Article

Individual Differences in Children's Scientific Reasoning

Erika Schlatter [1,*], Ard W. Lazonder [1], Inge Molenaar [1] and Noortje Janssen [2]

[1] Behavioural Science Institute, Radboud University, 6525 XZ Nijmegen, The Netherlands; ard.lazonder@ru.nl (A.W.L.); inge.molenaar@ru.nl (I.M.)

[2] Department of Instructional Technology, University of Twente, 7522 NB Enschede, The Netherlands; noortje.janssen@ru.nl

[*] Correspondence: erika.schlatter@ru.nl

Abstract: Scientific reasoning is an important skill that encompasses hypothesizing, experimenting, inferencing, evaluating data and drawing conclusions. Previous research found consistent inter- and intra-individual differences in children's ability to perform these component skills, which are still largely unaccounted for. This study examined these differences and the role of three predictors: reading comprehension, numerical ability and problem-solving skills. A sample of 160 upper-primary schoolchildren completed a practical scientific reasoning task that gauged their command of the five component skills and did not require them to read. In addition, children took standardized tests of reading comprehension and numerical ability and completed the Tower of Hanoi task to measure their problem-solving skills. As expected, children differed substantially from one another. Generally, scores were highest for experimenting, lowest for evaluating data and drawing conclusions and intermediate for hypothesizing and inferencing. Reading comprehension was the only predictor that explained individual variation in scientific reasoning as a whole and in all component skills except hypothesizing. These results suggest that researchers and science teachers should take differences between children and across component skills into account. Moreover, even though reading comprehension is considered a robust predictor of scientific reasoning, it does not account for the variation in all component skills.

Keywords: scientific reasoning; primary education; individual differences

1. Introduction

Science education is an important part of the curriculum in many countries [1,2]. Starting in primary school, children learn about the underlying principles and causal relationships of science domains as well as the processes through which this knowledge is created. This process of intentional knowledge-seeking is known as scientific reasoning [3,4] and is important for children because it prepares them for a society where science and the outcomes of scientific research are embedded in the culture [5]. In a school setting, scientific reasoning skills are particularly important for successful inquiry learning: 'minds-on' scientific reasoning skills [6] are instrumental to achieving meaningful outcomes from a 'hands-on' inquiry.

Scientific reasoning consists of multiple component skills, namely, hypothesizing, experimenting and evaluating evidence, the latter of which can be further divided into inferencing, evaluating data and drawing conclusions [7,8]. These component skills emerge at a different age, tend to develop at a different pace and are known to vary greatly between same-age children (e.g., [9]). However, most existing research either treats scientific reasoning as a unitary construct or looks at one specific component of scientific reasoning—most often experimenting [10]. Therefore, the inter- and intra-individual differences are not yet well understood, and to this date, few guidelines exist for addressing these differences in primary science classrooms.

An important challenge in understanding individual differences in scientific reasoning is the valid measurement of its component skills. Even though scientific reasoning is

often taught in hands-on settings, it is mostly measured by paper-and-pencil tests. As performance-based testing circumvents many of the problems typically associated with written tests (see, for an overview, Harlen [11]), it might shed new light on the development of scientific reasoning in children. Using one such performance-based test, the current study set out to advance our insight into children's proficiency in different component skills of scientific reasoning, when applied in a practical, coherent inquiry setting in order to ultimately aid the development of teaching materials for various groups of learners in primary education.

1.1. Variation in Scientific Reasoning

As mentioned above, scientific reasoning comprises the skills of hypothesizing (the articulation of ideas about possible outcomes of an investigation), experimenting (the skills to design and perform experiments to test these hypotheses) and evaluating evidence (i.e., drawing valid conclusions). Evidence evaluation, in turn, involves inferencing (making a verbal interpretation of the gathered data), evaluating data (assessing measurement quality, for instance, to decide whether there are enough data to base a conclusion on), and drawing conclusions (using this information to make causal statements to answer the research question).

This multidimensionality is confirmed by psychometric models [12] and studies investigating one or more component skills point to substantial variation. Experimenting, for example, is relatively easy for children to learn: most pre-schoolers are capable of some systematic testing [13,14], and older children can be taught this skill successfully by both direct instruction [15,16] and guided inquiry [17,18]. Hypothesizing is more difficult for young children to learn [9,13,19], whilst evidence evaluation is the most difficult skill for them to acquire [4] and is also experienced as such [20].

Most of the studies on which this tentative order of difficulty is based examined a single skill at a single time point. A positive exception is the study by Piekny and Maehler [9], who inferred the age at which children learn hypothesizing, experimenting and evidence evaluation from cross-sectional data collected with children from kindergarten to grade 5 and found a similar build-up as described above. Still, this study used different types of tasks for the different component skills rather than one task that encompassed all component skills. Thus, the relative difficulty of the component skills of scientific reasoning is not fully understood yet.

Other studies indicate that not all children develop scientific reasoning proficiency at the same pace. In a large-scale cross-sectional study using written tests in grades 4–6, Koerber, Mayer, Osterhaus, Schwippert and Sodian [21] distinguished between naïve, intermediate and advanced conceptions of scientific reasoning and found that, although older children more often had advanced conceptions and less often naïve conceptions than younger children, all proficiency levels were present at all participating grade levels. The results of Piekny and Maehler [9] further suggest that this variation increases with age. For example, both the means and standard deviations of 'hypothesizing' were low in kindergarten but increased from grade 1 onward. This finding indicates that, although children's hypothesizing skills grow, the inter-individual variation increases accordingly. Thus, although children improve in scientific reasoning over the years, not all children improve equally or at the same time as their peers. Acknowledging and understanding these differences is vital for good science education.

To conclude, the component skills of scientific reasoning improve considerably during the primary school years [9,21], albeit with substantial variation. As not all subskills emerge at the same point in time and not all children develop their scientific reasoning proficiency at the same pace, the teaching of scientific reasoning in primary education is a challenging task. A profound understanding of how the component scientific reasoning skills develop can help teachers make scientific reasoning accessible for all children.

1.2. Explaining Variation in Scientific Reasoning

Although differences in the development of scientific reasoning are known to exist, the roots of the differences between children as well as differences in developmental patterns within children (i.e., differences across skills) are less clear. Children's cognitive characteristics account for part of the variation in scientific reasoning proficiency. Previous research provides evidence that reading comprehension, numerical ability and problem-solving skills contribute to scientific reasoning.

Reading comprehension most consistently explains children's overall scientific reasoning performance on written tests [21,22] as well as their ability to set up unconfounded experiments using the Control-of-Variables Strategy [23,24]. Van de Sande, Kleemans, Verhoeven and Segers [25] found that reading comprehension explained the variance in all component scientific reasoning skills, albeit not to the same extent: effect sizes ranged from medium ($r = 0.30$) for experimentation and drawing conclusions to large ($r = 0.47$) for hypothesis validation. Why reading comprehension is such a strong predictor is not entirely clear. Possibly, reasoning ability transcends the domains of reading and science [24,25], or a general understanding of the language of science is important for science learning [26]. However, it is also possible that the influence of reading comprehension is a consequence of test item format: most of the studies cited above used written tests that likely call upon children's reading skills, even though questions were sometimes read out loud. In light of these findings, reading comprehension can be considered an important predictor of scientific reasoning, but because past research heavily relied on the use of paper and pencil tests, further scrutiny of its role is warranted.

Numerical ability is often named as a prerequisite for scientific reasoning by national curriculum agencies [27,28] as well as scientists [29]—likely because scientific reasoning, in particular the evidence evaluation skills, involves reasoning about numerical data [9,22,30,31]. Yet, empirical evidence for this relation is scarce. Early work by Bullock and Ziegler [32] demonstrated that numerical intelligence predicts the growth of experimentation skills in primary schoolchildren, explaining almost 35 percent of the variance in a quadratic growth model. More recent studies found significant correlations between numerical ability and scientific reasoning [10,33]. However, as the latter studies treated scientific reasoning as a unitary construct, it is yet unclear whether numerical ability also predicts children's scientific reasoning, and if so, if it predicts all component skills of scientific reasoning to the same extent.

Children's problem-solving skill is another possible predictor of scientific reasoning. Klahr and Dunbar [34] characterized scientific reasoning as a process of rule induction, which inherently involves problem-solving. One could even argue that scientific reasoning is a form of problem-solving in itself: the problem is a need for specific knowledge, which is resolved through a systematic process of knowledge-seeking. Furthermore, as with the previous predictors, it seems plausible that problem-solving calls upon a person's reasoning skills and therefore predicts scientific reasoning. Although upper-primary schoolchildren are still incapable of formal abstract reasoning, they can solve problems that involve reasoning with concrete objects such as the nine-dots problem and the Tower of Hanoi [35]. Recent research supports these ideas: Mayer et al. [22] found that problem-solving predicted a substantial portion of the variance in children's scientific reasoning. Van de Sande et al. [25] further showed that this effect does not apply to all subskills: hypothesis validation and experimenting depended on problem-solving, whereas generating conclusions did not. As such, problem-solving may explain some but not all component scientific reasoning skills, and the extent to which the different component skills are predicted is yet unclear.

1.3. Research Questions and Hypotheses

Although the cited literature points to notable differences in children's scientific reasoning, most studies either addressed scientific reasoning as a single, albeit multifaceted construct or examined one of its subskills in isolation. Furthermore, most extant research

has been conducted using written tests. These instruments neither resemble the learning context nor scientific practice and therefore may not accurately gauge children's true ability in scientific reasoning [36]. Moreover, written tests of scientific reasoning can confound with reading comprehension, as children with better reading comprehension might perform better on such tests because the test itself involves reading. In order to extend our understanding of the relations between scientific reasoning and the cognitive characteristics discussed above, the subskills should be studied in tandem, preferably in an authentic whole-task setting that does not require children to read.

This study, therefore, aimed to identify and explain differences in children's ability to reason scientifically by means of a performance-based task so as to maximize authenticity and minimize the influence of reading skills. A sample of 160 upper-primary schoolchildren performed this task to gauge their proficiency in five scientific reasoning skills: hypothesizing, experimenting, making inferences, evaluating data and drawing conclusions. Performance differences were related to reading comprehension, numerical ability and problem-solving skills in order to answer the following research questions:

1. What amount of variation can be found in children's scientific reasoning?
2. To what extent is this variation explained by reading comprehension, numerical ability and problem-solving skills?

Based on previous research using written tests, it was expected that children would differ considerably in their overall scientific reasoning proficiency. Differences across the five subskills were also predicted to occur. Specifically, children were expected to be most proficient in experimentation, less proficient in hypothesizing and least proficient in the three evidence evaluation skills (inferencing, evaluating data and drawing conclusions). Reading comprehension, numerical ability and problem-solving skills were expected to explain a unique portion of the variance in scientific reasoning. Considering the alleged differences across subskills, these characteristics were expected to have differential effects.

2. Materials and Methods

2.1. Participants

A sample of 166 children attending the two highest grades of a primary school in a suburban area of the Netherlands participated in this study. Ages ranged from 8 years 11 months to 12 years 8 months. About 80% of the parents held a degree from a research university or university of applied sciences, and almost all children had at least one parent who was born in the Netherlands. Complete data were obtained for 160 of the 166 participating children (54% boys, M_{age} = 11 years 0 months, SD = 9 months); 84 of these children were in grade 5 (52% boys, M_{age} = 10 years 5 months, SD = 7 months) and 76 of them in grade 6 (55% boys, M_{age} = 11 years 7 months, SD = 6 months).

The school participated in a large-scale longitudinal research project that was approved by the ethics committee of the Faculty of Behavioural, Management and Social Sciences of the University of Twente. All participating children had passive parental consent, meaning that parents were informed and did not object to their child's participation in the study. The findings reported here were gathered during the third wave of data collection, which means that the sample was familiar with most tests. The school's science curriculum contained five annual hands-on science projects which enabled children to practice their scientific reasoning.

2.2. Materials

2.2.1. Scientific Reasoning Task

Children's scientific reasoning skills were gauged during a 20 min performance-based scientific reasoning task under supervision of a test administrator [19]. The task contained 15 questions and assignments (hereafter referred to as 'items'), 3 for each component scientific reasoning skill, which were organized in four inquiry cycles of increasing difficulty (for example, see Table 1). The task was administered orally in order to minimize the effects of reading and writing ability, and handouts were used to ensure uniformity in the data

children used to make inferences, evaluate data and draw conclusions. Children's answers and actions were registered by the test administrator for later scoring. Each of the items was worth one point, and a child could thus earn a maximum of three points per subskill. Total test scores could range from 0 to 15 points. The Cohen's κ inter-rater agreement of the answer scoring was 0.84.

Table 1. Example inquiry cycle.

Subskill [1]	Question
Experimenting	Can you figure out if it matters whether the surface is hard or soft? So you can be sure whether the ball without the mat bounces more, less or as much as with the extra mat.
Inferencing	In this box, the outcome was 'tick-tick', and in this box, the outcome was 'tick-tick-tick-tick'. Can you explain what the outcome of the experiment was?
Drawing conclusions	Can you be sure that balls *always* bounce more often on a hard surface?
Hypothesizing	The student who is next will also be completing this experiment. Imagine that student asks you to predict what the outcomes of their experiment will be. What would you say?

[1] This is an example of the first inquiry cycle of the bouncing balls version of the test. In other versions, only the variables would be different. Evaluating data were assessed in subsequent research cycles.

Three versions of this task were available, which differed exclusively with regard to the topic of investigation. In the *rolling balls* version, adapted from Chen and Klahr [16], children interacted with two inclined planes to find out how four dichotomous input variables (slope, starting point, surface and mass of the ball) influenced the distance balls travel after leaving a ramp. In the *bouncing ball* version, children investigated how four dichotomous variables (starting height, surface, mass of the ball and whether the ball was solid) affected the number of times a ball would bounce; the *cars* version had children set four features of rubber-band-powered toy cars (size of back wheels, axle size, diameter of the rubber band and tightness of the winding of the rubber band) in order to examine how far a car drives.

Children were assigned to the version they had not received in previous waves of data collection, and scores did not differ significantly between the three versions, $F(2, 157) = 0.08$, $p = 0.925$. Furthermore, a validation study [19] showed no effects of prior domain knowledge on the performance of any of the versions. This study also demonstrated that the test scores conform to a two-parameter Item-Response theory model and have an acceptable expected a posteriori (EAP) reliability of 0.59. As the component skills were each assessed by only three items of increasing complexity, internal consistency of the subscales could not meaningfully be calculated.

2.2.2. Reading Comprehension Test

Reading comprehension was measured by a standardized progress evaluation measure developed by Cito, the Dutch national testing agency [37]. Different versions are available for different grades, and the test has a measurement accuracy between 0.87 and 0.89 [37]. In all versions of the test, children had to read different types of mostly pre-existing texts, such as short stories, newspaper articles, advertisements and instruction manuals. The test consisted of 55 multiple choice items that, for example, required children to fill in the blanks, explain what a particular line in the text means or choose an appropriate continuation of a story. As participants in the current study were drawn from different grades, the version corresponding to their grade level was administered. The One Parameter Logistic Model [38] was used to transform children's answers into a person proficiency score that can be meaningfully compared across grades.

2.2.3. Numerical Ability Test

Numerical ability was gauged by a standardized progress evaluation measure that required children to add, subtract, multiply or divide one- and two-digit numbers by heart [39]. The test consists of 200 items of increasing difficulty and is highly reliable ($\alpha = 0.97$). Children worked on the test for 5 min and obtained 1 point for each correct answer.

2.2.4. Problem-Solving Test

A digital version of the Tower of Hanoi (adapted from Welsh [40]) was developed to assess children's problem-solving skills. The test required children to solve as many problems as they could in 7 min. One point was awarded for each solved problem, and reliability was high ($\alpha = 0.85$). The 20 problems required children to move differently sized disks from their starting position to their target position on the rightmost peg. Three simple rules limited the possible moves children could make: only one disk could be moved at a time, the disk could only be moved to an adjacent peg and it could never be placed on top of a smaller disk. The starting position differed per problem in order to assure a gradual increase from a minimum of 3 moves to solve the puzzle at Problem 1 to a minimum 15 moves at Problem 19. The target solution for each of the problems was a three- or four-disk tower on the rightmost peg. In order to prevent trial-and-error and provide children with an opportunity for a fresh start if they had trouble solving a certain problem, each unsolved puzzle would be automatically reset after 20 moves were made. Manual reset was not possible. To ensure that children would not finish the task ahead of time, the final problem was a 5-disk, 31-move problem. In practice, none of the children reached this final problem.

2.2.5. Procedure

Children were tested in their regular classrooms. First, teachers administered the reading comprehension and numerical ability tests on a whole-class basis, using the guidelines provided by the test publishers. When standardized testing was completed, the researchers administered the problem-solving test and the scientific reasoning task. The problem-solving test was administered in small groups. After a short explanation, children worked on the test for 7 min. The scientific reasoning task was administered individually and lasted about 20 min per child.

2.2.6. Data Analysis

Data were analyzed using IBM SPSS 25. In order to answer the first research question, variation in scientific reasoning was explored using descriptive statistics; relations between the five scientific reasoning subskills were analyzed using Pearson correlations and a within-subject analysis of variance (ANOVA), controlled for grade and gender. The second research question, which sought to reveal what accounts for the observed differences in scientific reasoning, was answered by means of correlational analyses and multivariate multiple regression analysis.

Table 2 presents the descriptive statistics of children's test performance. Preliminary analyses of three predictor skills indicated that the sixth-graders outperformed the fifth-graders in reading comprehension, $F(1, 158) = 14.18$, $p < 0.001$, partial $\eta^2 = 0.08$, numerical ability, $F(1, 158) = 8.02$, $p = 0.005$, partial $\eta^2 = 0.05$ and problem-solving, $F(1, 158) = 4.35$, $p = 0.039$, partial $\eta^2 = 0.03$. The cross-grade differences in scientific reasoning were minor, and were tested for statistical significance in the main analysis reported below.

Table 2. Descriptive statistics of children's test scores.

Test Scores	Grade 5		Grade 6		Entire Sample	
	M	**SD**	**M**	**SD**	**M**	**SD**
Scientific reasoning	8.01	2.23	7.99	2.24	8.00	2.23
Hypothesizing	2.15	0.86	2.01	0.82	1.77	0.88
Experimenting	1.54	0.63	1.50	0.64	2.09	0.84
Inferencing	1.19	0.63	1.37	0.73	1.52	0.63
Evaluating data	1.32	0.95	1.38	0.80	1.28	0.68
Drawing conclusions	1.81	0.86	1.72	0.90	1.35	0.88
Reading comprehension	52.40	12.58	61.91	18.98	56.92	16.59
Numerical ability	84.42	19.84	95.26	28.23	89.57	24.72
Problem-solving	11.39	2.92	12.33	2.74	2.87	0.17

3. Results

In order to determine the extent to which scientific reasoning ability differs between children, the means and standard deviations of children's test scores were examined. Overall test scores ranged from 2 to 13 points with an average of 8.00 (*SD* = 2.23). Scores on the subskills ranged from 0 to 3 except for inferencing, where the minimum score was 1 point. Means and standard deviations confirmed this differential ability and warranted further exploration as to what could explain this difference in scientific reasoning proficiency.

The mean scores in Table 2 point to variation in proficiency on the different subskills: on average, children appeared to be most proficient in experimenting and least proficient in evaluating data and drawing conclusions, while hypothesizing and inferencing held the middle ranks. A within-subject ANOVA, controlling for gender and grade, was conducted to test whether these differences were statistically significant. Multivariate results revealed an overall effect of subskill (Pillai's trace = 0.46, $F(4, 153) = 32.50$, $p < 0.001$), but no interaction effects of subskill with gender (Pillai's trace = 0.02, $F(4, 153) = 0.60$, $p = 0.665$), and grade (Pillai's trace = 0.03, $F(4, 153) = 1.23$, $p = 0.300$). The differences between subskills were further explored in univariate analyses. Scores on experimenting were significantly higher than scores on all other subskills ($p < 0.01$). Scores on hypothesizing were significantly higher than scores on inferencing, evaluating data and drawing conclusions ($p < 0.05$). Scores on inferencing were significantly higher than scores on evaluating data ($p < 0.01$), but not scores on drawing conclusions ($p = 0.214$). Drawing conclusions and evaluating data, the two subskills with the lowest scores, were not significantly different from one another ($p = 0.993$).

Having established that there is variation in the extent to which children master the five scientific reasoning subskills, the next set of analyses sought to explain these differences from children's reading comprehension, numerical ability and problem-solving skills. As shown in Table 3, the total scientific reasoning score correlated with all three factors, albeit moderately. Correlations at the subskill level paint a mixed picture. Reading comprehension was associated with all subskills except hypothesizing, numerical ability only correlated with evaluating data and problem-solving did not correlate with any of the subskills.

Multivariate multiple regression was used to further scrutinize the relations between the three predictor variables and the five scientific reasoning subskills. Multivariate test results showed no main effect for the control variables gender, Pilai's trace = 0.01, $F(5, 150) = 0.35$, $p = 0.882$, partial $\eta^2 = 0.01$, and grade, Pilai's trace = 0.05, $F(5, 150) = 1.60$, $p = 0.164$, partial $\eta^2 = 0.51$. Regarding the explanatory variables, a significant contribution of reading comprehension on scientific reasoning was found, Pilai's trace = 0.17, $F(5, 150) = 6.28$, $p < 0.001$, partial $\eta^2 = 0.17$. Neither numerical ability, Pilai's trace = 0.02, $F(5, 150) = 0.57$, $p = 0.725$, partial $\eta^2 = 0.02$, nor problem-solving skills, Pilai's trace = 0.02, $F(5, 150) = 0.61$, $p = 0.694$, partial $\eta^2 = 0.02$, explained scientific reasoning to a significant degree. The between-subject effects of reading comprehension in Table 4 showed that reading comprehension accounted for a significant proportion of the variance in experi-

menting, inferencing, evaluating data and drawing conclusions, but not in hypothesizing. The regression coefficients further indicate that experimenting was most influenced by reading comprehension. Of the significantly predicted subskills, inferencing was least influenced by reading comprehension. Thus, although reading comprehension remains an important explanatory factor, it did not explain all scientific reasoning subskills uniformly.

Table 3. Correlations for predictors and scientific reasoning subskills.

	1	2	3	4	5	6	7	8	9
1. Scientific reasoning (total score)	—								
2. Hypothesizing	0.61 **	—							
3. Experimenting	0.58 **	0.15	—						
4. Inferencing	0.53 **	0.18 *	0.11	—					
5. Evaluating data	0.46**	0.14	0.14	0.06	—				
6. Drawing conclusions	0.63 **	0.17 *	0.16 *	0.28 **	0.08	—			
7. Reading comprehension	0.39 **	0.15	0.29 *	0.18 *	0.31 **	0.20 *	—		
8. Numerical ability	0.18 *	0.13	0.04	0.09	0.18 *	0.07	0.27 **	—	
9. Problem-solving	0.17 *	0.07	0.08	0.09	0.13	0.11	0.14	0.15	—

$* p < 0.05$, $** p < 0.01$.

Table 4. Reading comprehension as explanatory factor of the scientific reasoning subskills.

Subskills	β	t	p	95% CI	Partial η^2
Hypothesizing	0.008	1.82	0.071	(0.00, 0.017)	0.021
Experimenting	0.017	4.09	<0.001	(0.009, 0.025)	0.098
Inferencing	0.007	2.10	0.037	(0.000, 0.013)	0.028
Evaluating data	0.011	3.23	0.001	(0.004, 0.018)	0.064
Drawing conclusions	0.011	2.43	0.016	(0.002, 0.020)	0.037

4. Discussion

This study aimed to identify and explain differences in children's ability to reason scientifically. To this end, a performance-based scientific reasoning task was administered, and measures of reading comprehension, numerical ability and problem-solving skills were collected in a sample of 160 upper-primary children. Their scientific reasoning scores varied considerably, which indicates that not all children are equally proficient in performing these skills. Observed differences within children further suggest that the five scientific reasoning skills are not equally difficult to perform. These intra-individual differences were partially explained by reading comprehension but not by numerical ability or problem-solving skills.

Results regarding the first research question confirm the existence of variation in children's scientific reasoning: the inter-individual spread in total scores was considerable, and marked intra-individual differences were found for some subskills. The hypothesized proficiency pattern was confirmed: children in our sample were most proficient in experimenting, less proficient in hypothesizing and least proficient in inferencing, evaluating data and drawing conclusions. This is particularly important because, as Koerber and Osterhaus [10] argued, previous research has studied these component skills separately, often through written tests [22,25]. The present study thus confirms the differences in subskill difficulty during a comprehensive performance-based scientific reasoning task and suggests that children's relative proficiency at the subskill level is stable across test modalities (cf. [6]).

Of particular interest is that the component scientific reasoning skills were consistently but moderately associated with total task scores. This result raises the question as to what accounts for the error variance in these correlations. Part of it could be due to the psychometric qualities of the scientific reasoning task. As mentioned in Section 2.2.1, each component skill was assessed by only three items, so a meaningful analysis of internal scale consistency was deemed impossible. In the absence of this information, the mag-

nitude of the correlations should be considered with some caution. A more substantive interpretation is that the proficiency pattern described above does not apply similarly to all children: some will develop the component skills in the indicated order, whereas others will show a deviating developmental trajectory. As a consequence, fine-grained measures of separate component skills, if reliably measured, give a more accurate impression of children's proficiency in scientific reasoning than global measures and should be the preferred approach when assessment serves diagnostic purposes, for instance, to inform the design of instruction.

The observed variation in scientific reasoning was independent of children's grade level. This equivalence of task performance might be due to the fact that our sample had few opportunities to practice their scientific reasoning skills—the school offered them only five inquiry projects per year, whereas the daily language and math classes lead to grade differences in reading comprehension and numerical ability. A related explanation is that scientific reasoning develops slowly in general and in the upper-primary grades in particular (e.g., [9]). Although most children at this age advance in scientific reasoning [19], the inter-individual variation is considerable and prevents the minor cross-grade growth differences from becoming statistically significant. Alternative research methods such as longitudinal designs and person-centered approaches to data analysis are more sensitive to capturing developmental growth and are increasingly being applied in scientific reasoning research [41].

Reading comprehension explained part of the variance in scientific reasoning. This result is consistent with hypotheses and complements previous research that administered written tests of scientific reasoning (e.g., [22,25,26]). Thus, why did reading comprehension predict scientific reasoning on a performance-based test that makes minimal demands on reading skills? One explanation is that scientific reasoning and reading comprehension both draw on general language comprehension processes, in particular when scientific reasoning is measured through interactive dialogue. Another interpretation could be that reading comprehension is a proxy of general intelligence or academic attainment, which, in turn, is associated with scientific reasoning (e.g., [42]). In addition, relations have been found between scientific reasoning and verbal reasoning [24], as well as nonverbal reasoning [25] and conditional sentence comprehension [43]. In line with these findings, language-centered scientific reasoning interventions have been proposed [25,43] and have been found to be effective [44].

Our results further show that reading comprehension does not explain all component scientific reasoning skills to the same extent, which underscores the importance of assessing the constituent skills separately rather than merging them in a single overarching construct. The most striking finding in this regard is that hypothesizing was not related to reading comprehension, even though one would intuitively expect verbal reasoning to be associated with this skill. Although it is not entirely clear why hypothesizing and reading comprehension were not related, a possible explanation may lie in what children need to reason about: their own ideas about the world (as in hypothesizing) as opposed to building a situation model from given information (as in reading [45] as well as in interpreting outcomes). In hypothesizing, misconceptions and naive beliefs may interfere with the reasoning process, whereas the chance of such 'illogical' thoughts could be less pronounced when reasoning with given information.

Numerical ability did not predict children's scientific reasoning. Although there were sound theoretical reasons to assume that numerical ability would predict scientific reasoning, empirical evidence on this relation is either scarce and relatively recent [10] or involved a different math strand [32]. Thus, while numerical ability as operationalized in this study does not explain individual differences in scientific reasoning, future research might examine whether this independence generalizes across tasks and settings. Future research could also investigate whether different math skills (e.g., number sense, measurement) contribute to performance on a scientific reasoning task.

Children's problem-solving skills did not predict scientific reasoning either, possibly because of task incongruence. Jonassen [46] argued that the ease with which a problem is solved relies on individual differences between problem solvers and problem characteristics. A scientific inquiry is an ill-defined problem that requires a problem solver to combine strategies and rules to come to an unknown solution, whereas the Tower of Hanoi is a well-defined problem with a constrained set of rules and a known solution. Thus, although the Tower of Hanoi does involve problem-solving, it may be insufficiently sensitive to distinguish weak from strong problem solvers. Beyond problem characteristics, the problem representation [46] might explain why Mayer et al. [22] found that the very similar Tower of London problem explained scientific reasoning. Mayer et al. [22] used a multiple-choice paper-and-pencil version of this problem in which all manipulations had to be completed mentally, thus making a relatively straightforward problem rather difficult to solve. As such, this test may not have identified all children who could solve a Tower of London problem, but only those who were sufficiently good at reasoning to complete the problem mentally. The current study, by contrast, used a less demanding task that allowed for real-time manipulation and was programmed to make invalid moves impossible. This difference in task demands might explain why the current study did not show a relation between problem-solving and scientific reasoning while previous research showed such a relation. As understanding what explains specific subskills is only a recent endeavor [10,25], more research is needed to understand which component skills can be explained as well as why differential effects are found.

4.1. Limitations

This study has some limitations, which include the homogenous sample in terms of parental background and education, with highly educated parents being overrepresented. As these parents are more likely to intellectually stimulate their children, for example, by taking them to science museums [47], this might have given the participants in the current study a certain advantage compared to children whose parents are less educated. The observed variation in scientific reasoning was nevertheless considerable and would probably have been even more diverse if a more heterogeneous sample had been used. Future research should therefore incorporate more diverse samples to find out whether the present conclusions generalize to more typical groups of upper-primary schoolchildren.

Another limitation lies in the task used to assess numerical ability. Because there was no precedent as to what type of math skills would predict scientific reasoning, a lean task that assessed basic numerical operations was chosen because it seemingly matched the type of operations children had to carry out during the scientific reasoning task (e.g., counting, direct comparisons). A further advantage of this task was that it did not make demands on reading skills, which is particularly important because previous studies did not allow for untangling of scientific reasoning and reading comprehension. However, although the current task resembled the types of *operations* children had to carry out during the scientific reasoning task, no *reasoning* was required. The absence of any significant results suggests that numerical ability may not be the most relevant math skill to predict scientific reasoning, and further research is needed to identify if and what math skills relate to scientific reasoning.

4.2. Implications

The current study confirms that scientific reasoning is a multifaceted construct. This is not only evident from differences in children's proficiency in the component skills but also from the asymmetry in the extent to which reading comprehension predicts these skills. How children of different proficiency levels learn scientific reasoning in a classroom setting and can be taught to reach their best potential is something that needs to be attended to in future research. Studying all scientific reasoning skills together is particularly important. Previous research has predominantly focused on a single skill, most often experimenting [48], which stands to reason because experimenting is such

a fundamental skill. At the same time, these focused investigations do not capture the complexity of scientific inquiry, the relative proficiency of children in the different subskills and the relations between these skills. Therefore, future research should focus more on scientific reasoning in authentic inquiry settings while still distinguishing subskills.

The absence of grade-level differences suggests that scientific reasoning develops slowly in the upper-primary years and implies that sustained practice is needed to boost this development. In preparing weekly or bi-weekly inquiry-based science lessons, teachers should attend to differences between children and among subskills. Most children will be able to perform the relatively easy skill of experimenting themselves with minimal guidance, whereas more teacher guidance is needed in generating hypotheses. Inferencing, evaluating data and drawing conclusions, which are the most difficult subskills, should initially be taken over by the teacher, who can demonstrate the skills to the class and gradually decrease their involvement as the lesson series progresses.

Results of the multiple regression analysis imply that teachers who start an inquiry-based curriculum can infer children's entry levels from their reading comprehension scores—children's basic numerical skills and ability to solve mind puzzles that resemble the Tower of Hanoi (e.g., tangrams, sudokus) should not be used for this purpose because both are poor predictors of scientific reasoning. The regression data also suggest that proficient readers need less guidance in scientific reasoning, so teachers can devote more attention to the average and poor readers in the class. Teachers should, of course, monitor the progress of all children and adjust the level of guidance just-in-time on an as-needed basis. A final practical suggestion concerns the scheduling of inquiry-based science classes. As these lessons are often taught by specialist teachers with part-time contracts, schools can opt for flexible scheduling and combine the fifth- and sixth-grade lessons because the proficiency levels in these classes are comparable. Alternatively, the same lessons can be delivered in both grades, perhaps with some minor adjustments in the amount of guidance, which will ease the teachers' burden in lesson preparation.

5. Conclusions

This study found substantial overall differences in children's scientific reasoning as well as marked differences at the subskill level. This variation was in part explained by children's reading comprehension but not their numerical ability and problem-solving skills. These results confirm the importance of treating scientific reasoning as a multifaceted skill. Both teachers and researchers should address scientific reasoning in an integrated setting where its component skills are distinguished but not studied or taught in isolation. As reading comprehension explains scientific reasoning in general and most of its constituent skills, science teachers should give more guidance to the poor readers in their classes, and researchers should administer performance-based assessments of scientific reasoning that make minimal demands on reading skills.

References

1. Abd-El-Khalick, F.; BouJaoude, S.; Duschl, R.; Lederman, N.G.; Mamlok-Naaman, R.; Hofstein, A.; Niaz, M.; Treagust, D.; Tuan, H.-L. Inquiry in science education: International perspectives. *Sci. Educ.* **2004**, *88*, 397–419. [CrossRef]
2. Achieve. *International Science Benchmarking Report Taking the Lead in Science Education: Forging Next-Generation Science Standards*; Achieve: Washington, DC, USA, 2010.
3. Kuhn, D. What is scientific thinking and how does it develop? In *Blackwell Handbook of Childhood Cognitive Development*; Goswami, U., Ed.; Blackwell Publishers Ltd: Malden, USA, 2002; pp. 371–393.
4. Zimmerman, C. The development of scientific thinking skills in elementary and middle school. *Dev. Rev.* **2007**, *27*, 172–223. [CrossRef]
5. Kind, P.M.; Osborne, J. Styles of Scientific Reasoning: A Cultural Rationale for Science Education? *Sci. Educ.* **2017**, *101*, 8–31. [CrossRef]
6. Kruit, P.M.; Oostdam, R.J.; van den Berg, E.; Schuitema, J.A. Assessing students' ability in performing scientific inquiry: Instruments for measuring science skills in primary education. *Res. Sci. Technol. Educ.* **2018**, *36*, 413–439. [CrossRef]
7. Schiefer, J.; Golle, J.; Tibus, M.; Oschatz, K. Scientific reasoning in elementary school children: Assessment of the inquiry cycle. *J. Adv. Acad.* **2019**, *30*, 144–177. [CrossRef]
8. Pedaste, M.; Mäeots, M.; Siiman, L.A.; de Jong, T.; van Riesen, S.A.N.; Kamp, E.T.; Manoli, C.C.; Zacharia, Z.C.; Tsourlidaki, E. Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educ. Res. Rev.* **2015**, *14*, 47–61. [CrossRef]
9. Piekny, J.; Maehler, C. Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *Br. J. Dev. Psychol.* **2013**, *31*, 153–179. [CrossRef]
10. Koerber, S.; Osterhaus, C. Individual differences in early scientific thinking: Assessment, cognitive Influences, and their relevance for science learning. *J. Cogn. Dev.* **2019**, *20*, 510–533. [CrossRef]
11. Harlen, W. Inquiry-based learning in science and mathematics. *Rev. Sci. Math. ICT Educ.* **2013**, *7*, 9–33. [CrossRef]
12. Edelsbrunner, P.A.; Dablander, F. The Psychometric Modeling of Scientific Reasoning: A Review and Recommendations for Future Avenues. *Educ. Psychol. Rev.* **2018**, *31*, 1–34. [CrossRef]
13. Köksal-Tuncer, Ö.; Sodian, B. The development of scientific reasoning: Hypothesis testing and argumentation from evidence in young children. *Cogn. Dev.* **2018**, *48*, 135–145. [CrossRef]
14. Van der Graaf, J.; Segers, E.; Verhoeven, L. Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instr. Sci.* **2015**, *43*, 381–400. [CrossRef]
15. Lorch, R.F.; Lorch, E.P.; Freer, B.; Calderhead, W.J.; Dunlap, E.; Reeder, E.C.; Van Neste, J.; Chen, H.T. Very long-term retention of the control of variables strategy following a brief intervention. *Contemp. Educ. Psychol.* **2017**, *51*, 391–403. [CrossRef]
16. Chen, Z.; Klahr, D. All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Dev.* **1999**, *70*, 1098–1120. [CrossRef]
17. Kuhn, D.; Dean, D. Is developing scientific thinking all about learning to control variables? *Psychol. Sci.* **2005**, *16*, 866–870. [CrossRef]
18. Lazonder, A.W.; Wiskerke-Drost, S. Advancing scientific reasoning in upper elementary classrooms: Direct instruction versus task structuring. *J. Sci. Educ. Technol.* **2015**, *24*, 69–77. [CrossRef]
19. Lazonder, A.W.; Janssen, N. Development and initial validation of a performance-based scientific reasoning test for children. *Stud. Educ. Eval.* **2021**, *68*, 100951. [CrossRef]
20. van Uum, M.S.J.; Verhoeff, R.P.; Peeters, M. Inquiry-based science education: Scaffolding pupils' self-directed learning in open inquiry. *Int. J. Sci. Educ.* **2017**, *39*, 2461–2481. [CrossRef]
21. Koerber, S.; Mayer, D.; Osterhaus, C.; Schwippert, K.; Sodian, B. The development of scientific thinking in elementary school: A comprehensive inventory. *Child Dev.* **2015**, *86*, 327–336. [CrossRef]
22. Mayer, D.; Sodian, B.; Koerber, S.; Schwippert, K. Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learn. Instr.* **2014**, *29*, 43–55. [CrossRef]
23. Wagensveld, B.; Segers, E.; Kleemans, T.; Verhoeven, L. Child predictors of learning to control variables via instruction or self-discovery. *Instr. Sci.* **2014**, *43*, 365–379. [CrossRef]
24. Siler, S.A.; Klahr, D.; Magaro, C.; Willows, K.; Mowery, D. Predictors of transfer of experimental design skills in elementary and middle school children. In Proceedings of the 10th International Conference on Intelligent Tutoring Systems, Pittsburgh, PA, USA, 1 June 2010; pp. 198–208.

25. Van de Sande, E.; Kleemans, M.; Verhoeven, L.; Segers, E. The linguistic nature of children's scientific reasoning. *Learn. Instr.* **2019**, *62*, 20–26. [CrossRef]
26. Snow, C.E. Academic language and the challenge of reading for learning about science. *Science* **2010**, *328*, 450–452. [CrossRef]
27. van Graft, M.; Tank, M.K.; Beker, T.; van der Laan, A. *Wetenschap en Technologie in Het Basis- en Speciaal Onderwijs: Richtinggevend Leerplankader Bij Het Leergebied Oriëntatie op Jezelf en de Wereld.*; SLO (nationaal expertisecentrum leerplanontwikkeling): Enschede, The Netherlands, 2018.
28. Wong, V. Authenticity, transition and mathematical competence: An exploration of the values and ideology underpinning an increase in the amount of mathematics in the science curriculum in England. *Int. J. Sci. Educ.* **2019**, *41*, 1805–1826. [CrossRef]
29. Schauble, L. In the eye of the beholder: Domain-general and domain-specific reasoning in science. In *Scientific Reasoning and Argumentation*; Fischer, F., Chinn, C.A., Engelmann, K., Osborne, J., Eds.; Routledge: New York, NY, USA, 2018; pp. 11–33.
30. Krummenauer, J.; Kuntze, S. Primary students' reasoning and argumentation based on statistical data. In Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education, Utrecht, The Netherlands, 6–10 February 2019.
31. Makar, K.; Bakker, A.; Ben-Zvi, D. The reasoning behind informal statistical inference. *Math. Think. Learn.* **2011**, *13*, 152–173. [CrossRef]
32. Bullock, M.; Ziegler, A. Scientific reasoning: Developmental and individual differences. In *Individual Development from 3 to 12: Findings from the Munich Longitudinal Study*; Cambridge University Press: Cambridge, UK, 1999; pp. 38–54.
33. Tajudin, N.M.; Chinnappan, M. Exploring relationship between scientific reasoning skills and mathematics problem solving. In Proceedings of the 38th Annual Conference of the Mathematics Education Research Group of Australasia: Mathematics Education in the Margins, Sunshine Coast, Australia, 28 June–2 July 2015; pp. 603–610.
34. Klahr, D.; Dunbar, K. Dual space search during scientific reasoning. *Cogn. Sci.* **1988**, *12*, 1–48. [CrossRef]
35. Piaget, J. Part I: Cognitive development in children: Piaget development and learning. *J. Res. Sci. Teach.* **2003**, *40*, 8–18. [CrossRef]
36. Shavelson, R.J.; Baxter, G.P.; Pine, J. Performance assessment in science. *Appl. Meas. Educ.* **1991**, *4*, 347–362. [CrossRef]
37. Weekers, A.; Groenen, I.; Kleintjes, F.; Feenstra, H. *Wetenschappelijke Verantwoording Papieren Toetsen Begrijpend Lezen Voor Groep 7 en 8 [Scientific Justification Pen-and-Paper Tests Reading Comprehension Grade 5 and 6]*; CITO: Arnhem, The Netherlands, 2011.
38. Verhelst, N.D.; Glas, C.A.W. The one parameter logistic model. In *Rasch Models: Foundations, Recent Developments and Applications*; Fischer, G.H., Molenaar, I.W., Eds.; Springer: New York, NY, USA, 1995; pp. 215–239. [CrossRef]
39. De Vos, T. *Schoolvaardigheidstoets Hoofdrekenen [Arithmetic Proficiency Test for Primary School]*; Boom Test Uitgevers: Amsterdam, The Netherlands, 2006.
40. Welsh, M.C. Rule-guided behavior and self-monitoring on the tower of hanoi disk-transfer task. *Cogn. Dev.* **1991**, *6*, 59–76. [CrossRef]
41. Hickendorff, M.; Edelsbrunner, P.A.; McMullen, J.; Schneider, M.; Trezise, K. Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learn. Individ. Differ.* **2018**, *66*, 4–15. [CrossRef]
42. Veenman, M.V.J.; Wilhelm, P.; Beishuizen, J.J. The relation between intellectual and metacognitive skills from a developmental perspective. *Learn. Instr.* **2004**, *14*, 89–109. [CrossRef]
43. Svirko, E.; Gabbott, E.; Badger, J.; Mellanby, J. Does acquisition of hypothetical conditional sentences contribute to understanding the principles of scientific enquiry? *Cogn. Dev.* **2019**, *51*, 46–57. [CrossRef]
44. Van der Graaf, J.; van de Sande, E.; Gijsel, M.; Segers, E. A combined approach to strengthen children's scientific thinking: Direct instruction on scientific reasoning and training of teacher's verbal support. *Int. J. Sci. Educ.* **2019**, *41*, 1119–1138. [CrossRef]
45. Swart, N.M.; Muijselaar, M.M.L.; Steenbeek-Planting, E.G.; Droop, M.; de Jong, P.F.; Verhoeven, L. Cognitive precursors of the developmental relation between lexical quality and reading comprehension in the intermediate elementary grades. *Learn. Individ. Differ.* **2017**, *59*, 43–54. [CrossRef]
46. Jonassen, D.H. Toward a design theory of problem solving. *Educ. Technol. Res. Dev.* **2000**, *48*, 63–85. [CrossRef]
47. Archer, L.; Dawson, E.; Seakins, A.; Wong, B. Disorientating, fun or meaningful? Disadvantaged families' experiences of a science museum visit. *Cult. Stud. Sci. Educ.* **2016**, *11*, 917–939. [CrossRef]
48. Rönnebeck, S.; Bernholt, S.; Ropohl, M. Searching for a common ground—A literature review of empirical research on scientific inquiry activities. *Stud. Sci. Educ.* **2016**, *52*, 161–197. [CrossRef]