JOURNAL OF BALTIC SCIENCE EDUCATION

ISSN 1648-3898 /Print/
ISSN 2538-7138 /Online/

This is an open access article under the **Creative Commons Attribution 4.0 International License**

MEASURING SCIENTIFIC REASONING OF FOURTH GRADERS: VALIDATION OF THE SCIENCE-K INVENTORY IN PAPER-BASED AND COMPUTER-BASED TESTING ENVIRONMENTS

Mário Tibor Nagy,
Erzsébet Korom

Abstract. *Nowadays, the assessment of student performance has become increasingly technology-based, a trend that can also be observed in the evaluation of scientific reasoning, with more and more of the formerly paper-based assessment tools moving into the digital space. The study aimed to examine the reliability and validity of the paper-based and computer-based forms of the Science-K Inventory, which assesses children's scientific reasoning in three aspects: experimentation, data interpretation, and understanding of the nature of science. The pilot study involved 84 fourth-grade Hungarian students, with 39 students taking the paper-based test and 45 students taking the computer-based test. Rasch measurements and reliability tests have indicated that both the paper-based and computer-based test versions are equally valid for assessing the scientific reasoning skills of fourth graders. Students achieved high test scores in both mediums, and there were no significant differences between boys' and girls' scientific reasoning in either test type. The novelty of this research was that the Science-K Inventory had not yet been tested in a computer-based format. The results demonstrate that the Science-K Inventory can be effectively utilized in digital testing to provide teachers with rapid and valuable information for fostering the development of their students' scientific reasoning.*

Keywords: *computer-based testing, paper-based testing, primary school, Science-K Inventory, scientific reasoning*

Mário Tibor Nagy,
Erzsébet Korom
University of Szeged, Hungary

Introduction

Science education in the 21st century seeks to develop scientific literacy that will enable society to deal effectively with technological progress, the rapid and sometimes uncontrolled spread of information and global challenges. This requires, in addition to disciplinary knowledge, appropriate scientific thinking and inquiry skills, as well as an understanding of the nature of science. Scientific thinking is the set of mental processes that are used to solve scientific problems, acquire scientific knowledge, carry out investigations and reflect on research results (Dunbar & Fugelsang, 2005; Dunbar & Klahr, 2012; Kuhn, 2011). The domain-specific strand of research on scientific thinking focuses on children's conceptual development and conceptual change in different disciplinary domains, while the domain-general strand focuses more on identifying the reasoning processes involved in the acquisition of scientific knowledge (Zimmerman, 2007; Zimmerman & Klahr, 2018). The current study is concerned with the latter research area.

Scientific reasoning plays a central role in the acquisition of scientific literacy. Higher-order reasoning skills are needed to solve problems and make decisions, just as they are needed to understand complex concepts or the nature of science (Lawson, 2004). Research on the development of reasoning skills is rooted in the work of Piaget (Inhelder & Piaget, 1958), who investigated children's reasoning operations in various tasks (e.g., pendulum, balancing) and the process by which children systematically explore the world: formulating hypotheses, predicting outcomes, manipulating variables, observing, and making inferences. While early research (see, for example, Inhelder & Piaget, 1958; Kuhn et al., 1988) suggested that the acquisition of scientific reasoning skills could not begin until adolescence, recent research suggests that some skills can be developed earlier (Zimmerman, 2007). There is growing evidence that primary school students have a wide range of scientific reasoning skills:

https://doi.org/10.33225/jbse/23.22.1050

by the age of 8 years, students prefer controlled experiments over experiments with confounding features (Bullock et al., 2009); they can distinguish between hypotheses and evidence (Sodian et al., 1991); children have the ability to understand the control-of-variables strategy (CVS) (Peteranderl & Edelsbrunner, 2020); they can draw inferences from graphical data (Koerber & Sodian, 2009); and they have a rudimentary understanding of the Nature of Science (NOS) (Koerber et al., 2015).

Although scientific reasoning is rooted in the intuitive information-seeking process, it does not develop spontaneously. Its development and change are influenced by individual and contextual factors (Morris et al., 2012). Therefore, it is necessary to examine students' scientific reasoning skills and to elaborate activities that support the development of reasoning from early childhood. This requires valid and reliable instruments. Despite the fact that the development of scientific reasoning in pre-school and primary school is a dynamic area of research (O'Connor et al., 2021), relatively few measurement tools are available. One of these is the Science-K Inventory (SK-I), which has been the most widely used measure among German children (see Koerber & Osterhaus, 2019; Osterhaus & Koerber, 2023). Therefore, this research examines the usability of the Hungarian adaptation of the Science-K Inventory in both paper-based and computer-based testing environments. The paper-based survey is a modified version of the original survey procedure, while the development of the computer-based version represents a new innovation in the history of the Science-K Inventory because this test has not yet been applied in a digital environment.

Literature Review

Evaluating Early Scientific Reasoning

Common methods of assessing young people's scientific reasoning include experiments carried out independently with ordinary or simple hands-on physical apparatus, and problem-based and story-based simulations (Mayer et al., 2014). Students' performance on the test is influenced by the type, complexity, and level of abstraction of the tasks used (Lazonder & Kamp, 2012). The difficulty and complexity of the components of basic scientific reasoning can be effectively assessed using tasks that are embedded in problem-based stories, use few variables, are cognitively less demanding, and do not involve children's prior emotional beliefs about the story (Schulz & Gopnik, 2004). Research by Bullock and Ziegler (1999) showed that when identifying and controlling variables in tests of scientific reasoning, students performed better when they were given a choice of closed-response alternatives rather than having to formulate the correct answer themselves. In multiple-choice tests, students who performed well were able to explain their choice thus demonstrating their understanding of the controlled test (Mayer et al., 2014). Paper and pencil tests, which have recently transitioned to online platforms for data collection based on practical tasks, offer advantages such as practicality and the ability to carry out large-scale measurements with larger sample sizes. They also facilitate the examination of relationships with other cognitive and affective factors (Strand-Cary & Klahr, 2008).

The Science-K Inventory

The above criteria are well met by the Science-K Inventory developed by Koerber and Osterhaus (2019), which is designed to comprehensively measure scientific reasoning skills in pre-schoolers. The Science-K inventory comprises 30 multiple-choice items with a brief story and is divided into three components, namely experimentation, data interpretation, and understanding the nature of science. Each component consists of 10 items. In the experimentation dimension, students are asked to choose the most appropriate experiment to test the research problem being investigated and to solve tasks that require them to correctly apply the CVS to test hypotheses. CVS is a crucial indicator of the development of scientific reasoning (Schwichow et al., 2020). In the data interpretation component, children are asked to identify covariation patterns (i.e., perfect, imperfect or non-covariations) in the data and then use the identified patterns to formulate a hypothesis. They also need to understand that when they encounter confusing or contradictory data, it will be challenging to draw clear conclusions. The third part of the exercises focuses on understanding the nature of science. The tasks are designed to address questions such as what scientists do as part of their job and what research questions they ask.

The SK-I was validated in German (Koerber & Osterhaus, 2019) and has also demonstrated validity and reliability in other languages and cultures, such as Chinese (Osterhaus et al., 2023). Despite being primarily designed for kindergarten children, the SK-I has also been tested with third graders, where it performed well (Nyberg et al.,

https://doi.org/10.33225/jbse/23.22.1050

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

2020), and has been used in a longitudinal study of children from five to ten years of age (Osterhaus & Koerber, 2023). The data collection method used in previous studies was paper-based, utilizing interviews (where the items were presented in an illustrated booklet) for kindergarten children and group tests (read out by an experimenter and presented in a PowerPoint presentation) for primary school children (Koerber & Osterhaus, 2019; Nyberg et al., 2020; Osterhaus et al., 2023; Osterhaus & Koerber, 2023).

Computer-based Assessment of Scientific Reasoning

Computer-based assessment has now almost completely taken over the assessment of the learning process, surpassing the use of paper-based tests. Rapid advances in technology have made it possible to assess more complex skills (e.g., critical thinking (e.g., Rosen & Tager, 2014), problem-solving (e.g., Wu & Molnár, 2018), creativity (e.g., Pásztor et al., 2015) and collaboration (e.g., Rosen, 2015) more reliably through computer-based testing (Shute & Rahimi, 2017). Another advantage is that multimedia elements, simulations, and dynamic items can be integrated into the testing environment (Csapó et al., 2014). Furthermore, computer-based testing in large-scale assessment programs using item banks allows for targeted and adaptive test development and creation. This form of testing reduces the resources needed to carry out tests. Additionally, it enables rapid communication of test results and personalized evaluator feedback (Bennet, 2003; Quellmalz & Pellegrino, 2009). Recognizing these benefits, one of the most comprehensive international assessments, the OECD's PISA (Programme for International Student Assessment) test, also completely switched (except for nine countries) to a computer-based testing environment in 2018 (OECD, 2019).

There is a substantial body of research in the literature on the validation of tests in a digital environment regarding scientific reasoning and inquiry skills. The main research questions of these research directions are whether paper-based and computer-based tests are equivalent in terms of internal consistency and validity (Williamson et al., 2017). For example, Vo and Csapó (2023) conducted a study examining the performance of 731 students on the application of the control-of-variables strategy in basic physics. The study compared paper-based testing under teacher supervision with online testing without teacher supervision. The findings indicated that paper-based testing exhibited better reliability and total scores compared to online testing. However, the validity of the two tests was found to be equivalent regardless of the difference in delivery modalities (Vo & Csapó, 2023). Another study (Schroeders et al., 2013) found that the test medium (paper-pencil or technology-based) is not a determinant of comprehension ability in the natural sciences. Research conducted by Halldórsson and colleagues (2009) examines modality effects on gender performance by comparing the scientific literacy results in the PISA 2006 assessment, both in paper-and-pencil format and computer-based format. The findings indicate that in all three participating countries (Iceland, Denmark, and Korea), boys outperformed girls in science literacy when the test was administered via computer. However, there are studies that have shown no gender differences in student performance in science tests in either test setting (e.g., Adanır et al., 2020; Brallier et al., 2015) In addition, it is important to note that learners tend to prefer computer-based testing, which can impact their test-taking attitudes and self-efficacy (Nikou & Economides, 2016; OECD, 2010).

Research Aim and Research Questions

The aim of the present study has been to explore the potential of paper-based and computer-based assessments for the Science-K Inventory among fourth graders in primary school. For this purpose, the SK-I was adapted to Hungarian, transferred into an online interface, and the paper-based (PB) and computer-based (CB) test versions were compared by addressing the following research questions:

- RQ1: Does the computer-based test measure similar reliability to the paper-based format?
- RQ2: Is there any evidence of equivalence between the online and paper-based groups on the SK-I at the item and task levels?
- RQ3: Is there a gender difference in students' overall test scores and their scores across different test contexts?
- RQ4: How do students perceive the difficulty of the test, and how interesting did they find the tasks?

https://doi.org/10.33225/jbse/23.22.1050

Research Methodology

General Background

This cross-sectional study employed a quantitative approach and utilized the Rasch model, based on item response theory (IRT), to illustrate the probability of an individual successfully solving an item in two test modes. This probability is determined by the relationship between the latent variable of ability and the item's difficulty expressed on a linear scale (Rasch, 1960; Vo & Csapó, 2023). Internal consistency was measured using Cronbach's alpha to examine whether the reliability of the Science-K Inventory, which is employed for assessing the scientific reasoning of 4th-grade students in a computer-based testing environment, was consistent with that of the paper-based version.

It is important to emphasize that the research described in this paper represents the outcome of the initial data collection phase (pilot study) conducted on a small sample after adapting the instrument. This phase served as the initial validation step in a broader research project aimed at large-scale assessment of the scientific reasoning of primary school students within a digital environment, which is ending this year.

Participants

The pilot study involved 84 fourth-grade students (M_{age} = 10.18, SD = 0.47; 50.0% girls) from three classes in a primary school in Szeged. One of the classes participated in the Golden Gate English Playschool program (an accredited English language program), which is very popular among parents. As a result, in this particular class students of high ability and a good social background were overrepresented. To determine who would take paper-based tests and who would take computer-based tests, a random number generator was used to select students randomly for all three classes. Thus, 39 students (53.8% girls) completed the paper-based test, while 45 students (46.7% girls) completed the computer-based test. The absence of some students at the time of testing resulted in a slight imbalance in the proportions between the two test modes. At the beginning of the school year, the school administration asked parents for written permission for their children to participate in educational research.

Paper- and Computer-based Instruments for the Assessment of Scientific Reasoning

The students' scientific reasoning was assessed using the Science-K Inventory which comprises 30 multiple-choice items divided into 3 components (10 items each): experimentation (EXP), data interpretation (DAT), and understanding the nature of science (NOS) (Koerber & Osterhaus, 2019). For each item, children are required to choose the best out of three answer options. Tasks of varying difficulty are adapted to children's emerging scientific reasoning processes. For the purposes of scoring, each correct answer was worth 1 point, allowing students to obtain a maximum score of 30 points for answering all 30 items correctly. The test makers emphasized that the test is not a tool for measuring sub-skills, but rather an instrument for providing a more comprehensive and general assessment of scientific reasoning (Koerber & Osterhaus, 2019). For this reason, we considered the test as having a one-dimensional structure when analysing and evaluating the data.

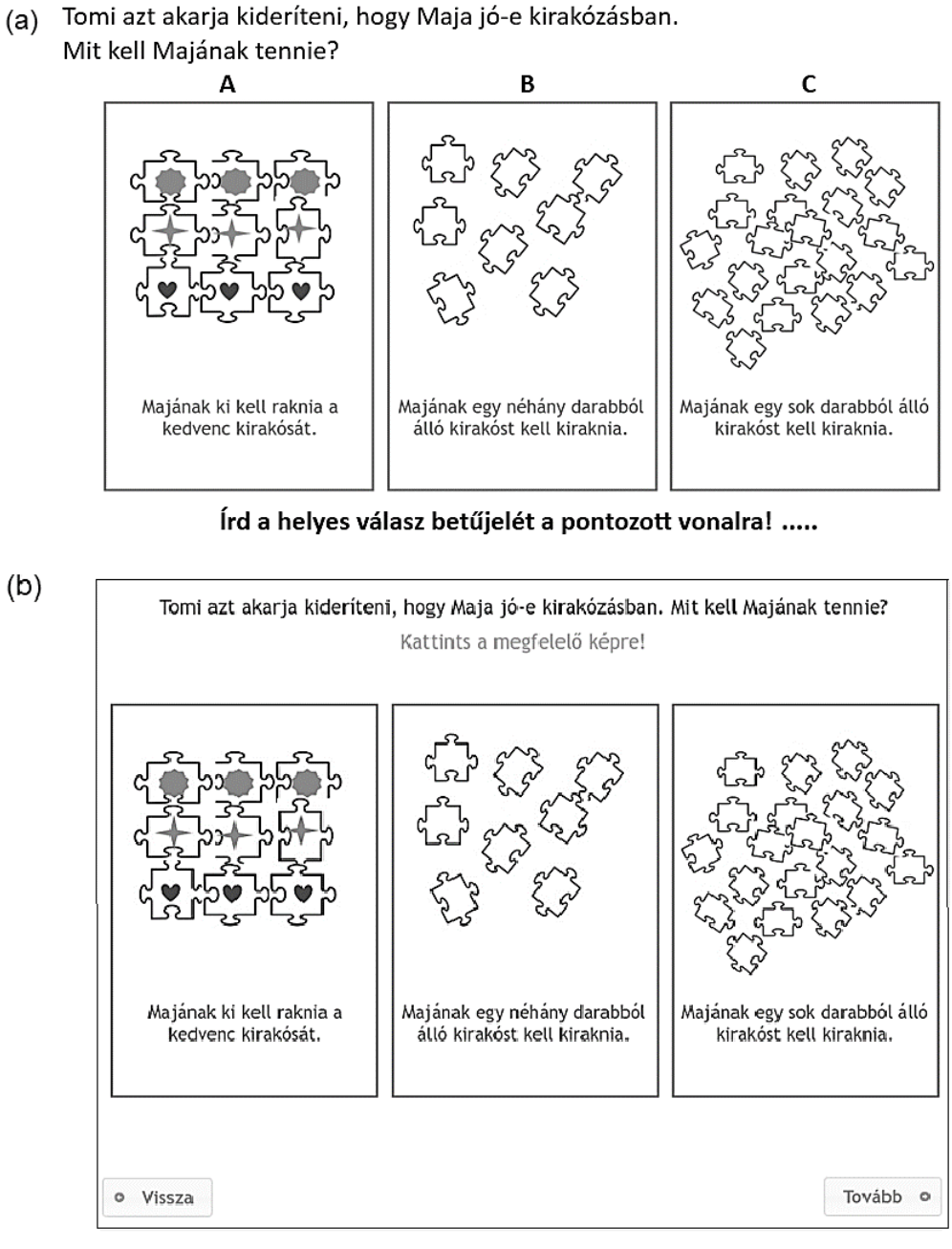
The adaptation procedure, which involved the contribution of Christopher Osterhaus, one of the developers of the instrument, was carried out in accordance with the established rules. Assistance was received from an English-speaking translator and an independent expert, both of whom provided support in the research undertaken. In contrast to previous group tests, where data was collected with the assistance of an experimenter who read the tasks aloud and the students had to mark the correct answer on a score sheet, in this study, students were asked to complete the test individually. In adapting the instrument, available slides were utilized to generate a paper-based test booklet and its digital version. Throughout the process, efforts were made to preserve the original structure and illustrations of the instrument. The only difference between the paper-based and digital versions of the test we created is the method of administration: in the paper-based test, the letter of the correct answer was written on the dotted line or encircled; in the digital version, the student had to click on the image of the correct answer in the experimentation and NOS subtests, while in the data interpretation subtest, the student had to click an icon to the left of the correct answer. Examples of these implementations are shown in Figures 1 and 2.

https://doi.org/10.33225/jbse/23.22.1050

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Figure 1

Sample Item from the Paper-based (a) and the Computer-based (b) Version of an Experimentation Task



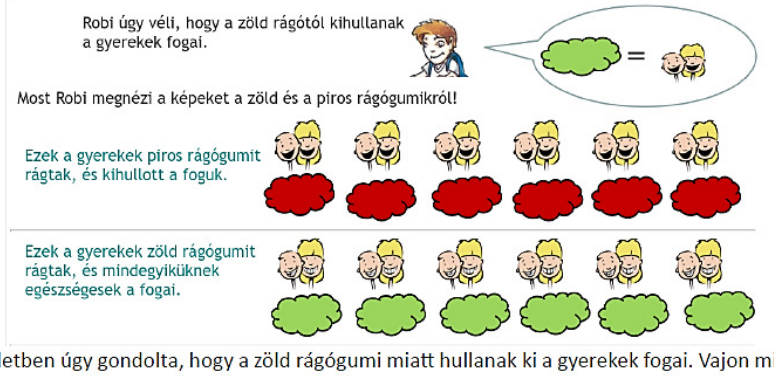
Note. Tom wants to find out if Mia is good at doing puzzles. What does she have to do? Write the letter of the correct answer on the dotted line/Click on the correct picture.

https://doi.org/10.33225/jbse/23.22.1050

Figure 2

Sample Item from the Paper-based (a) and the Computer-based (b) Version of a Data Interpretation Task

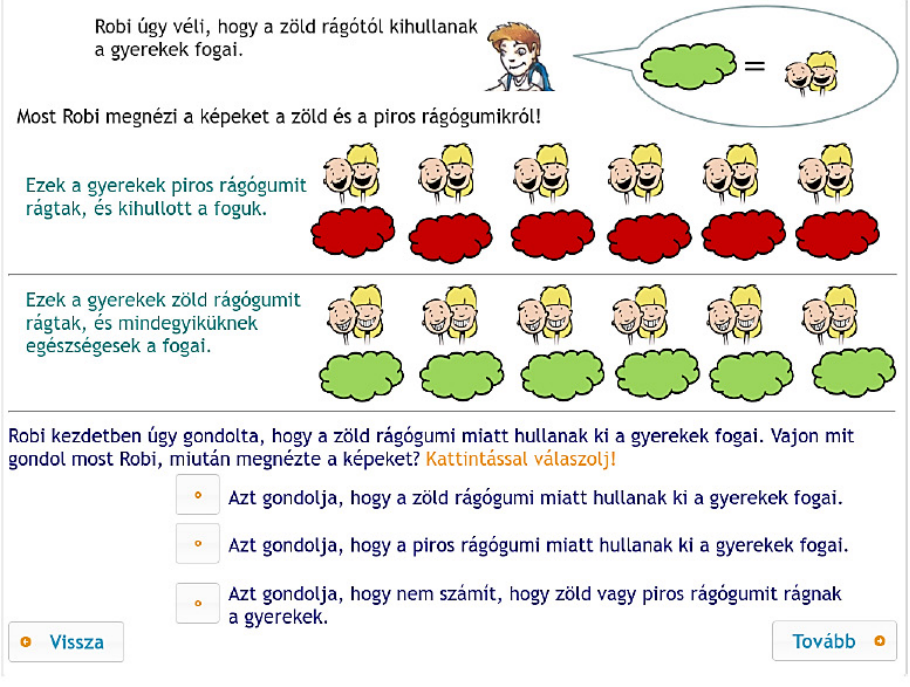
(a)



Robi kezdetben úgy gondolta, hogy a zöld rágógumi miatt hullanak ki a gyerekek fogai. Vajon mit gondol most Robi, miután megnézte a képeket? *Karikázd be a helyes válasz betűjelét!*

A) Azt gondolja, hogy a zöld rágó miatt hullanak ki a gyerekek fogai.
B) Azt gondolja, hogy a piros rágó miatt hullanak ki a gyerekek fogai.
C) Azt gondolja, hogy nem számít, hogy zöld vagy piros rágógumit rágnak a gyerekek.

(b)



Note. He initially thought that green chewing gum causes teeth to fall out. What does Robby believe now? Circle the correct answer/Click on the correct answer.

Measuring Test-Takers' Perceptions

To obtain students' views, at the end of the tests, students were asked to indicate how difficult they found the tasks and how much they liked them. The items were rated on a four-point scale (1 = not at all, 2 = not very, 3 = quite a bit, 4 = completely).

https://doi.org/10.33225/jbse/23.22.1050

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Procedure

The paper-based and computer-based data collection took place simultaneously at the school in May 2023 and lasted approximately 20-30 minutes. The paper-based test was administered with a colour test booklet, while the online data collection was conducted in the school's computer labs using the eDIA (Electronic Diagnostic Assessment) system (Csapó & Molnár, 2019).

Data Analysis

Jamovi (version 2.3.26) was utilized for conducting the Rasch model based on item response theory (IRT) and differential item functioning (DIF) analysis. The aim was to examine whether there was a difference between the two test modes at the item level. SPSS (version 27) was employed for conducting descriptive statistical analysis, such as calculating means, standard deviations and percentage, and also for calculating Cronbach's alphas to assess the reliability of the two test modes. Additionally, SPSS was used to perform *t*-tests to analyse differences between genders.

Research Results

Reliability and Validity

To test the reliability of the SK-I, Cronbach's alpha (α) values were computed. Alpha values were found to be similar for both the paper-based (α = .65) and computer-based (α = .70) tests. However, two items did not exhibit a good fit in the reliability analysis. Consequently, considering the item-rest correlations (*Rir*), it was deemed necessary to remove items DAT2 (Rir_{PB} = -.14, Rir_{CB} = .00) and DAT9 (Rir_{PB} = .04, Rir_{CB} = .00) from the analysis. After the necessary corrections – the possible causes of which are discussed in detail in the discussion section – the Cronbach's alphas with the 28 items were as follows: α_{PB} = .71, α_{CB} = .73. These values are found to be acceptable.

Table 1 summarizes the psychometric properties of the SK-I comparing the paper-based and computer-based groups. Considering the percentage of correct answers and item difficulty, it was observed that item 16 (DAT6) proved to be the most challenging in both the paper-based (diff. 0.47) and computer-based (diff. 0.57) tests. This particular item assessed the interpretation of confounded data. However, the results suggested that overall the test was easy for students to complete (all except item 16 have negative values), irrespective of the medium. In this case, the ceiling effect was not a problem, as the test was used to measure the level of students' scientific reasoning by the end of Grade 4.

The Rasch model analysis indicated that the items fit the model well for both test modes. In the paper-based test, the infit values (weighted mean squares, MNSQ) ranged from 0.86 to 1.18 (*M* = 0.99, *SD* = 0.10), whereas in the digital test, they ranged from 0.75 to 1.34 (*M* = 0.99, *SD* = 0.12). To address the question of whether there are any items with unexpected behaviour in the test, a differential item functioning (DIF) analysis was conducted using binary logistic regression in dichotomous items. The DIF analysis revealed no significant difference in the behaviour of any of the items between the two test modes. The same analysis method applied to gender did not identify any items exhibiting unexpected behaviour in any of the test modes.

Table 1
The Psychometric Parameters of the SK-I by Mode of Administration

No.	Item	Correct answer		Difficulty		Infit	
		Paper-based	Computer-based	Paper-based	Computer-based	Paper-based	Computer-based
1	EXP1	82.86	80.00	-1.79	-1.55	1.03	0.88
2	EXP2	85.71	62.22	-2.02	-0.56	0.92	0.92
3	EXP3	85.71	84.44	-2.02	-1.89	0.91	0.86
4	EXP4	62.86	53.33	-0.61	-0.14	0.94	0.93

https://doi.org/10.33225/jbse/23.22.1050

No.	Item	Correct answer		Difficulty		Infit	
		Paper-based	Computer-based	Paper-based	Computer-based	Paper-based	Computer-based
5	EXP5	82.86	86.67	-1.79	-2.09	1.14	0.96
6	EXP6	97.14	93.33	-3.85	-2.91	1.07	1.08
7	EXP7	97.14	97.78	-3.85	-4.09	0.87	1.07
8	EXP8	97.14	93.33	-3.85	-2.91	0.87	0.86
9	EXP9	97.14	97.78	-3.85	-4.09	0.87	0.96
10	EXP10	94.29	91.11	-3.10	-2.58	0.99	0.96
11	DAT1	94.29	84.44	-3.10	-1.89	1.09	0.96
12	DAT2	51.43	35.62	-	-	-	-
13	DAT3	85.71	84.44	-2.02	-1.89	1.00	0.90
14	DAT4	51.43	53.33	-0.07	-0.14	0.86	1.01
15	DAT5	60.00	46.67	-0.47	0.16	1.06	0.97
16	DAT6	40.00	37.78	0.47	0.57	1.10	1.05
17	DAT7	62.86	62.22	-0.61	-0.56	1.18	1.34
18	DAT8	77.14	86.67	-1.39	-2.09	0.96	1.08
19	DAT9	56.41	40.02	-	-	-	-
20	DAT10	85.71	84.44	-2.02	-1.89	1.02	0.75
21	NOS1	74.29	55.56	-1.21	-0.24	1.05	1.14
22	NOS2	85.71	82.22	-2.02	-1.71	0.88	0.87
23	NOS3	65.71	57.78	-0.75	-0.35	0.89	1.10
24	NOS4	97.14	97.78	-3.85	-4.09	0.96	0.99
25	NOS5	91.43	82.22	-2.64	-1.71	1.04	0.92
26	NOS6	60.00	57.78	-0.47	-0.35	1.06	1.03
27	NOS7	82.86	66.67	-1.79	-0.78	1.13	1.25
28	NOS8	91.43	93.33	-2.64	-2.91	0.89	1.10
29	NOS9	82.86	88.89	-1.79	-2.31	1.06	1.03
30	NOS10	85.71	91.11	-2.02	-2.58	1.12	0.90

Descriptive Comparison of the Students' Performance

The average performance of students on the paper-based test was 80.7% (*SD* = 12.4%) and on the computer-based test 77.1% (*SD* = 13.3%). An independent samples *t*-test showed that there was no significant difference ($t(82) = 1.38$, $p = .172$) between the student's performance in the two media. Gender differences were examined by comparing the results of girls and boys using an independent samples *t*-test. The results are being presented in Table 2. It can be said that there was no significant difference between the performance of boys and girls in the paper-based ($t(37) = 0.14$, $p = .889$) or computer-based ($t(43) = 0.69$, $p = .493$) testing modes. Furthermore, there was no significant difference in test modality between the boys' ($t(40) = 0.57$, $p = .573$) and girls' ($t(40) = 1.39$, $p = .172$) groups.

https://doi.org/10.33225/jbse/23.22.1050

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Table 2
Comparison of Student Performance on SK-I by Gender and Test Modality

Group	Paper-based (% correct)		Computer-based (% correct)		*t*	*p*	Cohen's d
	n	*M (SD)*	*n*	*M (SD)*			
All	39	80.7 (12.4)	45	77.1 (13.3)	1.38	.172	0.30
Boys	18	80.7 (12.3)	24	78.2 (13.3)	0.57	.573	0.17
Girls	21	81.1 (12.7)	21	75.4 (13.5)	1.39	.172	0.43
t		0.14		0.69			
p		.889		.493			
Cohen's d		0.05		0.21			

Test-Taker Perceptions

Table 3 displays the perceptions of paper-based and computer-based test-takers. The low mean scores (M_{PB} = 1.32, *SD* = 0.54; M_{CB} = 1.55, *SD* = 0.66) for test difficulty indicated that students found the test easy in both modes, which was consistent with the results of item difficulty. Consequently, it can be concluded that the fourth graders made a realistic assessment of the test's difficulty. The high average scores of test enjoyment (M_{PB} = 3.53, *SD* = 0.66; M_{CB} = 3.47, *SD* = 0.63) indicated that both paper and computer-based respondents found these types of tasks interesting and enjoyable to solve. When comparing test modes, no significant differences were found in either test difficulty ($t(82) = 1.59$, $p = .116$) or enjoyment of the test ($t(82) = 0.43$, $p = .668$).

Table 3
Students' Perceptions Across Modes

Item	Paper-based (*n* = 39)	Computer-based (*n* = 45)	*t*	*p*	Cohen's *d*
	M (SD)	*M (SD)*			
Test difficulty	1.32 (0.54)	1.55 (0.66)	1.59	.116	0.36
Test enjoyment	3.53 (0.66)	3.47 (0.63)	0.43	.668	0.09

Discussion

A key concern in transitioning from a paper-based test to a digital format is always whether the measurement instrument utilized in the new testing environment can maintain similar validity to its paper-based counterpart (Williamson et al., 2017). Consequently, the majority of research questions (RQ1–RQ3) were focused on addressing this matter. In assessing the validity of the measurement, we discovered that the reliability of the SK-I was acceptable for both paper-based and computer-based data collection. The Cronbach's alpha values were nearly identical to those of the instrument developed for the German sample in the same grade during group testing (Osterhaus & Koerber, 2023). This result may validate the efficacy of this data recording method for the SK-I, confirming the hypothesis that by grade 4, students' text comprehension skills have developed to the point where they can independently complete reasoning skills tests.

The two items (DAT2 and DAT9) that exhibited low item-rest correlations measured the interpretation of covariation data, including data of imperfect type. It can be observed that in both the computer- and paper-based versions, these two items were solved with the lowest accuracy by the participants, except for item DAT6 (see Table 1). It can be assumed that, in grade 4, students are not yet proficient in separating the effects of covariation data for different levels of complexity. Therefore, even though they solved the complete test with high scores, these two items (DAT2 and DAT9) proved to be too difficult for them to interpret. The German fourth-graders also made the highest number of mistakes on these two items (Osterhaus & Koerber, 2023), leading to unexpected behaviour and not fitting the pattern of their performance in the complete test. Consequently, this situation impacted the test's reliability.

https://doi.org/10.33225/jbse/23.22.1050

The internal validity of the SK-I, assessed through DIF analysis, confirmed that the two methods of data collection are comparably acceptable. As a result, no item-level differences in media effects were observed between the two testing methods. The reliability and validity of both the paper-based and computer-based test forms were found to be equivalent. Previous studies have also yielded similar results, in which paper-based tests of reasoning skills were converted to a digital format (e.g., Hassler Hallstedt & Ghaderi, 2018; Molnár et al., 2011; Vo & Csapó, 2023).

The sample exhibited a ceiling effect (Staus et al., 2021), with a significant proportion of students attaining high test scores. Nevertheless, this does not pose a problem for the utilization of the test, as the primary objective of the SK-I is to monitor the progression of scientific reasoning. Hence, the high scores indicate that the fourth graders in the study have already reached the expected level of scientific reasoning development for their age. The similarly high scores (75.9%) of German students in the same age group are interesting (Osterhaus & Koerber, 2023).

Before the PISA test was transitioned to an online format, a series of pilot studies were conducted to examine whether there were any differences in student performance when the test medium was altered (e.g., Björnsson, 2008; Jerrim et al., 2018; Kroehne et al., 2019). Jerrim's (2016) study found that, despite a strong correlation between paper-based and digital versions of the math test, significant differences exist in the results. However, it is important to note that while the assessment framework remained the same, certain test items incorporated specific task types that exploited the opportunities provided by digital testing (OECD, 2014). The results showed no significant difference in mean performance when the test medium was switched, indicating that the digital form of the test is suitable for assessing the scientific reasoning of fourth-grade students. Another study, analysing PISA 2006 data, reported that by changing the test medium, boys outperformed girls in scientific literacy in a manner that was not observed in the paper-and-pencil test. This phenomenon was attributed to boys having a lower reading load and experiencing greater test fatigue on low difficulty items in paper-based tests (Halldórsson et al., 2009). However, none of these phenomena were observed. Changing the test medium did not have a significant impact on test performance. This aligns with the initial assumption, as the only change made was in the medium, without altering the test's presentation or structure. While the influence of paper versus digital media on reading comprehension could have affected these results (Delgado et al., 2018), this was not observed in this sample. Furthermore, with the conversion of the test recording mode to digital, the boy's performance in the computer-based test did not surpass that of the girls.

Gender is a crucial variable in assessing scientific reasoning, so one of the research questions concerned gender differences. Several studies confirm that boys tend to perform better on scientific reasoning ability tests (see Ha et al., 2021; Lazonder et al., 2020; Luo et al., 2021), while other studies find no significant differences in reasoning ability (see Koerber et al., 2015; Mayer et al., 2014; Molnár, 2011; Osterhaus & Koerber, 2023). The results of this research support the latter perspective, as we did not find any significant differences in boys' and girls' scientific reasoning across any of the test types.

Some previous studies have indicated that students generally prefer computer-based testing over paper-based testing (e.g., Okocha, 2022; Tella & Bashorun, 2011). Therefore, one of the research questions (RQ4) aimed to examine whether there are differences in students' perceptions of test difficulty and test enjoyment as a function of the test medium. The results suggest that the test medium did not have a significant effect on test takers' perceptions.

Conclusions and Implications

The main aim of this research was to demonstrate that the Hungarian version of the Science-K Inventory can be effectively utilized to evaluate the scientific reasoning (experimentation, data interpretation, understanding the nature of science) of fourth-grade students, not only in a traditional paper-based setting but also in a computer-based testing environment. The novelty of this research was that the SK-I had not yet been tested in a computer-based format. Previously, data collection for third and fourth graders was conducted in a paper-based group setting with test assistants who read out the instructions for the visually presented tasks. It was therefore questionable whether unassisted computer-based data collection could be employed in this particular age group. Based on these results, it can be concluded that the test-taking method used is suitable for fourth graders and that both the paper-based and digital-based tests exhibit comparable reliability and validity indicators. As a result, both modes are equally valid for assessing fourth graders' scientific reasoning.

The study carries several implications. The development of scientific reasoning and research skills is crucial in the teaching and learning of science. However, there is a lack of reliable measures available to assess young children's understanding of scientific inquiry and scientific reasoning skills. Therefore, the introduction of reliable measurement tools such as the SK-I into teaching practice that allow the monitoring of changes in students' reasoning and individual differences in this area is valuable.

https://doi.org/10.33225/jbse/23.22.1050

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Computer-based testing offers several advantages that can be exploited, such as the facilitation of large-sample testing by reducing resource requirements. It can also provide rapid and valuable information for teachers to aid the planning of the development of scientific literacy competence. Furthermore, the research results can contribute to the design of programs aimed at improving scientific literacy and serve as an early diagnostic tool to better understand the decline in students' performance in international science assessments, such as PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study).

A limitation of this study is the small sample size, which limits the ability to obtain more robust item parameter estimates for Rasch analysis, and to test the construct validity of SK-I by confirmatory factor analysis (CFA). However, future research aims to carry out a study with an expanded sample size to acquire more precise model fit statistics for the Rasch model and CFA, along with more accurate item parameters. Also, it may be worth investigating whether the observed item fit problems persist in the case of a larger sample. If so, exploring the contextual issues or curricular and cultural differences that give rise to these problems could indicate a need for revisions to the Hungarian version of the SK-I. Additionally, there are plans to test SK-I with younger children, which, when narrated by voice, could be used to assess the scientific reasoning of first and second graders and thus conduct a longitudinal study. Exploration of the relationship between scientific reasoning and additional variables, including school performance, attitudes towards school subjects, motivation to learn science, and social background, is also part of the next research agenda.

Acknowledgements

Many thanks to the students who participated in the study and to the teachers who organised the data collection.

Funding

This study was funded by the Research Programme for Public Education Development of the Hungarian Academy of Sciences (KOZOKT2021-16) and supported by the ÚNKP-22-3-SZTE-61 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

Declaration of Interest

The authors declare no competing interest.

References

Adanır, G. A., Akmatbekova, A., & Muhametjanova, G. (2020). Longitudinal study of Kyrgyz students' general physics course performance in paper-based versus online–based tests. *Education and Information Technologies, 25,* 4197–4210. https://doi.org/10.1007/s10639-020-10175-7

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning* (ETS-RM-03-05). Educational Testing Service.

Björnsson, J. K. (2008). Changing Icelandic national testing from traditional paper and pencil to computer-based assessment: Some background, challenges, and problems to overcome. In F. Scheuermann & A. Guimaraes Pereira (Eds.), *Towards a research agenda in computer-based assessment: Challenges and needs for European Educational Measurement* (pp. 6–9). European Communities.

Brallier, S. A., Schwanz, K. A., Palm, L. J., & Irwin, L. N. (2015). Online testing: Comparison of online and classroom exams in an upper-level psychology course. *American Journal of Educational Research, 3*(2), 255–258. https://doi.org/10.12691/education-3-2-20

Bullock, M., & Ziegler, A. (1999). Scientific reasoning: Developmental and individual differences. In F. E. Weinert, & W. Schneider (Eds.), *Individual development from 3 to 12: Findings from the Munich longitudinal study* (pp. 38–54). Cambridge University Press.

Bullock, M., Sodian, B., & Koerber, S. (2009). Doing experiments and understanding science: Development of scientific reasoning from childhood to adulthood. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood: Findings from a 20-year longitudinal study* (pp. 173–197). Psychology Press.

Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school-readiness and reasoning skills. *Journal of Educational Psychology, 106*(2), 639–650. https://doi.org/10.1037/a0035756

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

Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review, 25*, 23–38. https://doi.org/10.1016/j.edurev.2018.09.003

Dunbar, K. & Fugelsang, J. (2005). Scientific thinking and reasoning. In Holyoak, K. J. & Morrison, R. G. (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 705–725). University of California.

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

https://doi.org/10.33225/jbse/23.22.1050

Ha, M., Sya'bandari, Y., Rusmana, A. N., Aini, R. Q., & Fadillah, S. M. (2021). Comprehensive analysis of the FORT instrument: Using distractor analysis to explore students' scientific reasoning based on academic level and gender difference. *Journal of Baltic Science Education, 20*(6), 906–926. https://doi.org/10.33225/jbse/21.20.906

Halldórsson, A. M., McKelvie, P., & Björnsson, J. K. (2009). Are Icelandic boys really better on computerized tests than conventional ones? In F. Scheuermann & J. K. Björnsson (Eds.), *The transition to computer-based assessment* (pp. 178–193). JRC Scientific and Technical Report EUR 23679 EN. Office for Official Publications of the European Communities.

Hassler-Hallstedt, M., & Ghaderi, A. (2018). Tablets instead of paper-based tests for young children? Comparability between paper and tablet versions of the mathematical Heidelberger Rechen Test 1-4. *Educational Assessment*, *23*(3), 195–210. https://doi.org/10.1080/10627197.2018.1488587

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. Basic Books.

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy and Practice*, *23,* 495–518. https://doi.org/10.1080/0969594X.2016.1147420

Jerrim, J., Micklewright, J., Heine, J. H., Sälzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education, 23*, 476–493. https://doi.org/10.1080/03054985.2018.1430025

Kuhn, D. (2011). What is scientific thinking and how does it develop? In Goswami, U. (Eds.), *Handbook of childhood cognitive development* (pp. 497–523). Wiley-Black-well. https://doi.org/10.1002/9781444325485.ch19

Kuhn, D., Amsel, E. & O'Loughlin, M. (1988). *The development of scientific thinking skills*. Academic Press.

Koerber, S., & Sodian, B. (2009). Reasoning from graphs in young children. Preschoolers' ability to interpret and evaluate covariation data from graphs. *Journal of Psychology of Science and Technology, 2*(2), 73–86. https://doi.org/10.1891/1939-7054.2.2.73

Koerber, S., Osterhaus, C., & Sodian, B. (2015). Testing primary-school children's understanding of the nature of science. *British Journal of Developmental Psychology, 33*(1), 57–72. https://doi.org/10.1111/bjdp.12067

Koerber, S., & Osterhaus, C. (2019). Individual differences in early scientific thinking: Assessment, cognitive influences, and their relevance for science learning. *Journal of Cognition and Development 20*(4), 510–533. https://doi:10.1080/15248372.2019.1620232

Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (2019). Construct equivalence of PISA reading comprehension measured with paper-based and computer-based assessments. *Educational Measurement: Issues and Practice, 38*, 97–111. https://doi.org/10.1111/emip.12280

Lawson, A. E. (2004). The nature and development of scientific reasoning: A synthetic view. *International Journal of Science and Mathematics Education, 2*, 307–338. https://doi.org/10.1007/s10763-004-3224-2

Lazonder, A. W., & Kamp, E. (2012). Bit by bit or all at once? Splitting up the inquiry task to promote children's scientific reasoning. *Learning and Instruction, 22*(6), 458–464. https://doi.org/10.1016/j.learninstruc.2012.05.005

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

Luo, M., Sun, D., Zhu, L., & Yang, Y. (2021). Evaluating scientific reasoning ability: Student performance and the interaction effects between grade level, gender, and academic achievement level. *Thinking Skills and Creativity, 41*, 100899. https://doi.org/10.1016/j.tsc.2021.100899

Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning and Instruction, 29*, 43–55. https://doi.org/10.1016/j.learninstruc.2013.07.005

Molnár, G. (2011). Playful fostering of 6- to 8-year-old students' inductive reasoning. *Thinking Skills and Creativity, 6*, 91–99. https://doi.org/10.1016/j.tsc. 2011.05.002

Molnár, G., R. Tóth, K., & Benő Csapó, B. (2011, April 8–12). *Comparing paper-based and computer-based testing in the first grade.* [Conference presentation]. 2011 AERA Annual Meeting, New Orleans, Louisiana, USA.

Morris, B. J., Croker, S., Masnick, A. M., & Zimmerman, C. (2012). The emergence of scientific reasoning. In H. Kloos, B. J. Morris & J. Amaral (Eds.), *Current topics in children's learning and cognition* (pp. 61–82). InTech. https://doi.org/10.5772/53885

Nikou, S. A., & Economides, A. A. (2016). The impact of paper-based, computer-based and mobile-based self-assessment on students' science motivation and achievement. *Computers in Human Behavior*, *55*(Part B), 1241–1248. https://doi.org/10.1016/j.chb.2015.09.025

Nyberg, K., Koerber, S., & Osterhaus, C. (2020). How to measure scientific reasoning in primary school: A comparison of different test modalities. *European Journal of Science and Mathematics Education 8*(3), 136–144. https://doi.org/10.30935/scimath/9552

O'Connor, G., Fragkiadaki, G., Fleer, M., & Rai, P. (2021). Early childhood science education from 0 to 6: a literature review. *Education Sciences*, *11*(4), 178. https://doi.org/10.3390/educsci11040178

OECD (2010). PISA computer-based assessment of student skills in science. OECD Publishing. https://doi.org/10.1787/9789264082038-en

OECD (2014). PISA 2012 results: What students know and can do – Student performance in mathematics, reading and science (Volume I, revised edition, February 2014). OECD Publishing. https://doi.org/10.1787/9789264201118-en

OECD (2019). PISA 2018 Technical Report. OECD Publishing.

Okocha, F. (2022). Student perception of computer-based testing in Kwara State, Nigeria. *International Journal of Web-Based Learning and Teaching Technologies*, *17*(1) 1–11. http://doi.org/10.4018/IJWLTT.294575

Osterhaus, C., & Koerber, S. (2023). The complex associations between scientific reasoning and advanced theory of mind. *Child Development, 94*(1), 18–42. https://doi.org/10.1111/cdev.13860

Osterhaus, C., Lin, X., & Koerber, S. (2023). Measuring scientific reasoning in kindergarten and elementary school: Validating the Chinese version of the Science-K Inventory. *Educational Research for Policy and Practice.* https://doi.org/10.1007/s10671-023-09332-9

Pásztor, A., Molnár, G., & Csapó, B. (2015). Technology-based assessment of creativity in educational context: the case of divergent thinking and its relation to mathematical achievement. *Thinking Skills and Creativity*, *18*, 32–42. https://doi.org/10.1016/j.tsc.2015.05.004

ISSN 1648–3898 /Print/
ISSN 2538–7138 /Online/

Peteranderl, S., & Edelsbrunner, P. A. (2020). The predictive value of children's understanding of indeterminacy and confounding for later mastery of the control-of-variables strategy. *Frontiers in psychology, 11*, 531565. https://doi.org/10.3389/fpsyg.2020.531565

Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science, 323*(5910), 75–79. https://doi.org/10.1126/SCIENCE.1168046

Rasch, G. (1960). *Probabilistic model for some intelligence and achievement tests.* Danish Institute for Educational Research.

Rosen, Y., & Tager, M. (2014). Making student thinking visible through a concept map in computer-based assessment of critical thinking. *Journal of Educational Computing Research, 50*, 249–270. https://doi.org/10.2190/EC.50.2.f

Rosen, Y. (2015). Computer-based assessment of collaborative problem solving: Exploring the feasibility of human-to-agent approach. *International Journal Artificial Intelligence in Education, 25*, 380–406. https://doi.org/10.1007/s40593-015-0042-3

Schroeders, U., Bucholtz, N., Formazin, M., & Wilhelm, O. (2013). Modality specificity of comprehension abilities in the sciences. *European Journal of Psychological Assessment*, *29*(1). https://doi.org/10.1027/1015-5759/a000114

Schulz, L., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology, 40*(2), 162–176. http://doi.org/10.1037/0012-1649.40.2.162

Schwichow, M., Osterhaus, C., Edelsbrunner, P. A. (2020). The relation between the control-of-variables strategy and content knowledge in physics in secondary school. *Contemporary Educational Psychology, 63*. https://doi.org/10.1016/j.cedpsych.2020.101923

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, *33*, 1–19. https://doi.org/10.1111/jcal.12172

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development, 62*(4), 753–766. https://doi.org/10.2307/1131175

Staus, N. L., O'Connell, K., & Storksdieck, M. (2021). Addressing the ceiling effect when assessing STEM out-of-school time experiences. *Frontiers in Education*, *6,* 690431. https://doi.org/10.3389/feduc.2021.690431

Strand-Cary, M., & Klahr, D. (2008). Developing elementary science skills: instructional effectiveness and path independence. *Cognitive Development, 23*, 488–511. http://dx.doi.org/10.1016/j.cogdev.2008.09.00

Tella, A., & Bashorun, M. (2012). Attitude of undergraduate students towards computer-based test (CBT): A case study of the University of Ilorin, Nigeria. *International Journal of Information and Communication Technology Education, 8*(2), 33–45. https://doi.org/10.4018/jicte.2012040103

Vo, D. V., & Csapó, B. (2023). Effects of multimedia on psychometric characteristics of cognitive tests: A comparison between technology-based and paper-based modalities. *Studies in Educational Evaluation, 77*. https://doi.org/10.1016/j.stueduc.2023.101254

Williamson, K. C., Williamson, V. M. & Hinze, S. R. (2017). Administering spatial and cognitive instruments in-class and on-line: Are these equivalent? *Journal of Science Education and Technology 26*, 12–23. https://doi.org/10.1007/s10956-016-9645-1

Wu, H., & Molnár G. (2018). Computer-based assessment of Chinese students' component skills of problem Solving: A pilot study. *International Journal of Information and Education Technology, 8*(5), 381– 356. https://doi.org/10.18178/ijiet.2018.8.5.1067

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 J. T. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 223–248). Wiley & Sons. https://doi.org/10.1002/9781119170174.epcn407

Received: *August 24, 2023* Revised: *October 06, 2023* Accepted: *November 08, 2023*

Cite as: Nagy, M. T., & Korom, E. (2023). Measuring scientific reasoning of fourth graders: Validation of the Science-K Inventory in paper-based and computer-based testing environments. *Journal of Baltic Science Education*, *22*(6), 1050-1062. https://doi.org/10.33225/jbse/23.22.1050

Mário Tibor Nagy
(Corresponding author)
PhD Student, Doctoral School of Education, MTA–SZTE Digital Learning Technologies Research Group, University of Szeged, Petőfi sgt. 32–34, H–6722 Szeged, Hungary.
E-mail: nagy.mario.tibor@edu.u-szeged.hu
ORCID: https://orcid.org/0000-0001-6465-3797

Erzsébet Korom
PhD, Associate Professor, Department of Learning and Instruction, Institute of Education, MTA–SZTE Digital Learning Technologies Research Group, University of Szeged, Petőfi sgt. 32–34, H–6722 Szeged, Hungary.
E-mail: korom@edpsy.u-szeged.hu
ORCID: https://orcid.org/0000-0001-9534-8146