, DR, or CoV. Thus, the tests were optimized to create more valid and independent tools for assessing SR, DR, and CoV as distinct dimensions of scientific thinking. The three-dimensional model with the optimized tests fits the data well (χ^2 (1025) = 1419.608, $p < .001$, CFI = .938, TLI = .935, RMSEA = .045 [.039, .050], as shown in Table 3.

In the optimized model shown in Fig. 6, SR (β = .933), DR (β = .914), and CoV (β = .975) proved to be strong indicators of the scientific thinking of students. Model optimization involves the removal of items that had weak factor loading identified in the earlier model. The high standardized loadings in the final model indicate that each component contributes substantially to the higher-order scientific thinking construct, with CoV showing the strongest.

Lastly, time-on-task analyzes confirmed the applicability of the optimized test battery in the classroom context. The average cumulative time that the students spent on the SR, DR, and CoV tests was 31.06 min (SD = 6.94), indicating that the optimized test battery can be completed during a standard classroom period.

5. Discussion

The purpose of this study was to develop and validate a test battery that can be used to evaluate the scientific thinking of junior high school students in a classroom environment, with immediate feedback to them. The tests were developed in a multistage process. After the framework and elements were developed, external experts were asked to validate the content of the items based on the theoretical framework. Then, to validate the constructs and optimize the test battery, an empirical study was conducted with the developed SR, DR, and CoV tests.

Table 2
Fit indices for the one-dimensional models of the original and revised instruments assessing SR, DR, and CoV.

Skill	Model	χ^2	df	P	CFI	TLI	RMSEA
SR	Original	555.631	169	<.001	.884	.870	.111
	Revised	335.928	164	<.001	.949	.940	.075
DR	Original	594.351	299	<.001	.901	.892	.078
	Revised	520.170	295	<.001	.925	.917	.069
CoV	Original	266.013	44	<.001	.589	.486	.178
	Revised	76.878	40	<.001	.932	.906	.076

Table 3
Comparison of model fit indices before and after test optimization.

Model	χ^2 (df)	P	CFI	TLI	RMSEA [90% CI]
First-order three-dimensional model (Before optimization)	2304.727 (1626)	<.001	.893	.887	.047 [.042, .051]
Second-order three-dimensional model (After optimization)	1419.608 (1025)	<.001	.938	.935	.045 [.039, .050]

The reliability of the three instruments was found to be good, and it was then further enhanced by excluding some items with lower item-total score correlations. This supports the generalizability of the test results. A first-order CFA also supported the construct validity of the three instruments; that is, their final scores were determined to be meaningful indicators of the SR, DR, and CoV skills of the students. Then, the fit indices demonstrated that the test items captured the latent constructs as unidimensional models. In sum, the findings affirm that the revised and optimized tests are reliable and valid tools for separately assessing SR, CoV, and DR as unidimensional constructs.

Subsequently, a second-order CFA demonstrated that SR, DR, and CoV are latent dimensions of scientific thinking. Specifically, strong loadings of SR, DR, and CoV in the second-order model confirmed that these skills jointly capture essential aspects of how students engage in scientific thinking (Kuhn, 2010; Zimmerman, 2007). That is, not only are these tests designed to evaluate student SR, DR, and CoV skills theoretically sound individually, but they are also appropriate to evaluate student scientific thinking at a latent level. These findings are consistent with those of previous research that has described scientific thinking as a complex cognitive process involving hypothesis generation, experimental design, prediction, and conclusion drawing (Osborne, 2013; Zhou et al., 2016). The strong contribution of DR corroborates Johnson-Laird and Byrne's (1991) view that deductive logic provides the foundation for coherence and rational decision-making in scientific contexts, while empirical studies have confirmed that students with higher DR skills perform better on tasks that require data interpretation and hypothesis testing (Csapó et al., 2014). Similarly, the SEM results align with the views of Lawson (1978), Osborne (2013), and Kuhn (2010) that scientific thinking is best understood as a higher-order construct made up of distinct but closely connected reasoning skills.

In sum, the test battery is practical for classroom use based on time-on-task analyses. Furthermore, it is practical for classroom use since the average completion time of 31.05 min means that the test can be completed during a single classroom period. Thus, the optimized battery is useable and feasible in a classroom setting, supporting the formative assessment of SR, DR, and CoV during regular instruction.

5.1. Limitations

This study had some limitations. First, its small sample size, distributed across the three grade levels, affects the determination of item parameter estimates of the construct, as well as the construct validity. Therefore, to determine the performance of the students in each test, the number of students would be limited, affecting cross-validation. No further measurement invariance could be examined across grade levels, which limits the replicability.

Second, overreliance on modification indices carries the risk of capitalization on chances based on confirmatory factor analysis, with the resulting model being interpreted as provisional and subject to further validation. Third, the study only focused on subsets of skills from the component of scientific thinking, namely, cognitive variables of SR,

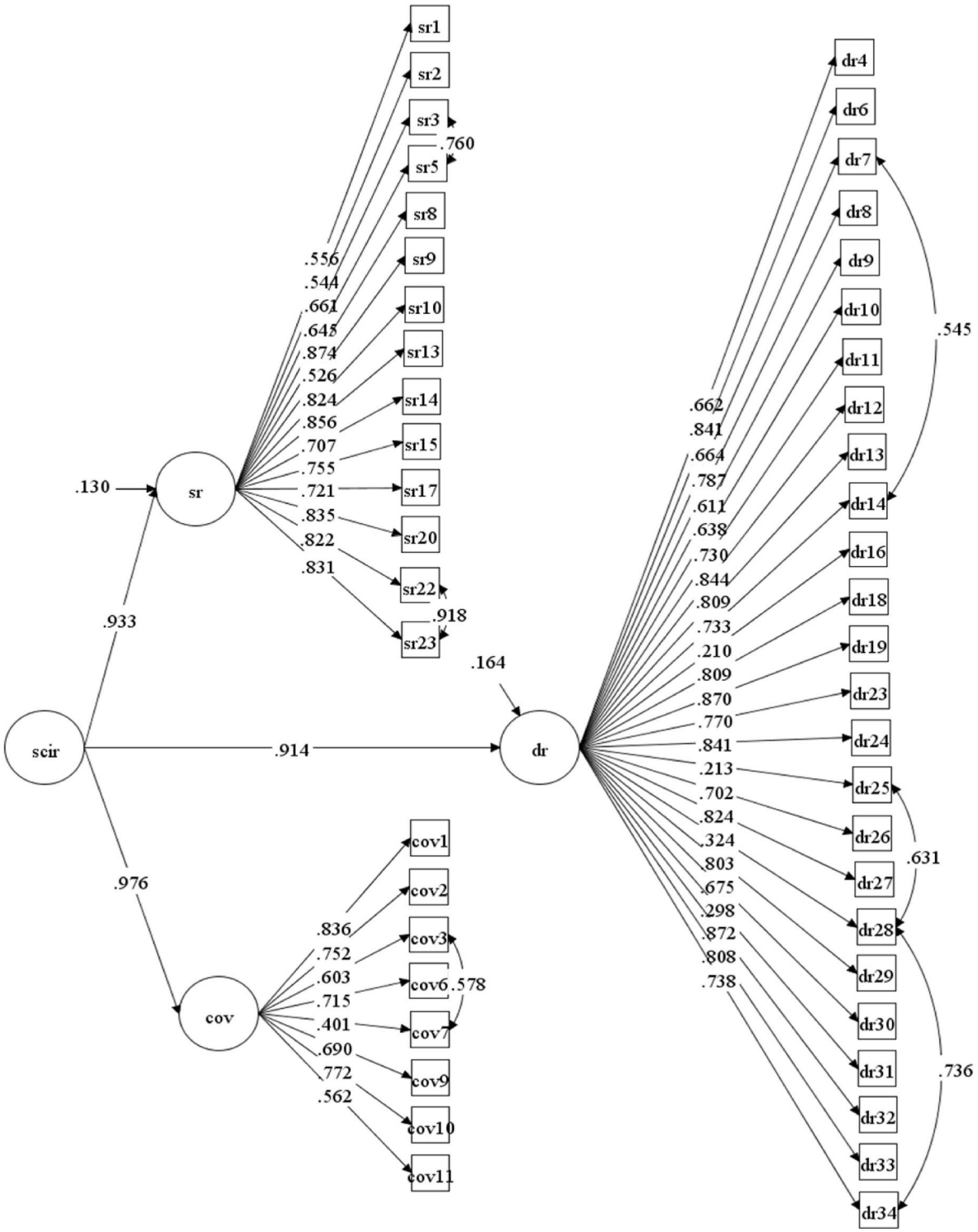

Fig. 6. Second-order CFA: a three-dimensional model of scientific thinking, with SR, DR, and CoV as distinct dimensions.

DR, and CoV; therefore, it did not monitor other potential variables, such as attitude towards science.

Fourth, although scientific thinking was considered a higher-order construct, a unidimensional model may assume that all other abilities are the same. While the second-order model provided a theoretically coherent representation of SR, DR, and CoV, it did not allow for direct comparison with alternative structural representations.

Fifth, an expert review was performed during the development of the item to determine the relevance of the content; however, the content validity measures were not determined.

6. Conclusions

This study aimed to develop and validate instruments for assessing scientific thinking in a classroom context. The newly developed SR, DR, and CoV tests demonstrated good reliability, confirming their internal consistency. Although the first-order CFA supported the construct validity and unidimensionality of the instruments separately, the second-

order CFA demonstrated that SR, DR, and CoV are not only good indicators but also three important dimensions of scientific thinking at the latent level.

In general, the research presented and summarized the steps involved in developing an assessment instrument suitable for measuring scientific thinking in a classroom context. This included every step from framework development to validation, as well as the optimization process, enabling the test battery to be used in classrooms. The findings offer a foundation for improving test design, improving measurement practices, and supporting SR development in educational settings. Therefore, it forms a foundation for future studies to apply item response theory (IRT) approaches to determine the students' ability levels and item characteristics.

CRediT authorship contribution statement

Langat Gilbert Cheruiyot: Writing – original draft, Data curation, Conceptualization. **Gyöngyvér Molnár:** Writing – review & editing, Funding acquisition, Formal analysis, Data curation.

Data

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declaration of generative AI tools

The manuscript was prepared without the application of generative AI tools. Grammarly for proofreading and editing purposes to improve grammar, spelling, and readability. The content was edited and reviewed for publication based on Grammarly's suggestions.

Ethical statement

The study was conducted in accordance with the ethical standards, with ethical approval obtained from the institutional review board before data collection was conducted in Kenya. The participants were junior secondary school students; no personal identifiers were collected, and all the data were treated confidentially and used only for academic purposes.

Ethical approval

The Institutional Review Board (IRB) of the Doctoral School of Education, University of Szeged has recently reviewed your application for an ethical approval (Title of the Research Project: "*Computer*-supported *collaborative training of scientific reasoning on academic achievement in science in Kenya*", supervisor: Prof. Dr. Gyöngyvér Molnár). This proposal was evaluated in light of the requirements of the ethical conducts on social research with human subjects of the Doctoral School of Education, University of Szeged.

IRB decision: approved

Justification:

This proposal is deemed to meet the requirements of the ethical conducts on social research with human subjects of the Doctoral School of Education, University of Szeged. The study aims to monitor the effectiveness of the competency-based curriculum by comparing students' level of scientific knowledge and their development of reasoning skills, and to develop a computer-supported collaborative training of scientific reasoning and test their longitudinal effects. Participants are boys and girls from different Junior secondary schools in Kenya (age bracket 13-14 years). Online, edia system will be used in administering tests and questionnaires applying a true experimental design. Participation is voluntary. Written informed consent will be signed by the participants and their parents. Data collection is registered by code, data collected will be managed using edia system. Procedure of the data collection does not harm their privacy law, it does not have an impact on the participants' mental or physical health. Data cannot be handled by persons to whom they are not concerned.

Funding statements

This study has been conducted with support from the National Research, Development and Innovation Fund of Hungary, financed under the OTKA 152413 funding scheme, and was supported by a grant from the Hungarian Academy of Sciences KOZOKT2025-4 and the Open Access Grant of the University of Szeged.

Declaration of competing interest

There are no conflicts of interest to declare.

Acknowledgment

This research was supported by a Hungarian Academy of Sciences Research Programme for Public Education Development grant (KOZOKT2021-16) and by the Humanities and Social Sciences Cluster of the Centre of Excellence for Interdisciplinary Research, Development, and Innovation of the University of Szeged. The authors are members of the Digital Learning Technologies Incubation Research Group. The authors thank the junior secondary schools in Kenya for their participation and for providing the necessary support and resources, and acknowledge the University of Szeged for financial support.

Appendix A. Confirmatory Factor Analysis (CFA) Diagrams

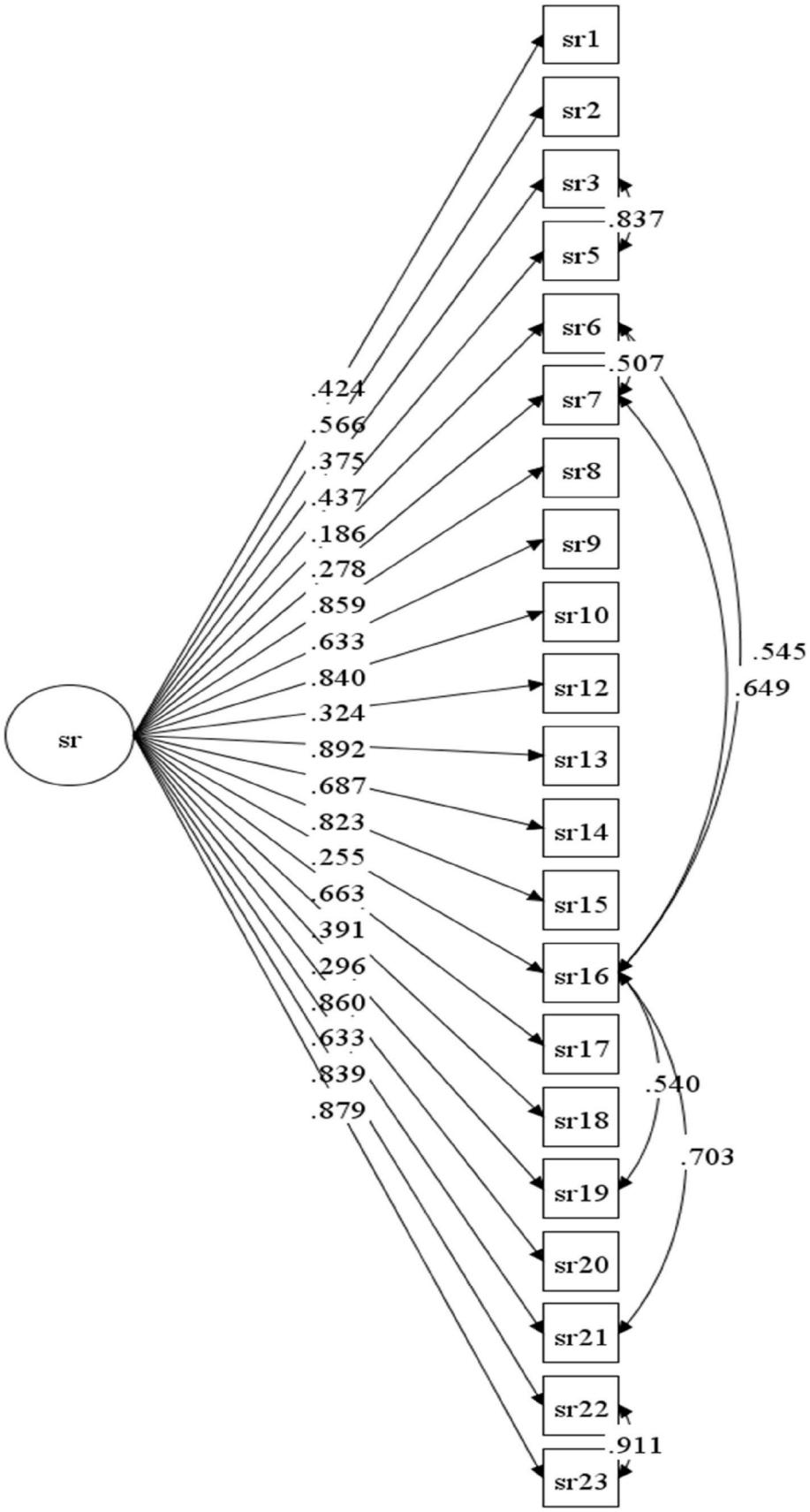

Fig. A1. CFA model for Scientific Reasoning (SR).

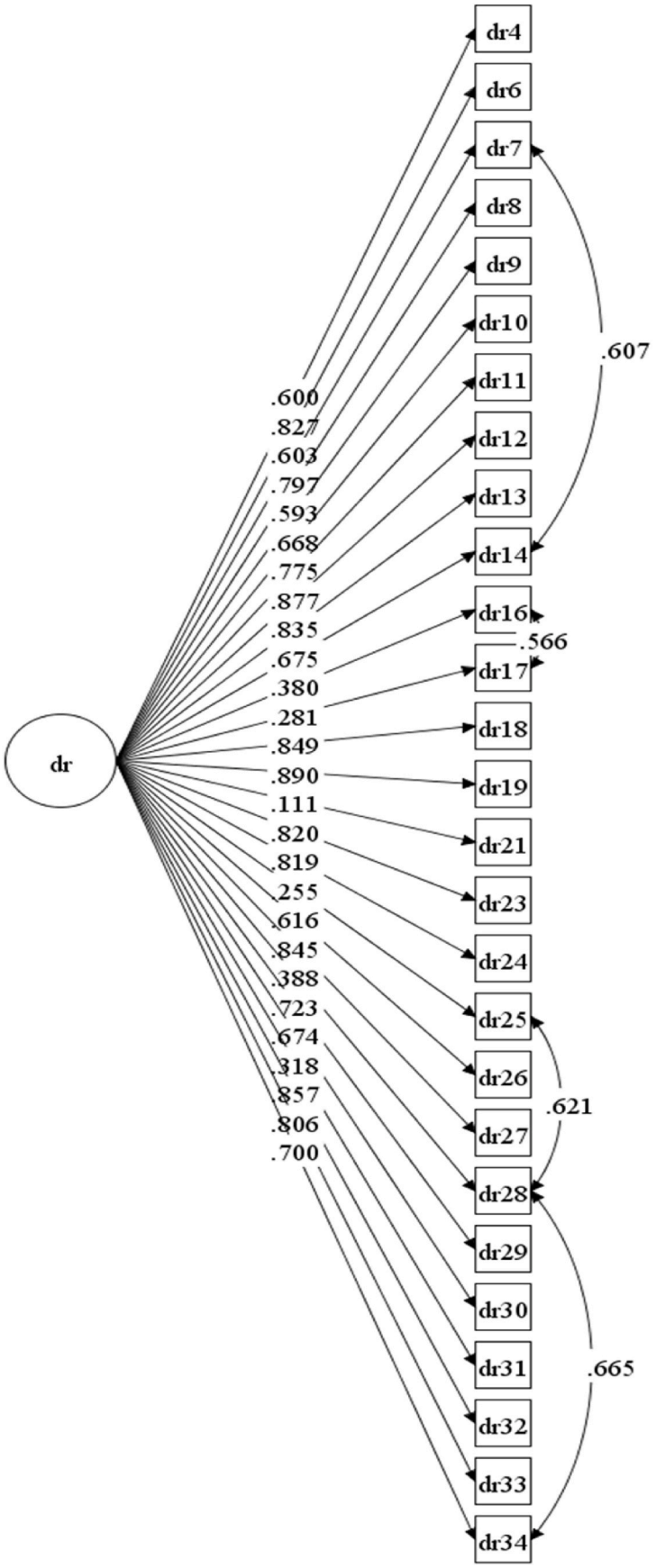

Fig. A2. CFA model for Deductive Reasoning (DR).

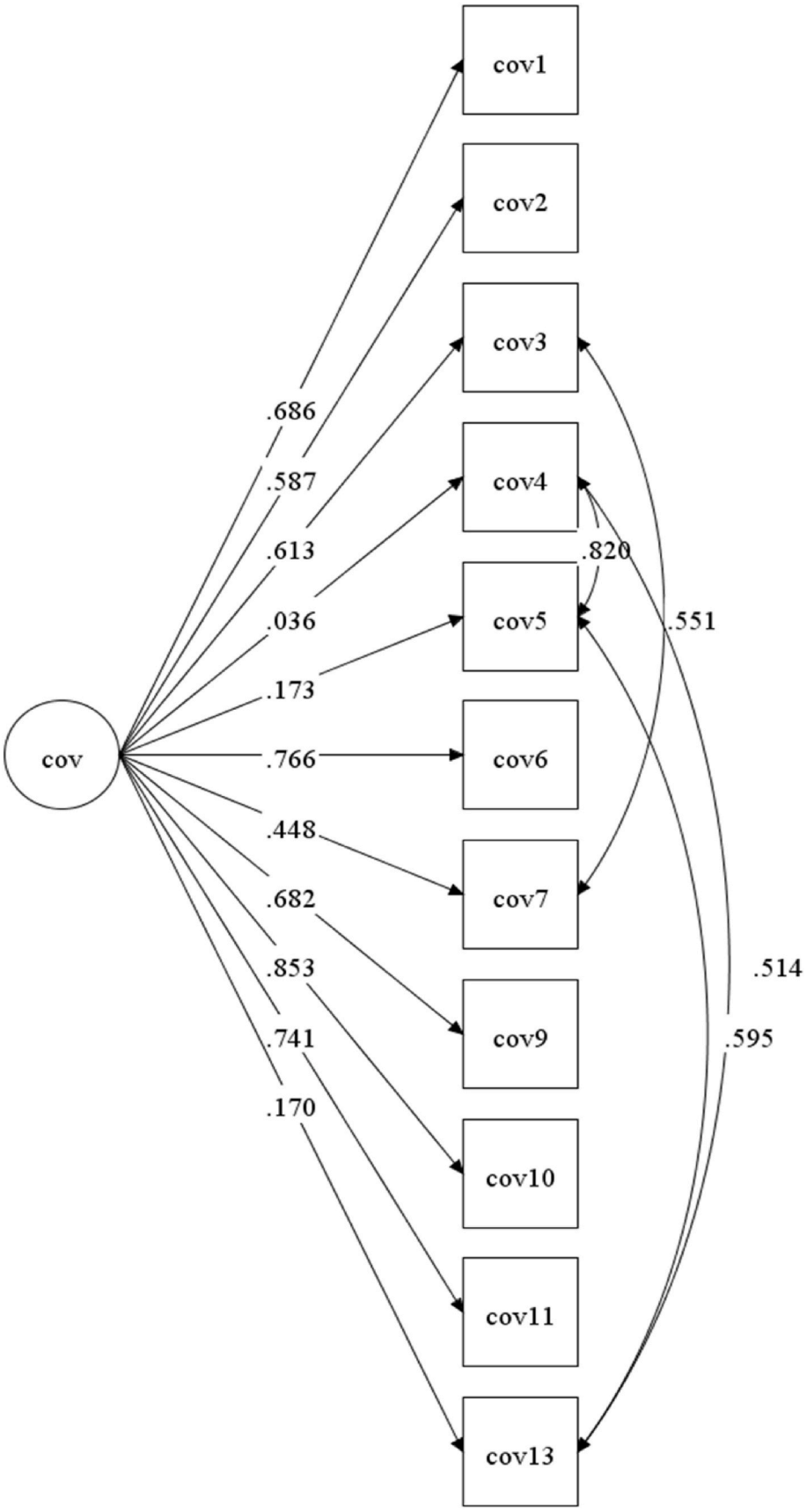

Fig. A3. CFA model for Control of Variables (CoV).

Appendix B. Standardized Factor Loadings

Table B1
Scientific Reasoning

Item	Factor loading	p-
SR1	.437	<.001
SR2	.564	<.001
SR3	.637	<.001
SR5	.675	<.001
SR6	.218	.013
SR7	.310	<.001
SR8	.862	<.001
SR9	.613	<.001
SR10	.826	<.001
SR12	.293	.002
SR13	.896	<.001
SR14	.676	<.001
SR15	.815	<.001
SR16	.328	<.001
SR17	.665	<.001
SR18	.313	<.001
SR19	.302	.001
SR20	.806	<.001
SR22	.830	<.001
SR23	.867	<.001

Table B2
Deductive reasoning

Item	Factor loading	p
DR01	.410	<.001
DR02	.011	.910
DR04	.605	<.001
DR06	.820	<.001
DR07	.646	<.001
DR08	.797	<.001
DR10	.657	<.001
DR11	.766	<.001
DR12	.862	<.001
DR13	.820	<.001
DR14	.697	<.001
DR16	.342	<.001
DR18	.835	<.001
DR19	.877	<.001
DR23	.806	<.001
DR24	.810	<.001
DR26	.621	<.001
DR27	.828	<.001
DR28	.394	<.001
DR29	.706	<.001
DR30	.651	<.001
DR32	.846	<.001
DR33	.807	<.001
DR34	.691	<.001

Table B3
Control of variable factor loading

CoV	Factor loadings	p
CoV1	.686	<.001
CoV2	.587	<.001
CoV3	.613	<.001
CoV4	.036	.765
CoV5	.173	.187
CoV6	.766	<.001
CoV7	.448	<.001
CoV9	.682	<.001
CoV10	.853	<.001
CoV11	.741	<.001
CoV13	.170	.094

Note. Standardized factor loadings values obtained from confirmatory factor analysis and are interpreted as effect-size of the relationships between items and the latent factor.

References

Althouse, L. A. (2000). Test development: Ten steps to a valid and reliable certification exam. In *Proceedings of the twenty-Fifth Annual SAS users group international conference (SUGI 25)*. SAS Institute Inc.

Ayalon, M., & Even, R. (2010). Mathematics educators' views on the role of mathematics learning in developing deductive reasoning. *International Journal of Science and Mathematics Education, 8*(6), 1131–1154. https://doi.org/10.1007/s10763-010-9238-z

Bao, L., Koenig, K., Xiao, Y., Fritchman, J., Zhou, S., & Chen, C. (2022). Theoretical model and quantitative assessment of scientific thinking and reasoning. *Physical Review Physics Education Research, 18*(1), Article 010115. https://doi.org/10.1103/PhysRevPhysEducRes.18.010115

Carreira, S., Amado, N., & Jacinto, H. (2020). Venues for analytical reasoning problems: How children produce deductive reasoning. *Education Sciences, 10*(6), 169. https://doi.org/10.3390/educsci10060169

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of the variables strategy. *Child Development, 70*(5), 1098–1120. https://doi.org/10.1111/1467-8624.00081

Cohen, L., Manion, L., & Morrison, K. (2017). *Research methods in education (8th)*. Routledge. https://doi.org/10.4324/9781315456539

Csapó, B., & Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in Psychology, 10*, 1522. https://doi.org/10.3389/fpsyg.2019.01522

Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school readiness and early reasoning. *Journal of Educational Psychology, 106*(3), 639–650. https://doi.org/10.1037/a0035756

De Chantal, P.-L., & Markovits, H. (2017). The capacity to generate alternative ideas is more important than inhibition for logical reasoning in preschool-age children. *Memory & Cognition, 45*(2), 208–220. https://doi.org/10.3758/s13421-016-0653-4

Dunbar, K. N., & Klahr, D. (2012). Scientific thinking and reasoning. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 701–718). London: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0035.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Echevarria, M. (2003). Anomalies as a catalyst for middle school students' knowledge construction and scientific reasoning during science inquiry. *Journal of Educational Psychology, 95*, 357–374. https://doi.org/10.1037/0022-0663.95.2.357

Evans, J., & St, B. T. (2005). *Deductive reasoning*. Cambridge University Press. https://www.cambridge.org/9780521824170.

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science, 3*(4), 484–501. https://doi.org/10.1177/2515245920951747

Hogan, K., & Fisherkeller, J. (2005). Dialogue as data: Assessing students' scientific reasoning with interactive protocols. In J. D. Novak (Ed.), *Assessing science understanding* (pp. 95–127). Burlington: Academic Press. https://doi.org/10.1016/B978-012498365-6/50007-X.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1–55. https://doi.org/10.1080/10705519909540118

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin, 138*(3), 427–457. https://doi.org/10.1037/a0026841

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. MIT Press. https://doi.org/10.7551/mitpress/2939.001.0001

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). The Guilford Press.

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development, 86*, 327–336. https://doi.org/10.1111/cdev.12298

Korom, E., Németh, M. B., Nagy, L., & Csapó, B. (2017). Diagnostic assessment frameworks for science: Theoretical background and practical issues. In B. Csapó, & G. Szabó (Eds.), *Framework for diagnostic assessment of science* (pp. 147–174). Budapest: Nemzeti Tankönyvkiadó.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Kuhn, D. (2005). *Education for thinking*. Cambridge, MA: Harvard University Press.

Kuhn, D. (2010). What is scientific thinking, and how does it develop? In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 371–393). Blackwell Publishers. https://doi.org/10.1002/9780470996652.ch17.

Lawson, A. E. (1978). Development and validation of the classroom test of formal reasoning. *Journal of Research in Science Teaching, 15*(1), 11–24. https://doi.org/10.1002/tea.3660150103

Lawson, A. E. (2000). *Classroom test of scientific reasoning*. Arizona State University.

Lawson, A. E. (2005). What is the role of induction and deduction in reasoning and scientific inquiry? *Journal of Research in Science Teaching, 42*(6), 716–740. https://doi.org/10.1002/tea.20067

Lazonder, A. W., & Janssen, N. (2018). Development and initial validation of a performance-based scientific reasoning test for children. *Journal of Experimental Child Psychology, 171*, 105–123. https://doi.org/10.1016/j.stueduc.2020.100951

Lazonder, A. W., Janssen, N., Gijlers, H., & Walraven, A. (2021). Patterns of development in children's scientific reasoning: Results from a three-year longitudinal study. *Journal of Cognition and Development, 22*(1), 108–124. https://doi.org/10.1080/15248372.2020.1814293

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods, 28*(1), 61–88. https://doi.org/10.1037/met0000425

Millar, R., & Driver, R. (1987). Beyond processes. *Studies in Science Education, 14*(1), 33–62. https://doi.org/10.1080/03057268708559938

Molnár, G., & Csapó, B. (2018). The efficacy and development of students' problem-solving strategies during compulsory schooling: Log file analyses. *Frontiers in Psychology, 9*, 302. https://doi.org/10.3389/fpsyg.2018.00302.

Osborne, J. (2013). The 21st-century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10*, 265–279. https://doi.org/10.1016/j.tsc.2013.07.006

Pedaste, M., Maeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A. N., Kamp, E. T., … Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review, 14*, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003

Radulović, B., & Stojanović, M. (2018). Research and evaluation of hypothetically deductive student reasoning in the Republic of Serbia. *Facta Universitatis – Series: Physics, Chemistry and Technology, 16*(3), 249–256. https://doi.org/10.2298/FUPCT1803249R

Rogers, P. (2024). Best practices for your confirmatory factor analysis: A JASP and lavaan tutorial. *Behavior Research Methods, 56*(7), 6634–6654. https://doi.org/10.3758/s13428-024-02375-7

Ross, J. A. (1988). Controlling variables: A meta-analysis of training studies. *Review of Educational Research, 58*(4), 405–437. https://doi.org/10.3102/00346543058004405

Schwichow, M., Christoph, S., Boone, W. J., & Härtig, H. (2016). The impact of sub-skills and item content on students' skills in the control-of-variables strategy. *International Journal of Science Education, 38*(2), 216–237. https://doi.org/10.1080/09500693.2015.1137651

Sternberg, R. J. (2012). Intelligence. *Dialogues in Clinical Neuroscience, 14*(1), 19–27. https://doi.org/10.31887/DCNS.2012.14.1/rsternberg

Tytler, R., & Peterson, S. (2003). Tracing young children's scientific reasoning. *Research in Science Education, 33*(4), 433–465. https://doi.org/10.1023/B:RISE.0000005250.04426.67

Walton, D. (2008). *Informal logic: A pragmatic approach*. Cambridge University Press.

Yaşar, O. (2022). Scientific thinking: A mindset for everyone. In N. Rezaei (Ed.), *Integrated education and learning. Integrated science*. Cham: Springer. https://doi.org/10.1007/978-3-031-15963-3_3.

Zhou, S., Han, J., Koenig, K., Raplinger, A., Pi, Y., Li, D., & Bao, L. (2016). Assessment of scientific reasoning: The effects of task context, data, and design on student reasoning in control of variables. *Thinking Skills and Creativity, 19*, 175–187. https://doi.org/10.1016/j.tsc.2015.11.004

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*(1), 99–149. https://doi.org/10.1006/drev.1999.0497

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172–223. https://doi.org/10.1016/j.dr.2006.12.001

Zimmerman, C., & Klahr, D. (2018). Development of scientific thinking. In S. Ghetti (Ed.), *The stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 223–248). Wiley.