

# **Individual differences in children's scientific reasoning**

Erika Schlatter

The research presented in this dissertation was carried out at the Behavioral Science Institute of Radboud University, the Netherlands.

*Coverdesign:* Maartje de Goede | [www.maartjedegoede.nl](http://www.maartjedegoede.nl)

*Printing:* Ridderprint | [www.ridderprint.nl](http://www.ridderprint.nl)

© Erika Schlatter, 2022

# Individual differences in children's scientific reasoning

Proefschrift ter verkrijging van de graad van doctor  
aan de Radboud Universiteit Nijmegen  
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,  
volgens besluit van het college voor promoties  
in het openbaar te verdedigen op

vrijdag 17 juni 2022  
om 12.30 uur precies

door  
Erika Schlatter  
geboren 25 september 1988  
te Amsterdam

**Promotor**

Prof. dr. A. W. Lazonder

**Copromotor**

Dr. I. Molenaar

**Manuscriptcommissie**

Prof. dr. P. C. J. Segers

Prof. dr. M. Raijmakers (*Vrije Universiteit Amsterdam*)

dr. A. Raes (*KU Leuven, België*)

# Contents

Chapter 1	General introduction	7
Chapter 2	Individual differences in children's scientific reasoning	21
Chapter 3	Individual differences in children's development of scientific reasoning through inquiry-based instruction: Who needs additional guidance?	47
Chapter 4	Learning scientific reasoning: A latent transition analysis	81
Chapter 5	Adapting scientific reasoning instruction to children's needs: effects on learning processes and learning outcomes	109
Chapter 6	General discussion	137
Appendices	A: Worksheet example for Chapters 3 and 4	154
	B: Support levels during inquiry sessions	160
	C: Worksheet example for Chapters 3 and 4	164
	Nederlandstalige samenvatting	174
	Dankwoord	186
	Author biography	190
	Publication list	191



# Chapter 1

## **General introduction**

## Introduction

Asking questions and figuring things out oftentimes seems to be the natural state of children. Their curiosity is wholeheartedly embraced in primary school science education today: walk into any science classroom and you will likely encounter children involved in hands-on science projects. Through such projects, children are expected to gain both science content knowledge and research skills. These research skills encompass practical skills, such as handling specialized equipment like a timer or a scale, as well as cognitive skills, such as hypothesizing, experimenting and evaluating outcomes. These cognitive skills are also known as scientific reasoning.

Scientific reasoning is often conceived of as a process of intentional knowledge seeking (Kuhn, 2002; Zimmerman, 2007). This process comprises multiple phases that call upon different yet interrelated skills. For example, research questions and hypotheses should be thoughtfully formulated in order to design an informative experiment and collect meaningful data, which then needs accurate interpretation in order to draw valid conclusions and, ultimately, find new knowledge. As such, scientific reasoning is an inherently whole-task activity.

Scientific reasoning is not just a skill for scientists. In general, it helps people understand the world around them (Chinn & Golan Duncan, 2018; Kind & Osborne, 2017) and make informed decisions about socio-scientific issues relevant to daily life (Sadler, 2004), thus enabling them to participate in modern society (Trilling & Fadel, 2009). For children in particular, scientific reasoning is important because it helps them acquire science content knowledge in school (Edelsbrunner et al., 2018; Zimmerman, 2007). Good scientific reasoning skills are also crucial to meaningful inquiry-based learning, a pedagogical approach in which content knowledge is acquired through scientific investigation (Lazonder & Harmsen, 2016).

However, teaching scientific reasoning in the primary classroom can be challenging. One of the main challenges is a multifaceted diversity: differences exist in scientific reasoning proficiency of same-age children, as well as in the difficulty of component scientific reasoning skills (Lazonder et al., 2021). The studies in this dissertation therefore sought to answer two main questions:

- (1) How can differences in upper primary school children's scientific reasoning be characterized and predicted?
- (2) How can these differences be addressed in upper primary classrooms?



## Research context

The research presented in this thesis was carried out in the Netherlands between 2017 and 2021. In this period, two developments regarding science and technology in primary education have been playing out. First, 2020 was the target year for a technology pact ('Techniekpact 2020'), which among other things included the ambition to teach science and technology in all Dutch primary schools by 2020. Second, in 2018 the Netherlands started a large and still ongoing curriculum renewal under the name Curriculum.nu. As this new curriculum is slowly taking shape, it is becoming clear that both inquiry and scientific reasoning will be important aspects in several domains, such as People and Nature ('Mens & Natuur') and People and Society ('Mens & Maatschappij').

Even though these developments showcase clear ambitions regarding science education, these aspirations do not automatically translate into curricular activities at the classroom level – primarily because Dutch schools traditionally have a large autonomy in the interpretation of curricular requirements. The curricular goals (Dutch: kerndoelen; Greven & Letschert, 2006) for primary education are very concise, and currently the only goal with regard to scientific reasoning (goal 42) reads: 'The pupils learn to investigate materials and natural phenomena, such as light, sound, electricity, power, magnetism and temperature'.

SLO, the national centre of expertise for curricular development, offers more specific subgoals and learning pathways (TULE: tussendoelen en leerlijnen) with regard to the science content of this curricular goal, as well as more general guidelines for science, technology and inquiry-based learning (Van Graft et al., 2018). Nonetheless, by 2018 only a small percentage of schools had implemented inquiry-based learning in their curriculum (Van Graft et al., 2018), often in the form of long-term projects that revolve around subject-specific content. In such projects, children are expected to develop research skills while at the same time employing these skills to acquire content knowledge.

Although content knowledge and research skills can strengthen each other (Schwichow et al., 2020; Zimmerman, 2007), a good basis in scientific reasoning is needed in order for children to learn from inquiry: poor research skills can result in poor research outcomes (Lazonder & Wiskerke-Drost, 2015) and, hence, poor learning of science content (Edelsbrunner et al., 2018). Because these skills do not develop automatically, it remains important to devote explicit attention to both practical research skills and scientific reasoning skills. It is for this reason that the current dissertation homes in on learning scientific reasoning in Dutch upper primary school grades.

## Components of scientific reasoning

In order to bring scientific reasoning to the Dutch primary classroom, a thorough understanding of scientific reasoning is needed, and a description of the skills that comprise scientific reasoning and their interrelatedness is an important first step in this effort. Klahr and Dunbar (1988) were probably the first to acknowledge the mutual dependency of distinct scientific reasoning skills. In their observations of undergraduate students working to figure out an unknown function of a robot tank, they found that hypotheses and experiments inform each other. In this process, which was named Scientific Discovery as Dual Search (SDDS), some students used experiments to inform their hypotheses, whereas others formulated hypotheses based on their own beliefs or prior knowledge in order to inform their experiments. Both strategies eventually led to the discovery of the unknown function, albeit in fundamentally different ways.

Although these differences in inquiry strategy could not have been uncovered by looking at either hypothesizing or experimenting in isolation, most contemporary research focuses on a single component skill – often experimenting (Koerber & Osterhaus, 2019). A few studies have examined multiple component skills, but used separate tasks for each skill rather than a single comprehensive task that involved all component skills (e.g., Koerber et al., 2015; Piekny & Maehler, 2013; Van de Sande et al., 2019). This is problematic from the conceptual viewpoint outlined above, as well as from a practical perspective: both in authentic scientific research and in inquiry-based learning, component skills of scientific reasoning are seldomly applied in isolation.

Considering the importance of studying scientific reasoning's constituent skills in a whole-task setting, the question rises as to what these skills are. Generally, three main skills are distinguished: hypothesizing, experimenting and evaluating outcomes (Kind, 2013; Klahr & Dunbar, 1988; Zimmerman, 2007). However, researchers often give a personal twist to these catch-all terms, resulting in a large set of different terms for similar skills. In a review of 32 studies on inquiry-based learning, Pedaste et al. (2015) found 109 different terms for the various skills. Although minor terminological variations were inevitable even in this dissertation (see Table 1), the target skills can be ranged under the three main categories of hypothesizing, experimenting and evaluating outcomes.

Hypothesizing involves the articulation of ideas about possible outcomes of an investigation, which often occurs at the onset of the research process (Van Joolingen & De Jong, 1991). Experimenting concerns the ability to design and perform systematic comparisons in order to test or inform hypotheses (Chen & Klahr, 1999). To evaluate

outcomes, children should consider their data in light of their prior knowledge and hypotheses, thereby coordinating theory and evidence (Koslowski et al., 2008). For the purpose of this dissertation, this broad subskill (Zimmerman, 2007) was decomposed into three constituent parts: interpreting data, evaluating data characteristics and drawing conclusions. Interpreting data was defined as the ability to identify patterns or regularities in one or more sets of numerical or visual representations of the outcomes of an experiment (Kanari & Millar, 2004). Evaluating data characteristics was defined as the assessment of the reliability and trustworthiness of a dataset on the basis of properties such as sample size and spread (Lubben & Millar, 2007; Masnick & Morris, 2008). Drawing conclusions was defined as the use of the previously gathered and interpreted data to make causal statements and evaluate previously formulated hypotheses (Moritz, 2003).

Throughout this thesis, different instruments were used to assess scientific reasoning. These instruments all measured various component skills, albeit not always the same set of skills (see Table 1). In Chapter 2, all five component skills mentioned above were assessed by means of a performance-based test (Lazonder & Janssen, 2021). In Chapters 3 through 5, where the learning of scientific reasoning was studied in a classroom setting, the worksheets children filled out were used as an instrument to gauge children's learning throughout the lesson series. The component skill 'evaluating data characteristics' was omitted from these worksheets because it was found to be rather difficult for children in the upper primary grades, and virtually impossible to incorporate in an inquiry where children collect their own data. In addition to the worksheets, a validated paper-and-pencil test (Van de Sande et al., 2019) was used as a pre- and post-test in Chapters 3 and 5. This test comprised three scales: experimenting and drawing conclusions, as well as a combined scale named hypothesis-evidence coordination, which comprised items on data interpretation and hypothesizing.

**Table 1***Instruments used to assess scientific reasoning*

Scientific reasoning skill	Performance-based			
	test <sup>1</sup>	Worksheets	Paper-and-pencil test <sup>2</sup>	
Hypothesizing	•	•	•	} Combined scale: hypothesis-evidence coordination
Experimenting	•	•	•	
Interpreting data	•	•	•	
Evaluating data	•			
Drawing conclusions	•	•	•	

<sup>1</sup>Lazonder & Janssen (2021) <sup>2</sup>Scientific Reasoning Inventory; Van de Sande et al. (2019)

## Diversity in scientific reasoning

### *Differences across skills*

Previous research points to substantial variation in the difficulty of the component scientific reasoning skills introduced above. Experimenting, for example, is relatively easy for children to learn: most pre-schoolers are capable of differentiating between confounded and unconfounded experimental comparisons and some are even able to perform systematic testing (Köksal-Tuncer & Sodian, 2018; Van der Graaf et al., 2015). Older children can be successfully taught to set up controlled experiments, both through direct instruction (Chen & Klahr, 1999; Lorch et al., 2017) and guided inquiry (Kuhn & Dean, 2005; Lazonder & Wiskerke-Drost, 2015). Hypothesizing is more difficult for young children to learn. Pre-schoolers have particularly poor hypothesizing skills, and although lower-primary school pupils do slightly better it is only by the end of primary education that children are actually starting to master this skill (Piekny & Maehler, 2013). Children's ability to evaluate the outcomes of an investigation is highly dependent on the type of data: although pre-schoolers are able to identify perfect covariation, it is only in the highest grades of primary education that most children can interpret imperfect covariation and non-covariation (Lazonder et al., 2021; Piekny & Maehler, 2013). Characteristics like sample size and spread of data are also difficult for children to attend to (Masnick & Morris, 2008), making the evaluation of outcomes the hardest skill to acquire.

Although this tentative order has its foundations in a solid basis of empirical evidence, most of these studies either examined a single skill at a single time point or used

separate tasks for each component skill – instead of one overarching task in which all component skills had to be applied. Chapter 2 took this possible methodological limitation into account by administering a performance-based test that engaged children in four inquiry cycles, thus enabling them to repeatedly demonstrate their ability to perform all component scientific reasoning skills in the context of a single investigation.

### *Differences between children*

As with many school subjects, there are also differences between same-aged children: some are better at scientific reasoning than others. Studies suggest that these differences grow over the course of primary education (Piekný & Maehler, 2013), making it particularly important for upper primary school teachers to address these differences in the classroom. To enable teachers in doing so, it is important to further characterise and explain the differences between children at the level of the component skills.

Previous studies point towards reading comprehension as an important predictor of scientific reasoning (Mayer et al., 2014; Van de Sande et al., 2019). Although reading comprehension seems a robust predictor, this relationship has been established using paper-and-pencil tests that require children to read questions and answers – thereby introducing a possible confound. Problem solving could be another predictor of children’s scientific reasoning: scientific reasoning can be characterized as a process of rule induction (Klahr & Dunbar, 1988), which inherently involves problem solving. Although some evidence points towards problem solving as a predictor of at least some component skills (Mayer et al., 2014; Van de Sande et al., 2019), this predictor is less established than reading comprehension. Lastly, although scientific reasoning involves reasoning about numerical data (Krummenauer & Kuntze, 2019; Makar et al., 2011), numerical ability had not been studied as a predictor of scientific reasoning at the start of this thesis research. Because the cognitive characteristics described above are either already known to predict scientific reasoning in particular circumstances, or show potential to do so, reading comprehension, problem solving and numerical ability were used to further explain differences in scientific reasoning.

In order to help teachers support children in a classroom environment, it is important to know which of these predictors explains specific component skills, and whether they do so in a classroom setting. Chapter 2 therefore examined differences between children using a performance-based test at a single point in time, and assessed the value of reading comprehension, numerical ability and problem solving skill as predictors of the component scientific reasoning skills. Chapter 3 elaborates on this by examining how

scientific reasoning skills *develop* in the short term in response to instruction and guided practice. To this end, both pre-and post-test scores and process data from children's worksheets were analysed, as well as two predictors on which information is readily available in most schools: reading comprehension and, as a broad measure of numerical ability, mathematical skilfulness.

### *Innovative approaches for understanding individual differences*

Chapters 2 and 3 used traditional, variable centred analyses to predict scientific reasoning from stable cognitive indicators. Although this deterministic approach is appropriate for analysing individual differences, it unlikely captures the full spectrum of variation in scientific reasoning. Useful complementary evidence can be provided by probabilistic, person-centred analysis techniques capable of uncovering latent proficiency profiles (Hickendorff et al., 2018). Rather than relying on stable traits, person-centred analyses allow for the identification of subgroups based solely on the variables of interest. As such, subgroups of children can be distinguished based on their similarity to one another as well as their difference from children in other subgroups (Hickendorff et al., 2018). In the domain of scientific reasoning, person-centred analyses have been used to establish patterns of conceptual change (Schneider & Hardy, 2013; Van der Graaf, 2020) and proficiency profiles in secondary students' understanding of variable control (Schwichow et al., 2020).

Latent Transition Analysis (LTA) is a person-centred analysis technique particularly suited for analysing learning (Reimann, 2009), as it uncovers both proficiency profiles and learning pathways. With LTA, it can be determined whether individuals remain in the same profile through time or transition from one profile to another. Such analyses are particularly promising in multifaceted domains like scientific reasoning: at different points in time, the identified subgroups might need different amounts of support for different component skills. Therefore, the worksheets used in the lesson series described in Chapter 3 were further scrutinized in Chapter 4, where LTA was used to establish proficiency profiles and progression through these profiles throughout the lesson series.

### **Adaptivity as a means to address individual differences**

The multifaceted diversity addressed in the first three studies of this thesis makes scientific reasoning a challenging subject to teach. Yet, how these observed differences between skills and across individuals are best addressed in primary science classrooms has hardly been studied. Adaptive instruction, a teaching strategy in which learning materials and teaching

strategies are adjusted based on information about students (Aleven et al., 2016) could be a fruitful means to mitigate this challenge. Previous research has shown that adaptive teaching can promote science content learning (e.g., McCrea Simpkins et al., 2009). However, its effect on the development of scientific reasoning *skills* has to our knowledge not been studied before. Furthermore, adaptivity is often software-based or software-supported (Deunk et al., 2018) whereas scientific reasoning is often taught using physical materials (Evangelou & Kotsis, 2018). Thus, although adaptive teaching materials are promising for science education, their application and effectiveness in unplugged (non-digital) settings has yet to be shown.

Therefore, two adaptive variants of the lesson series used in Chapters 3 and 4 were developed, for which children did not have to use a computer. These adaptive lessons either used information on children's ongoing task performance (the micro-adaptive condition) or on their reading comprehension and mathematical skilfulness (the macro-adaptive condition) to adjust scientific reasoning support in an unplugged setting. In Chapter 5, the learning processes and learning outcomes of children participating in these adaptive lesson series were compared to a non-adaptive control condition.

## Thesis overview

The studies presented in this dissertation aimed to unravel differences in scientific reasoning – both between children and across skills – in order to ultimately address these differences in the classroom. As can be seen in Table 2, each chapter focuses on one or two of the main research questions, and research methods were chosen carefully in order to paint a complete picture of children's scientific reasoning proficiency and its development. Throughout the chapters, insights are shared from individually administered one-time assessments (Chapter 2) as well as longitudinal data from classroom studies (Chapters 3 through 5). A range of instruments and analysis techniques were used to carefully build towards the adaptive intervention presented in Chapter 5.

The ultimate goal of these studies was to aid the improvement of scientific reasoning instruction in upper primary classrooms, so that the lively science classroom situations described at the beginning of this chapter can be even more instructive in the future – both in terms of science content learning and the acquisition of scientific reasoning skills.

**Table 2***Characteristics of the studies presented in this thesis*

	Chapter			
	2	3	4	5
Research questions				
Characterizing and predicting	•	•	•	
Addressing differences		•	•	•
Participants				
N	160	154	166	153
Age range	8-12	8-12	8-12	9-12
% boys	54	55	55	54
Scientific reasoning instruments				
Performance-based test <sup>1</sup>	•			
Paper-and-pencil test <sup>2</sup>		•		•
Worksheets		•	•	•
Additional instruments				
Tower of Hanoi	•			
Basic numeracy test (SVT-HR)	•			
Reading comprehension (Cito)	•	•		•
Mathematical skilfulness (Cito)		•		•
Analyses				
Correlations	•			
Analysis of Variance	•	•		•
Regression	•			
Person-centred analysis			•	
Non-parametric tests				•

*Note.* In Chapter 4, Cito test scores were used in the adaptivity mechanism, but not in the analyses.

<sup>1</sup>Lazonder & Janssen (2021) <sup>2</sup>Scientific Reasoning Inventory; Van de Sande et al. (2019)



## References

- Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016). Instruction based on adaptive learning technologies. In R. E. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 522-560). Routledge.
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098-1120. <https://doi.org/10.1111/1467-8624.00081>
- Chinn, C. A., & Golan Duncan, R. (2018). What is the value of general knowledge of scientific reasoning? In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation* (pp. 77-101). Routledge.
- Deunk, M. I., Smale-Jacobse, A. E., De Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education. *Educational Research Review*, *24*, 31-54. <https://doi.org/10.1016/j.edurev.2018.02.002>
- Edelsbrunner, P. A., Schalk, L., Schumacher, R., & Stern, E. (2018). Variable control and conceptual change: A large-scale quantitative study in elementary school. *Learning and Individual Differences*, *66*, 38-53. <https://doi.org/10.1016/j.lindif.2018.02.003>
- Evangelou, F., & Kotsis, K. (2018). Real vs virtual physics experiments: Comparison of learning outcomes among fifth grade primary school students. A case on the concept of frictional force. *International Journal of Science Education*, *41*(3), 330-348. <https://doi.org/10.1080/09500693.2018.1549760>
- Greven, J., & Letschert, J. (2006). *Kerndoelen primair onderwijs* [Curricular goals for primary education]. Ministerie van Onderwijs, Cultuur en Wetenschap.
- Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2018). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning and Individual Differences*, *66*, 4-15. <https://doi.org/10.1016/j.lindif.2017.11.001>
- Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*(7), 748-769. <https://doi.org/10.1002/tea.20020>
- Kind, P. M., & Osborne, J. (2017). Styles of scientific reasoning: A cultural rationale for science education? *Science Education*, *101*(1), 8-31. <https://doi.org/10.1002/sce.21251>
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*(1), 1-48. [https://doi.org/10.1207/s15516709cog1201\\_1](https://doi.org/10.1207/s15516709cog1201_1)
- Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development*, *86*(1), 327-336. <https://doi.org/10.1111/cdev.12298>
- 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.org/10.1080/15248372.2019.1620232>
- Köksal-Tuncer, Ö., & Sodian, B. (2018). The development of scientific reasoning: Hypothesis testing and argumentation from evidence in young children. *Cognitive Development*, *48*, 135-145. <https://doi.org/10.1016/j.cogdev.2018.06.011>
- Koslowski, B., Marasia, J., Chelenza, M., & Dublin, R. (2008). Information becomes evidence when an explanation can incorporate it into a causal framework. *Cognitive Development*, *23*(4), 472-487. <https://doi.org/10.1016/j.cogdev.2008.09.007>

- Krummenauer, J., & Kuntze, S. (2019, February). *Primary students' reasoning and argumentation based on statistical data*. Eleventh Congress of the European Society for Research in Mathematics Education, Utrecht, the Netherlands. <https://hal.archives-ouvertes.fr/hal-02398118/>
- Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 371-393). Blackwell Publishers Ltd.
- Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, 16(11), 866-870. <https://doi.org/10.1111/j.1467-9280.2005.01628.x>
- Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning. *Review of Educational Research*, 86(3), 681-718. <https://doi.org/10.3102/0034654315627366>
- Lazonder, A. W., & Janssen, N. (2021). Development and initial validation of a performance-based scientific reasoning test for children. *Studies in Educational Evaluation*, 68, Article 100951. <https://doi.org/10.1016/j.stueduc.2020.100951>
- Lazonder, A. W., Janssen, N., Gijlers, H., & Walraven, A. (2021). Patterns of development in children's scientific reasoning: results from a three-year longitudinal study. *Journal of Cognition and Development*, 22(1), 108-124. <https://doi.org/10.1080/15248372.2020.1814293>
- Lazonder, A. W., & Wiskerke-Drost, S. (2015). Advancing scientific reasoning in upper elementary classrooms: Direct instruction versus task structuring. *Journal of Science Education and Technology*, 24(1), 69-77. <https://doi.org/10.1007/s10956-014-9522-8>
- Lorch, R. F., Lorch, E. P., Freer, B., Calderhead, W. J., Dunlap, E., Reeder, E. C., Van Neste, J., & Chen, H. T. (2017). Very long-term retention of the control of variables strategy following a brief intervention. *Contemporary Educational Psychology*, 51, 391-403. <https://doi.org/10.1016/j.cedpsych.2017.09.005>
- Lubben, F., & Millar, R. (2007). Children's ideas about the reliability of experimental data. *International Journal of Science Education*, 18(8), 955-968. <https://doi.org/10.1080/0950069960180807>
- Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, 13(1-2), 152-173. <https://doi.org/10.1080/10986065.2011.538301>
- Masnack, A., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development*, 79(4), 1032-1048. <https://doi.org/10.1111/j.1467-8624.2008.01174.x>
- 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>
- McCrea Simpkins, P., Mastropieri, M. A., & Scruggs, T. E. (2009). Differentiated curriculum enhancements in inclusive fifth-grade science classes. *Remedial and Special Education*, 30(5), 300-308. <https://doi.org/10.1177/0741932508321011>
- Moritz, J. (2003). Reasoning about covariation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 227-257). Kluwer Academic Publishers.
- Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology*, 31(2), 153-179. <https://doi.org/10.1111/j.2044-835X.2012.02082.x>
- Reimann, P. (2009). Time is precious: Variable- and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning*, 4(3), 239-257. <https://doi.org/10.1007/s11412-009-9070-z>

- Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching*, 41(5), 513-536. <https://doi.org/10.1002/tea.20009>
- Schneider, M., & Hardy, I. (2013). Profiles of inconsistent knowledge in children's pathways of conceptual change. *Developmental Psychology*, 49(9), 1639-1649. <https://doi.org/10.1037/a0030976>
- 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, Article 101923. <https://doi.org/10.1016/j.cedpsych.2020.101923>
- Trilling, B., & Fadel, C. (2009). 21st Century Skills - learning for life in our times. Jossey-Bass.
- Van de Sande, E., Kleemans, M., Verhoeven, L., & Segers, E. (2019). The linguistic nature of children's scientific reasoning. *Learning and Instruction*, 62(1), 20-26. <https://doi.org/10.1016/j.learninstruc.2019.02.002>
- Van der Graaf, J. (2020). Inquiry-based learning and conceptual change in balance beam understanding. *Frontiers in Psychology*, 11, Article 1621. <https://doi.org/10.3389/fpsyg.2020.01621>
- Van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instructional Science*, 43(3), 381-400. <https://doi.org/10.1007/s11251-015-9344-y>
- Van Graft, M., Klein Tank, M., Beker, T., & Van der Laan, A. (2018). Wetenschap en technologie in het basis- en speciaal onderwijs: Richtinggevend leerplankader bij het leergebied oriëntatie op jezelf en de wereld. [Science and technology in primary and special education: directive educational framework for orientation on self and world.]. SLO (nationaal expertisecentrum leerplanontwikkeling).
- Van Joolingen, W., & De Jong, T. (1991). Supporting hypothesis generation by learners exploring an interactive computer simulation. *Instructional Science*, 20, 389-404. <https://doi.org/10.1007/BF00116355>
- 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>



# Chapter 2

## **Individual differences in children's scientific reasoning**

## Abstract

Scientific reasoning is an important skill that encompasses hypothesizing, experimenting, interpreting data, evaluating data characteristics 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 school children completed a practical scientific reasoning test 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 test to measure their problem solving skills. As expected, children differed substantially from one another. Generally, scores were highest for experimenting, lowest for evaluating data characteristics and drawing conclusions, and intermediate for hypothesizing and interpreting data. 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.

This chapter is based on:

Schlatter, E., Lazonder, A. W., Molenaar, I., & Janssen, N. (2021).

Individual differences in children's scientific reasoning.

*Education Sciences*, 11(9), Article 471.

<https://doi.org/10.3390/educsci11090471>

## Introduction

Science education is an important part of the curriculum in many countries (Abd-El-Khalick et al., 2004; Achieve, 2010). 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 (Kuhn, 2002; Zimmerman, 2007) 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 (Kind & Osborne, 2017). In a school setting, scientific reasoning skills are particularly important for successful inquiry-based learning: 'minds-on' scientific reasoning skills (Kruit et al., 2018) are instrumental to achieving meaningful outcomes from a 'hands-on' inquiry.

Scientific reasoning consists of multiple component skills, namely hypothesizing, experimenting and evaluating outcomes, the latter of which can be further divided in interpreting data, evaluating data characteristics and drawing conclusions (Pedaste et al., 2015; Schiefer et al., 2019). 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., Piekny & Maehler, 2013). However, most existing research either treats scientific reasoning as a unitary construct or looks at one specific component skill of scientific reasoning – most often experimenting (Koerber & Osterhaus, 2019). 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 school 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 with physical materials, 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, 2013), 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 in children's proficiency in different component skills of scientific reasoning, when applied in a practical, coherent inquiry-based setting in order to ultimately aid the development of teaching materials for various groups of learners in primary education.

*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 outcomes (drawing valid conclusions). Evaluation of outcomes, in turn, involves interpreting data (making a verbal interpretation of the gathered data), evaluating data characteristics (assessing measurement quality, for instance to decide whether there is 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 (Edelsbrunner & Dablander, 2018) 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 (Köksal-Tuncer & Sodian, 2018; Van der Graaf et al., 2015) and older children can be taught this skill successfully by both direct instruction (Chen & Klahr, 1999; Lorch et al., 2017) and guided inquiry (Kuhn & Dean, 2005; Lazonder & Wiskerke-Drost, 2015). Hypothesizing is more difficult for children to learn (Köksal-Tuncer & Sodian, 2018; Lazonder & Janssen, 2021; Piekny & Maehler, 2013), whilst evaluating outcomes is the most difficult skill for them to acquire (Zimmerman, 2007) and is also experienced as such (Van Uum et al., 2017).

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 (2013), who inferred the age at which children learn hypothesizing, experimenting and evaluating outcomes 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 each component skill rather than a single test 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 to 6, Koerber et al. (2015) 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 (2013) 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 do 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 (Koerber et al., 2015; Piekny & Maehler, 2013), albeit with substantial variation. As not all component skills 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.

### *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 suggests that reading comprehension, numerical ability and problem solving skill contribute to scientific reasoning.

Reading comprehension most consistently explains children's overall scientific reasoning performance on written tests (Koerber et al., 2015; Mayer et al., 2014) as well as their ability to set up unconfounded experiments using the Control-of-Variables Strategy (Siler et al., 2010; Wagensveld et al., 2014). Van de Sande et al. (2019) found that reading comprehension explained the variance in all component scientific reasoning skills, albeit not to the same extent: effect sizes ranged from medium ( $r = .30$ ) for experimentation and drawing conclusions, to large ( $r = .47$ ) for hypothesis-evidence coordination. Why reading comprehension is such a strong predictor is not entirely clear. Possibly, reasoning ability transcends the domains of reading and science (Siler et al., 2010; Van de Sande et al., 2019), or a general understanding of the language of science is important for science learning (Snow, 2010). 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 relied heavily on the use of paper-and-pencil tests a further scrutiny of its role is warranted.

Numerical ability is often named as a prerequisite for scientific reasoning by national curriculum agencies (Van Graft et al., 2018; Wong, 2019) as well as scientists (Schauble, 2018) – likely because scientific reasoning, in particular skills involved in evaluating outcomes require reasoning about numerical data (Krummenauer & Kuntze, 2019; Makar et al., 2011; Mayer et al., 2014; Piekny & Maehler, 2013). Yet, empirical evidence for this relation is scarce. Early work by Bullock and Ziegler (1999) demonstrated that numerical intelligence predicts the growth of experimentation skills in primary school children, explaining almost 35 percent of the variance in a quadratic growth model. More recent studies found significant correlations between numerical ability and scientific reasoning (Koerber & Osterhaus, 2019; Tajudin & Chinnappan, 2015). 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 to the same extent.

Children’s problem solving skill is another possible predictor of scientific reasoning. Klahr and Dunbar (1988) 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 school children 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 (Piaget, 2003). Recent research supports these ideas: Mayer et al. (2014) found that problem solving predicts a substantial portion of the variance in children’s scientific reasoning. Van de Sande et al. (2019) further showed that this effect does not apply to all component skills: hypothesis-evidence coordination and experimenting depended on problem solving, whereas drawing 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.

### *Current study*

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 component skills 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 (Shavelson et al., 1991). Moreover, written tests of scientific reasoning can confound with reading comprehension: 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 component skills 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 test so as to maximize authenticity and minimize the influence of reading skills. A sample of 160 upper primary school children took this test to gauge their proficiency in five scientific reasoning skills: hypothesizing, experimenting, making inferences, evaluating data characteristics 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 component skills were also predicted to occur. Specifically, children were expected to be most proficient in experimenting, less proficient in hypothesizing and least proficient in the three outcome evaluation skills (interpreting data, evaluating data characteristics 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 component skills, these characteristics were expected to have differential effects.

## Method

### *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 was obtained for 160 of the 166 participating children (54% boys,  $M_{\text{age}} = 11$  years 0 months,  $SD = 9$  months); 84 of these children were in grade 5 (52% boys,  $M_{\text{age}} = 10$  years 5 months,  $SD = 7$  months) and 76 of them in grade 6 (55% boys,  $M_{\text{age}} = 11$  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 under project number 15460. 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 children were familiar with most tests. The school's science curriculum contained five annual hands-on science projects which enabled children to practice their scientific reasoning.

### *Materials*

#### *Scientific reasoning test*

Children's scientific reasoning skills were gauged during a 20 minute performance-based scientific reasoning test under supervision of a test administrator (Lazonder & Janssen, 2021). The test 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 an example, see Table 1). The test 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 component skill. Total test scores could range from 0 to 15 points. The Cohen's  $\kappa$  inter-rater agreement of the answer scoring was .84.

**Table 1***Example inquiry cycle*

Component skill	Question/assignment
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?
Interpreting data	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 <i>always</i> bounce more often on a hard surface?
Hypothesizing	The student who's next will also be doing this experiment. Imagine that student asks you to predict what the outcomes of their experiment will be. What would you say?

*Note.* 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 characteristics was assessed in subsequent research cycles.

Three versions of this test were available, which differed exclusively regarding the topic of investigation. In the *rolling balls* version, adapted from Chen and Klahr (1999), 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,  $F(2, 157) = 0.08, p = .925$ . Furthermore, a validation study (Lazonder & Janssen, 2021) showed no effects of prior content knowledge on the performance on 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 .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.

*Reading comprehension test*

Reading comprehension was measured by a standardized progress monitoring test developed by Cito, the Dutch national testing agency (Weekers et al., 2011). Different versions are available for different grades, and the test has a measurement accuracy between .87 and .89 (Weekers et al., 2011). 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 (Verhelst & Glas, 1995) was used to transform children's answers into a person proficiency score that can be meaningfully compared across grades.

*Numerical ability test*

Numerical ability was gauged by a standardized progress monitoring test that required children to add, subtract, multiply or divide one and two-digit numbers by heart (De Vos, 2006). The test consists of 200 items of increasing difficulty and is highly reliable ( $\alpha = .97$ ). Children worked on the test for 5 minutes and obtained 1 point for each correct answer.

*Problem solving test*

A digital version of the Tower of Hanoi adapted from Welsh (1991) was developed to assess children's problem solving skills. The test required children to solve as many problems as they could in seven minutes. One point was awarded for each solved problem and reliability was high ( $\alpha = .85$ ). The 20 problems required children to move differently sized disks from their starting position to their target position. Three simple rules limited the possible moves children could make: (1) only one disk could be moved at a time, (2) the disk could only be moved to an adjacent peg, and (3) 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 three 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 test ahead of time, the final problem was a 5-disk, 31-move problem. In practice, none of the children reached this final problem.

### *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 test. The problem solving test was administered in small groups. After a short explanation, children worked individually on the test for seven minutes. The scientific reasoning test was administered individually and lasted about 20 minutes per child.

### *Data analysis*

Data was analysed using IBM SPSS 25 (IBM, 2017). In order to answer the first research question, variation in scientific reasoning was explored using descriptive statistics; relations between the five component scientific reasoning skills were analysed 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**  
*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
Interpreting data	1.19	0.63	1.37	0.73	1.52	0.63
Evaluating data char.	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

## Results

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,  $F(1, 158) = 14.18, p < .001$ , partial  $\eta^2 = .08$ , numerical ability,  $F(1, 158) = 8.02, p = .005$ , partial  $\eta^2 = .05$ , and problem solving,  $F(1, 158) = 4.35, p = .039$ , partial  $\eta^2 = .03$ . The cross-grade differences in scientific reasoning were minor and were tested for statistical significance in the main analysis reported below.

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 component skills ranged from 0 to 3 except for interpreting data, where the lowest achieved score was 1 point. Means and standard deviations confirmed this differential ability and warrant 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 component skills: on average, children appeared to be most proficient in experimenting,



least proficient in evaluating data characteristics and drawing conclusions, while hypothesizing and interpreting data 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 component skill, Pillai's trace = .46,  $F(4, 153) = 32.50$ ,  $p < .001$ , but no interaction effects of component skill with gender, Pillai's trace = .02,  $F(4, 153) = 0.60$ ,  $p = .665$ , and grade, Pillai's trace = .03,  $F(4, 153) = 1.23$ ,  $p = .300$ . The differences between component skills were further explored in univariate analyses. Scores on experimenting were significantly higher than scores on all other component skills,  $p < .01$ . Scores on hypothesizing were significantly higher than scores on interpreting data, evaluating data characteristics and drawing conclusions,  $p < .05$ . Scores on interpreting data were significantly higher than scores on evaluating data characteristics,  $p < .01$ , but not than scores on drawing conclusions,  $p = .214$ . Drawing conclusions and evaluating data characteristics, the two component skills with the lowest scores, were not significantly different from one another,  $p = .993$ .

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

**Table 3***Correlations for predictors and scientific reasoning component skills.*

	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Scientific reasoning <sup>1</sup>	—								
2. Hypothesizing	.61**	—							
3. Experimenting	.58**	.15	—						
4. Interpreting data	.53**	.18*	.11	—					
5. Evaluating data char.	.46**	.14	.14	.06	—				
6. Drawing conclusions	.63**	.17*	.16*	.28**	.08	—			
7. Reading comprehension	.39**	.15	.29*	.18*	.31**	.20*	—		
8. Numerical ability	.18*	.13	.04	.09	.18*	.07	.27**	—	
9. Problem solving	.17*	.07	.08	.09	.13	.11	.14	.15	—

*Note.*  $N = 160$ .<sup>1</sup>Total score\* $p < .05$  \*\* $p < .01$ 

Multivariate multiple regression was used to further scrutinize the relations between the three predictor variables and the five scientific reasoning component skills. Multivariate test results showed no main effect of the control variables gender, Pillai's trace = .01,  $F(5, 150) = 0.35$ ,  $p = .882$ , partial  $\eta^2 = .01$ , and grade, Pillai's trace = .05,  $F(5, 150) = 1.60$ ,  $p = .164$ , partial  $\eta^2 = .51$ . Regarding the predictor variables, a significant contribution of reading comprehension on scientific reasoning was found, Pillai's trace = .17,  $F(5, 150) = 6.28$ ,  $p < .001$ , partial  $\eta^2 = .17$ . Neither numerical ability, Pillai's trace = .02,  $F(5, 150) = 0.57$ ,  $p = .725$ , partial  $\eta^2 = .02$ , nor problem solving skills, Pillai's trace = .02,  $F(5, 150) = 0.61$ ,  $p = .694$ , partial  $\eta^2 = .02$ , explained scientific reasoning to a significant degree. The between-subject effects of reading comprehension in Table 4 show that reading comprehension accounted for a significant proportion of the variance in experimenting, interpreting data, evaluating data characteristics and drawing conclusions, but not in hypothesizing. The regression coefficients further indicate that experimenting was most influenced by reading comprehension. Of the significantly predicted component skills, interpreting data was least influenced by reading comprehension. So, although reading comprehension remains an important explanatory factor, it did not explain all scientific reasoning component skills uniformly.

**Table 4***Reading comprehension as predictor of the scientific reasoning component skills.*

Component skills	$\beta$	t	p	95% CI	Partial $\eta^2$
Hypothesizing	.01	1.82	.071	[-.00, .02]	.02
Experimenting	.02	4.09	<.001	[.01, .03]	.10
Interpreting data	.01	2.10	.037	[.00, .01]	.03
Evaluating data char.	.01	3.23	.001	[.00, .02]	.06
Drawing conclusions	.01	2.43	.016	[.00, .02]	.04

## Discussion

This study aimed to identify and explain differences in children's ability to reason scientifically. To this end, a performance-based scientific reasoning test was administered and measures of reading comprehension, numerical ability and problem solving skill 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 skill.

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 component skills. The hypothesized proficiency pattern was confirmed: children in our sample were most proficient in experimenting, less proficient in hypothesizing, and least proficient in interpreting data, evaluating data characteristics and drawing conclusions. This is particularly important because, as Koerber and Osterhaus (2019) argued, previous research has studied these component skills separately, often through written tests (Mayer et al., 2014; Van de Sande et al., 2019). The present study thus confirms the differences in component skill difficulty during a comprehensive performance-based scientific reasoning test, and suggests that children's relative proficiency at the component skill level is stable across test modalities (cf. Kruit et al., 2018).

The observed variation in scientific reasoning was independent of children's grade level. This equivalence of test 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 did 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., Piekny & Maehler, 2013). Although most children at this age do advance in scientific reasoning (Lazonder & Janssen, 2021), the inter-individual variation is considerable and prevents the minor cross-grade growth differences to become statistically significant. Alternative research methods such as longitudinal designs and person-centred approaches to data analysis are more sensitive to capturing developmental growth and are increasingly being applied in scientific reasoning research (Hickendorff et al., 2018).

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., Mayer et al., 2014; Snow, 2010; Van de Sande et al., 2019). So 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 an 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., Veenman et al., 2004). In addition, relations have been found between scientific reasoning and verbal reasoning (Siler et al., 2010) as well as nonverbal reasoning (Van de Sande et al., 2019) and conditional sentence comprehension (Svirko et al., 2019). In line with these findings, language-centred scientific reasoning interventions have been proposed (Svirko et al., 2019; Van de Sande et al., 2019) and found effective (Van der Graaf et al., 2019).

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 component 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, which they do in

hypothesizing, as opposed to building a situation model from given information, which they do in reading (Swart et al., 2017) as well as in evaluating 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 (Koerber & Osterhaus, 2019) or involved a different math strand (Bullock & Ziegler, 1999). 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 test.

Children's problem solving skills did not predict scientific reasoning either, possibly because of task incongruence. Jonassen (2000) 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. So 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 (Jonassen, 2000) might explain why Mayer et al. (2014) did find the very similar Tower of London problem to explain scientific reasoning. Mayer et al. (2014) used a multiple choice paper-and-pencil version of this problem in which all manipulations had to be done 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 do so 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 did. As understanding what explains specific component skills is only a recent endeavour (Koerber & Osterhaus, 2019; Van de Sande et al., 2019), more research is needed to understand which component skills can be explained as well as why differential effects exist.

### *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 (Archer et al., 2016), this might have given the participants in the current study a certain advantage compared to children whose parents are less highly 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 school children.

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

### *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 (Rönnebeck et al., 2016), which stands to reasons 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 component skills, and the relations between these skills. Therefore, future research should focus more on scientific reasoning in authentic inquiry-based learning settings, while still distinguishing component skills.

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 biweekly inquiry-based science lessons, teachers should attend to differences between children and among component skills. 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. Interpreting data, evaluating data characteristics and drawing conclusions, which are the most difficult component skills, 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, sudoku's) 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 a flexible scheduling and combine the fifth- and sixth-grade lessons because the proficiency levels in these classes is 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.

To conclude, this study found substantial overall differences in children's scientific reasoning as well as marked differences at the component skill 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 component 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

- Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., Niaz, M., Treagust, D., & Tuan, H. L. (2004). Inquiry in science education: International perspectives. *Science Education*, 88(3), 397-419. <https://doi.org/10.1002/scs.10118>
- Achieve. (2010). International science benchmarking report Taking the Lead in Science Education: Forging Next-Generation Science Standards.
- Archer, L., Dawson, E., Seakins, A., & Wong, B. (2016). Disorientating, fun or meaningful? Disadvantaged families' experiences of a science museum visit. *Cultural Studies of Science Education*, 11(4), 917-939. <https://doi.org/10.1007/s11422-015-9667-7>
- 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.
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, 70(5), 1098-1120. <https://doi.org/10.1111/1467-8624.00081>
- De Vos, T. (2006). Schoolvaardigheidstoets hoofdrekenen [arithmetic proficiency test for primary school]. Boom Test Uitgevers.
- Edelsbrunner, P. A., & Dablander, F. (2018). The psychometric modeling of scientific reasoning: A review and recommendations for future avenues. *Educational Psychology Review*, 31(1), 1-34. <https://doi.org/10.1007/s10648-018-9455-5>
- Harlen, W. (2013). Inquiry-based learning in science and mathematics. *Review of Science, Mathematics and ICT education*, 7(2), 9-33. <https://doi.org/10.26220/rev.2042>
- Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2018). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning and Individual Differences*, 66, 4-15. <https://doi.org/10.1016/j.lindif.2017.11.001>
- IBM. (2017). IBM SPSS Statistics for Windows. In (Version 25)
- Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, 48(4), 63-85. <https://doi.org/10.1007/BF02300500>
- Kind, P. M., & Osborne, J. (2017). Styles of scientific reasoning: A cultural rationale for science education? *Science Education*, 101(1), 8-31. <https://doi.org/10.1002/scs.21251>
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12(1), 1-48. [https://doi.org/10.1207/s15516709cog1201\\_1](https://doi.org/10.1207/s15516709cog1201_1)
- Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development*, 86(1), 327-336. <https://doi.org/10.1111/cdev.12298>
- 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.org/10.1080/15248372.2019.1620232>
- Köksal-Tuncer, Ö., & Sodian, B. (2018). The development of scientific reasoning: Hypothesis testing and argumentation from evidence in young children. *Cognitive Development*, 48, 135-145. <https://doi.org/10.1016/j.cogdev.2018.06.011>
- Kruit, P. M., Oostdam, R. J., Van den Berg, E., & Schuitema, J. A. (2018). Assessing students' ability in performing scientific inquiry: Instruments for measuring science skills in primary education. *Research in Science & Technological Education*, 36(4), 413-439. <https://doi.org/10.1080/02635143.2017.1421530>

- Krummenauer, J., & Kuntze, S. (2019, February). *Primary students' reasoning and argumentation based on statistical data*. Eleventh Congress of the European Society for Research in Mathematics Education, Utrecht, the Netherlands. <https://hal.archives-ouvertes.fr/hal-02398118/>
- Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 371-393). Blackwell Publishers Ltd.
- Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, 16(11), 866-870. <https://doi.org/10.1111/j.1467-9280.2005.01628.x>
- Lazonder, A. W., & Janssen, N. (2021). Development and initial validation of a performance-based scientific reasoning test for children. *Studies in Educational Evaluation*, 68, Article 100951. <https://doi.org/10.1016/j.stueduc.2020.100951>
- Lazonder, A. W., & Wiskerke-Drost, S. (2015). Advancing scientific reasoning in upper elementary classrooms: Direct instruction versus task structuring. *Journal of Science Education and Technology*, 24(1), 69-77. <https://doi.org/10.1007/s10956-014-9522-8>
- Lorch, R. F., Lorch, E. P., Freer, B., Calderhead, W. J., Dunlap, E., Reeder, E. C., Van Neste, J., & Chen, H. T. (2017). Very long-term retention of the control of variables strategy following a brief intervention. *Contemporary Educational Psychology*, 51, 391-403. <https://doi.org/10.1016/j.cedpsych.2017.09.005>
- Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, 13(1-2), 152-173. <https://doi.org/10.1080/10986065.2011.538301>
- 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>
- 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. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, 14, 47-61. <https://doi.org/10.1016/j.edurev.2015.02.003>
- Piaget, J. (2003). Development and learning. *Journal of Research in Science Teaching*, 40(1), 8-18. <https://doi.org/10.1002/tea.3660020306> (Original work published 1964)
- Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology*, 31(2), 153-179. <https://doi.org/10.1111/j.2044-835X.2012.02082.x>
- Rönnebeck, S., Bernholt, S., & Ropohl, M. (2016). Searching for a common ground – A literature review of empirical research on scientific inquiry activities. *Studies in Science Education*, 52(2), 161-197. <https://doi.org/10.1080/03057267.2016.1206351>
- Schauble, L. (2018). In the eye of the beholder: Domain-general and domain-specific reasoning in science. In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation* (pp. 11-33). Routledge.
- Schiefer, J., Golle, J., Tibus, M., & Oschatz, K. (2019). Scientific reasoning in elementary school children: Assessment of the inquiry cycle. *Journal of Advanced Academics*, 30(2), 144-177. <https://doi.org/10.1177/1932202x18825152>
- Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, 4(4), 347-362. [https://doi.org/10.1207/s15324818ame0404\\_7](https://doi.org/10.1207/s15324818ame0404_7)
- Siler, S. A., Klahr, D., Magaro, C., Willows, K., & Mowery, D. (2010, June). *Predictors of transfer of experimental design skills in elementary and middle school children*. 10th International Conference on Intelligent Tutoring Systems, Pittsburgh, PA.

- Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science*, 328(5977), 450-452. <https://doi.org/10.1126/science.1182597>
- Svirko, E., Gabbott, E., Badger, J., & Mellanby, J. (2019). Does acquisition of hypothetical conditional sentences contribute to understanding the principles of scientific enquiry? *Cognitive Development*, 51, 46-57. <https://doi.org/10.1016/j.cogdev.2019.05.008>
- Swart, N. M., Muijselaar, M. M. L., Steenbeek-Planting, E. G., Droop, M., De Jong, P. F., & Verhoeven, L. (2017). Cognitive precursors of the developmental relation between lexical quality and reading comprehension in the intermediate elementary grades. *Learning and Individual Differences*, 59, 43-54. <https://doi.org/10.1016/j.lindif.2017.08.009>
- Tajudin, N. M., & Chinnappan, M. (2015). *Exploring relationship between scientific reasoning skills and mathematics problem solving*. 38th annual conference of the Mathematics Education Research Group of Australasia: Mathematics education in the margins, Sunshine Coast, Australia.
- Van de Sande, E., Kleemans, M., Verhoeven, L., & Segers, E. (2019). The linguistic nature of children's scientific reasoning. *Learning and Instruction*, 62(1), 20-26. <https://doi.org/10.1016/j.learninstruc.2019.02.002>
- Van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instructional Science*, 43(3), 381-400. <https://doi.org/10.1007/s11251-015-9344-y>
- Van der Graaf, J., van de Sande, E., Gijssel, M., & Segers, E. (2019). A combined approach to strengthen children's scientific thinking: Direct instruction on scientific reasoning and training of teacher's verbal support. *International Journal of Science Education*, 41(9), 1119-1138. <https://doi.org/10.1080/09500693.2019.1594442>
- Van Graft, M., Klein Tank, M., Beker, T., & Van der Laan, A. (2018). Wetenschap en technologie in het basis- en speciaal onderwijs: Richtinggevend leerplankader bij het leergebied oriëntatie op jezelf en de wereld. [Science and technology in primary and special education: directive educational framework for orientation on self and world.]. SLO (nationaal expertisecentrum leerplanontwikkeling).
- Van Uum, M. S. J., Verhoeff, R. P., & Peeters, M. (2017). Inquiry-based science education: Scaffolding pupils' self-directed learning in open inquiry. *International Journal of Science Education*, 39(18), 2461-2481. <https://doi.org/10.1080/09500693.2017.1388940>
- Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction*, 14(1), 89-109. <https://doi.org/10.1016/j.learninstruc.2003.10.004>
- Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 215-239). Springer. [https://doi.org/10.1007/978-1-4612-4230-7\\_12](https://doi.org/10.1007/978-1-4612-4230-7_12)
- Wagensveld, B., Segers, E., Kleemans, T., & Verhoeven, L. (2014). Child predictors of learning to control variables via instruction or self-discovery. *Instructional Science*, 43(3), 365-379. <https://doi.org/10.1007/s11251-014-9334-5>
- Weekers, A., Groenen, I., Kleintjes, F., & Feenstra, H. (2011). *Wetenschappelijke verantwoording papieren toetsen begrijpend lezen voor groep 7 en 8* [Scientific validation paper-and-pencil tests reading comprehension for grade 5 and 6]. Cito.
- Welsh, M. C. (1991). Rule-guided behavior and self-monitoring on the tower of hanoi disk-transfer task. *Cognitive Development*, 6(2), 59-76. [https://doi.org/10.1016/0885-2014\(91\)90006-Y](https://doi.org/10.1016/0885-2014(91)90006-Y)

- Wong, V. (2019). 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. *International Journal of Science Education*, 41(13), 1805-1826. <https://doi.org/10.1080/09500693.2019.1641249>
- 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>





# Chapter 3

**Individual differences in  
children's development of  
scientific reasoning  
through inquiry-based  
instruction:  
Who needs additional  
guidance?**

## Abstract

Scientific reasoning involves a person's ability to think and act in ways that help advance their understanding of the natural world. Young children are naturally inclined to engage in scientific reasoning and display an emerging competence in the component skills of hypothesizing, experimenting and evaluating outcomes. Developmental psychology research has shown that same-age children often differ considerably in their proficiency to perform these skills. Part of this variation comes from individual differences in cognition; another part is due to the fact that the component skills of scientific reasoning emerge at a different age and mature at a different pace. Significantly less attention has been paid to children's capacity to improve in scientific reasoning through instruction and deliberate practice. Although primary science lessons are generally effective to raise the skill level of a group of learners, not all children benefit equally from the instructional treatment they receive. Knowing what causes this differential effectiveness is important as it can inform the design of adaptive instruction and support. The present study therefore aimed to identify and explain how fifth graders ( $N = 138$ ) improve their scientific reasoning skills over the course of a five-week inquiry-based science lesson series. In line with our expectations, significant progress was observed in children's achievements on a written scientific reasoning test, which was administered prior to and after the lessons, as well as in their responses to the questions and assignments that appeared on the worksheets they filled out during each lesson. Children's reading comprehension and mathematical skilfulness explained a portion of the variance in children's gain from pre- to post-test. As these overall results did not apply equally to all component skills of scientific reasoning, we recommend science teachers to adapt their lessons based on children's past performance in reading and math *and* their actual performance of each scientific reasoning skill. The orchestration and relative effectiveness of both adaptive science teaching approaches is an interesting topic for future research.

This chapter is based on:

Schlatter, E., Molenaar, I., & Lazonder, A. W. (2020).

Individual differences in children's development of scientific reasoning through inquiry-based instruction: Who needs additional guidance?

*Frontiers in Psychology*, 11, Article 904. <https://doi.org/10.3389/fpsyg.2020.00904>



## Introduction

Primary science education acquaints children with fundamental science concepts such as buoyancy, motion and electricity, and introduces them to the basics of doing scientific research. School science lessons make ample use of inquiry-based teaching methods, which enable children to learn to think and act in ways that help advance their understanding of the natural world (Kind & Osborne, 2017). This ability is commonly referred to as scientific reasoning and involves the skills of hypothesizing, experimenting and evaluating outcomes (Kind, 2013; Klahr & Dunbar, 1988; Zimmerman, 2007). The main purpose of this study was to investigate how these skills develop during an inquiry-based science lesson series, and which cognitive characteristics predict children's level of skilfulness at the end of the lesson series.

The teaching and learning of scientific reasoning is a challenging task for both teachers and children. One complicating factor is that considerable individual variation exists among children in the same classroom (Koerber et al., 2015; Lazonder et al., 2021). To complicate matters further, the component skills of scientific reasoning are known to emerge at different ages and develop at a different pace (Pieknny & Maehler, 2013). Kindergartners already show some initial proficiency in basic experimentation and evidence evaluation whereas the more difficult skill of hypothesizing usually starts developing around the age of 12. These accumulating differences point to a clear need for adaptive instruction, but until now few evidence-based guidelines for designing and delivering adaptive and age-appropriate science lessons seem to exist.

In working toward establishing such guidelines, the present study sought to unveil whether and to what extent the progress monitoring data available in schools can help predict differences in children's ability to learn scientific reasoning. Many schools have access to rich data records that portray children's developmental trajectories in the foundation skills of language and math. As these skills are related to children's scientific reasoning *performance* (Tajudin & Chinnappan, 2015; Van de Sande et al., 2019), it seems worthwhile to investigate their predictive powers for the *development* of scientific reasoning in an instructional setting. Additionally, process data collected during the lessons was analysed in order to identify key learning moments. The insights that result from such investigations can help teachers to respond adequately to individual differences during their science lessons.

*Development of scientific reasoning*

Scientific reasoning is a multidimensional process that consists of several component skills. Although scholars diverge on the definition and labelling of these skills (see, for an overview, Pedaste et al., 2015), consensus seems to exist on the core skills of hypothesizing, experimenting and evaluating outcomes (Kind, 2013; Zimmerman, 2007). Even though these are difficult skills even for adults (Zimmerman, 2007), children at the pre-school age already show an emerging proficiency in some skills (Köksal-Tuncer & Sodian, 2018; Piekny et al., 2014; Sodian et al., 1991; Van der Graaf et al., 2015) that develops steadily but slowly during the primary school years (Koerber et al., 2015; Kuhn, 2002; Piekny & Maehler, 2013).

Although developmental growth occurs in all component skills, their emergence and pace of development varies. Experimenting is relatively easy to learn and even young children can be rather proficient in the basics of experimentation (Cook et al., 2011; Schalk et al., 2019; Van der Graaf et al., 2018). Hypothesizing is more difficult for children to learn (Piekny & Maehler, 2013; Schlatter et al., 2021 [chapter 2]) and this skill generally emerges late and develops slowly. Results regarding evaluating outcomes are mixed. First-graders can already draw correct conclusions from perfectly covarying data (Koerber et al., 2005; Piekny & Maehler, 2013; Van der Graaf et al., 2018), but the evaluation of nonperfect covarying evidence in light of hypotheses remains difficult throughout primary school (Piekny & Maehler, 2013).

In addition to this variation across component skills, same-age children are not equally well versed in scientific reasoning either. In a large-scale cross-sectional study using written tests in grades 2 to 4, Koerber et al. (2015) distinguished between naïve, intermediate, and advanced conceptions of scientific reasoning. Although older children generally had a more sophisticated view, all three proficiency levels were present in all participating grade levels. The cross-sectional results of Piekny and Maehler (2013) further suggest that these inter-individual differences increase with age in all component skills. For example, both the means and standard deviations of hypothesizing were low in Kindergarten, but increased from grade 1 onward. These findings indicate that, although children's hypothesizing skills undergo a steady growth, the variation among peers grows accordingly. Thus, children improve in scientific reasoning with age, but not all children improve at the same pace. Acknowledging these individual differences alongside the dissimilar difficulty levels of the component skills is vital for good science education.

*Predictors of scientific reasoning*

Several studies have examined what accounts for observed differences in children's scientific reasoning (Koerber et al., 2015; Mayer et al., 2014; Siler et al., 2010; Wagenveld et al., 2014). Reading comprehension was a significant predictor in all these studies, whereas cognitive characteristics such as spatial reasoning, problem solving skill, and general intelligence had a less prominent and less consistent impact. Although mathematical skilfulness has been shown to correlate with scientific reasoning (Bullock & Ziegler, 1999; Koerber & Osterhaus, 2019), to the best of our knowledge, no studies have examined whether mathematical skilfulness *predicts* scientific reasoning. This seems remarkable because scientific reasoning tasks often require children to handle numerical data (Kanari & Millar, 2004), which, in turn, could be the reason why national curriculum agencies consider mathematical skilfulness as a prerequisite for scientific reasoning instruction (e.g., Van Graft et al., 2018; Wong, 2019).

Research into the predictors of scientific reasoning either treated scientific reasoning as a unitary construct or focused on one of its components, with experimenting being the most widely studied skill. Studies that assess and report children's performance on multiple component skills are clearly underrepresented in the literature and, as a consequence, little is known about how well reading comprehension and mathematical skilfulness predict children's proficiency in separate scientific reasoning skills. Initial evidence suggests that both predictors may have differential effects. Schlatter et al. (2021) [Chapter 2] established that reading comprehension predicts performance on all component skills except hypothesizing, whereas Van de Sande et al. (2019), who administered a written test, found a strong explanatory effect of reading comprehension on this skill and lower impacts on experimenting and drawing conclusions. Osterhaus et al. (2017) found no effect of language abilities on experimenting – but it did influence children's understanding of the nature of science. These findings, although apparently contradictory, emphasize the importance of analyses at the component skill level. Furthermore, as these studies examined children's scientific reasoning *performance*, research still has to determine whether and to what extent reading comprehension and mathematical skilfulness affect and predict children's *learning* of the component skills of scientific reasoning in regular science classrooms.

*In-school learning of scientific reasoning*

Studies examining the development of scientific reasoning in an instructional setting predominantly target children's ability to design and conduct controlled experiments. The

natural development of this skill can be boosted in a short period of time through various instructional methods that often yield long-term effects. Implicit methods such as giving hints to focus the investigation on a single variable (Kuhn & Dean, 2005), dividing the research question in single-variable subquestions (Lazonder & Wiskerke-Drost, 2015), providing scaffolds (Van Riesen et al., 2018) or opportunities for sustained practice (Schalk et al., 2019) all improve children's experimentation skills. Explicit instructional methods that explain and/or demonstrate the design of controlled experiments have similar benefits (Chen & Klahr, 1999; Lorch et al., 2017). A recent meta-analysis substantiated that implicit and explicit methods are equally effective for promoting experimenting skills (Schwichow et al., 2016).

The skills of hypothesizing and evaluating outcomes have less often been trained in isolation, but are included in integrated studies of scientific reasoning, often using microgenetic designs. In a three-year longitudinal study, Kuhn and Pease (2008) found that repeated practice alone promotes children's evidence evaluation skills throughout grades 4 to 6. Hypothesizing skills improved only when children were in sixth grade, despite frequent opportunities for practice in the preceding years, and individual change patterns in both skills varied considerably – including relapses to old, less-effective routines. More explicit instructional support can accelerate children's natural pace of development. Greven and Letschert (2006) showed that sixth graders who merely investigated a multivariable system did improve their ability to evaluate evidence over the course of a five-week inquiry-based lesson series. However, significantly higher learning gains were observed in children who received additional prediction practice exercises (that focused their attention on integrating the impact of multiple variables) or explicit instruction on the concept of multivariable causality.

To conclude, the cited studies exemplify that even short instructional interventions can promote children's scientific reasoning. Prolonged opportunities for practice have similar beneficial effects but seem more difficult to realize in regular science classrooms. Striking the right balance between independent practice and instructional guidance thus seems a major challenge for primary science teachers. This orchestration of instructional support is complicated further by the substantial variation across the component skills and among same-aged children. As a large share of this variance remains unexplained, the present study aimed to describe and explain children's development of scientific reasoning skills in inquiry-based classrooms.

*Research questions and hypotheses*

Previous research has shown that the component skills of scientific reasoning are not equally well developed and learned in upper primary science classrooms. Although these individual differences are explained in part by children's cognitive characteristics, with reading comprehension being the most robust predictor, questions remain as to how the core scientific reasoning skills of hypothesizing, experimenting and evaluating outcomes develop in an instructional setting, and how developmental differences can be adequately accommodated by primary science teachers. The present study therefore aimed to find out:

- (1) To what extent do fifth graders improve their scientific reasoning skills during a five-week inquiry-based lesson series?
- (2) Are the observed differences in learning gains contingent on children's reading comprehension and mathematical skilfulness?
- (3) Are there any key moments during this lesson series where children make marked progress in their application of the component scientific reasoning skills?

These research questions were examined in a sample of Dutch fifth graders, who engaged in five weekly science lessons. Each lesson revolved around a hands-on investigation using physical materials, which enabled children to practice the component skills of hypothesizing, experimenting and evaluating outcomes. Children's investigations were guided by worksheets and a whole-class introduction to the steps of the inquiry cycle. Learning gains were assessed by a written scientific reasoning pre-test and post-test. Learning process data were collected from the children's worksheets, and children's scores on standardized progress monitoring tests of reading comprehension and mathematical skilfulness were obtained from the schools' administration.

Hypotheses regarding the first research question predicted that children would make progress in all scientific reasoning skills – but not to the same degree. As previous research has shown that hypothesizing and evaluating outcomes is largely beyond fifth graders' reach, these skills were expected to improve marginally and comparably in just five lessons. Experimenting, on the other hand, is known to be relatively easy, so children might already be rather adept in this skill and, hence, have less opportunity for improvement compared to the other skills. However, in absence of a national science curriculum and instigated by recent policy measures, many Dutch primary schools are just beginning to systematically incorporate science in their curriculum (Inspectie van het Onderwijs, 2017), a rival hypothesis therefore predicted that children's experimentation skills are initially lower than

expected based on international benchmarking studies, but will improve more rapidly over the course of the five lessons than the other component skills – a result more often observed in intervention studies (Lorch et al., 2014; Peteranderl, 2019).

The second set of hypotheses related to the prediction of learning progress. Even when no overall learning gain is found, part of the sample could have made significant progress. To explain such possibly differential progress, two predictor variables were used: reading comprehension and mathematical skilfulness. Previous studies have shown that the former consistently predicts individual differences in scientific reasoning performance. We therefore felt it safe to assume that reading comprehension would explain learning gains in all three component skills. Evidence regarding the impact of children's mathematical skilfulness is limited, but existing studies suggest that 'being good with numbers' serves as an advantage when interpreting the numerical outcomes of science experiments (Bullock & Ziegler, 1999; Koerber & Osterhaus, 2019). Children's mathematical skilfulness was therefore expected to predict learning gains in evaluating outcomes.

Thirdly, children's worksheets were scrutinized for evidence of possible growth spurts in children's learning of the three component scientific reasoning skills. In absence of any theoretical and empirical underpinnings, no explicit hypothesis was made regarding the outcome of this analysis.

## Method

This study was carried out in accordance with the recommendations of the ethics code for research with human participants in the social and behavioural sciences, as agreed upon by the Deans of Social Sciences in the Netherlands. The protocol was approved by the Ethics Committee of the Faculty of Social Sciences at Radboud University, under number 2018-074R1. Descriptive data (gender and year of birth) were collected anonymously while other data (pre- and post-test, worksheets and standardized test scores) were pseudonymized.

### *Participants*

In the Fall of 2018, eight fifth grade classes (in Dutch: 'groep 7') from six schools in the central and northern part of the Netherlands participated in this study. All children in these classrooms received five one-hour lessons as part of their regular science curriculum. Passive parental consent was sought with the exception of one school, whose principal preferred active parental permission for participation. Children with parental consent ( $N = 154$ ) also took a scientific reasoning pre- and post-test; the worksheets they filled out during the

lessons were collected for analysis, and their progress monitoring scores on standardized tests of reading comprehension and mathematical skilfulness were obtained from the school. Sixteen children were excluded from the analyses, either because they missed more than one lesson, had not taken the pre- or post-test, or because their reading and math progress monitoring records could somehow not be obtained. The final sample thus consisted of 138 participants (55% boys) who were between 8 and 12 years of age; the majority of the sample was 10 years old.

## *Materials*

### *Lesson materials*

Children engaged in five science lessons that addressed primary-school physics topics (see Figure 1) through an inquiry-based teaching approach, taught by the principal investigator. All lessons were structured similarly and contained two types of activities: whole-class discussion and small-group work. Each lesson started with a plenary introduction (lesson 1) or refresher (lessons 2-5) of the inquiry cycle and introduced children to the topic of inquiry. Children then started their first inquiry, which they completed in 20 minutes. In order to mimic authentic classroom practice, children conducted their investigation in dyads, which they formed themselves on an ad-hoc basis. As children chose their learning partners based on friendship rather than academic achievement and partnerships rotated during the lesson series, the chances of any systematic bias due to group formation were assumed to be negligible. The first inquiry was wrapped up during a short whole-class discussion that addressed questions such as ‘who found an answer to the research question?’ and ‘who found a different result than hypothesized?’. After the second 20-minute inquiry cycle, children reconvened for a final whole-class discussion of the outcomes of the inquiry and the underlying physics principles.

The lessons were designed to practice four scientific reasoning skills: hypothesizing, experimenting, interpreting data, and drawing conclusions. Each lesson centred around a different subject-specific topic (see Figure 1) that children could learn about through experimentation. All experiments had three dichotomous input variables and one continuous output variable. For example, the pendulum swing experiment enabled children to manipulate the length of the rope (long or short), the weight of the pendulum (heavy or light), and the amplitude (far or close). Children used a stopwatch to measure the time it took to make five swings. In a typical lesson, the experimental equipment was used during

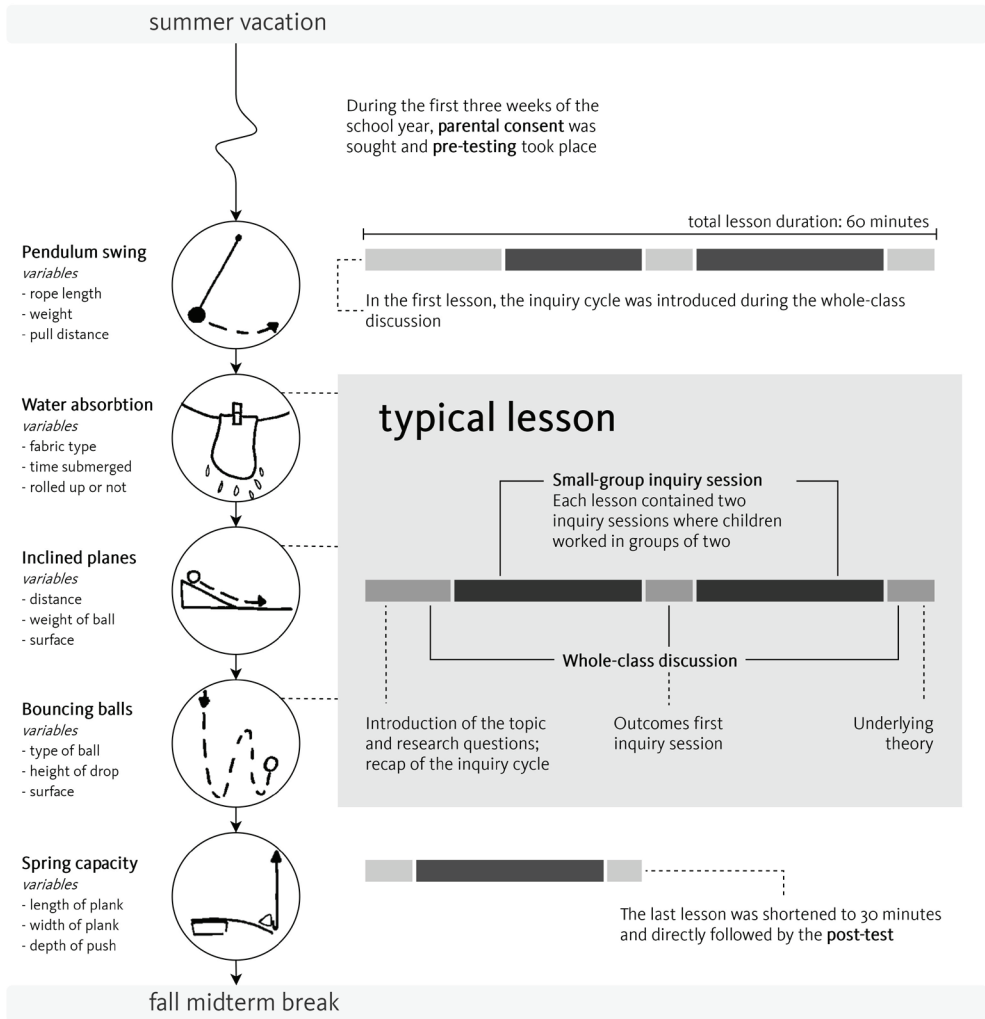
two inquiry sessions that were structured according to the inquiry cycle and enabled children to investigate two distinct research questions.

All inquiry sessions were supported by worksheets (see Appendix A) that assisted children in performing the four scientific reasoning skills. This guidance consisted of a pre-specified research question and several scaffolds that structured children's inquiry without explicitly instructing them what to do and why. Specifically, children could complete sentence starters to make their hypotheses and conclusions, and complete pre-structured tables to set up an experiment and interpret the results. The worksheets contained text and pictures that served to remind children of the research question and the variables under investigation (see, for an overview of inquiry topics and variables, see Figure 1).

This amount of support, defined by Bell et al. (2005) as guided inquiry, purposefully constrained the number of strategies children could apply, and has been shown to facilitate the learning of scientific reasoning skills (e.g., Van Riesen et al., 2018). For example, providing a research question minimized the risk of children conceiving a research question that could not be investigated, while still providing them with a fair degree of autonomy in their inquiry. The worksheets thus had a dual purpose: in addition to being a supportive device, they served as a measure of children's progress in scientific reasoning. Even though children conducted their investigations in dyads, they wrote down what they themselves thought to be the best hypothesis, experiment, interpretation of the data they gathered, and conclusion. As such, this process data could be used to identify where additional support was needed, and thus inform future research on adaptive science instruction.



**Figure 1**  
*Outline of the lesson series*



Supplementary to the worksheets, more elaborate support was given during whole-class discussions and to individual children who indicated they were struggling with the assignments. Children who struggled were first prompted to write down what they thought was best. If they were still hesitant to work on the worksheet, guidance was slowly increased following the protocol in Appendix B. In practice, children rarely asked for help and no child asked help repeatedly for the same component skill. During the whole-class

discussions, children were invited to share what they remembered about the inquiry cycle, what they found out during their investigations and what they thought were the underlying scientific principles. If answers were limited (e.g., ‘we found that it made a difference’), children were encouraged to provide more detail (e.g., ‘can you explain more precisely what you found?’).

### *Scientific Reasoning Inventory*

Children’s scientific reasoning skills were assessed at pre- and post-test using the Scientific Reasoning Inventory (SRI; Van de Sande et al., 2019), a paper-and-pencil test consisting of 24 multiple-choice items with three to four answer options each. Items were thematically embedded in five cover stories that are meaningful and appealing to children, such as the living conditions of wildlife and sports activities.

During the original validation of the SRI, three scales emerged: hypothesis validation (which included data interpretation), experimentation and drawing conclusions (Van de Sande et al., 2019). Confirmatory factor analysis was performed and results, including the comparative fit index (CFI), root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR) are reported below. In our pre-test data, a single-factor solution had a rather poor fit,  $\chi^2(252) = 437.04$ ,  $p < .001$ , CFI = 0.608, RMSEA = 0.067, SRMR = 0.081. The original three-factor model had a better fit,  $\chi^2(249) = 354.99$ ,  $p < .001$ , CFI = 0.776, RMSEA = 0.051, SRMR = 0.075, and the four-factor model, with data interpretation as a separate factor, yielded comparable fit statistics,  $\chi^2(246) = 352.15$ ,  $p < .001$ , CFI = 0.775, RMSEA = 0.051, SRMR = 0.074. While the improvement from the single-factor model to the three-factor model was significant,  $\chi^2_{\text{diff}}(3) = 82.05$ ,  $p < .001$ , the improvement from the three-factor model to the four-factor model was not,  $\chi^2_{\text{diff}}(3) = 2.84$ ,  $p = .417$ . We therefore decided to use the original three scales in the analyses. As a consequence, there was no one-on-one match between the SRI-scales and the skills addressed by the worksheets. Specifically, hypothesizing and interpreting outcomes were separate skills on the worksheets but combined in one SRI-scale, which we labelled ‘hypothesis-evidence coordination’.

This *hypothesis-evidence coordination* scale (9 items,  $\alpha_{\text{pre-test}} = .66$ ,  $\alpha_{\text{post-test}} = .74$ ), consisted of two types of items. Five items presented children with four research questions, and asked them to select the question that best matched the research purpose described in the cover story. The nature of these items closely resembled the way in which the skill of hypothesizing was addressed during the lessons. Four additional items measured children’s

ability to interpret a table with research data. These questions related to the skill of interpreting data as was addressed during the lessons. Although these nine items loaded on the same scale in the SRI, they were practiced separately during the intervention because they took place at a different stage of the inquiry cycle.

The second scale, *experimenting* (7 items,  $\alpha_{\text{pre-test}} = .47$ ,  $\alpha_{\text{post-test}} = .81$ ), required children to select the best experiment based on the cover story. Each item presented children with three experimental designs with either two variables (2 items) or three variables (5 items). For each experiment only one experimental setup allowed for valid causal conclusions. The other experiments were either confounded, did not change any variables, or were controlled but did not manipulate the target variable.

Items on the third scale, *drawing conclusions* (8 items,  $\alpha_{\text{pre-test}} = .64$ ,  $\alpha_{\text{post-test}} = .77$ ), contained two premises and a question about those premises children could answer with 'yes', 'no' or 'maybe'. These syllogisms were embedded in the overarching cover story. For example, one of the syllogisms in the sports storyline was: 'All children who will go rowing, are wearing shorts. Anna will go rowing. Is she wearing shorts?'

#### *Reading comprehension and mathematic ability*

Most schools in the Netherlands participate in the student monitoring program of the National Institute for Educational Testing and Assessment (Cito). This program includes standardized assessments of children's cognitive abilities, which are administered twice a year. The tests of reading comprehension and mathematical skilfulness were used in the present study.

The reading comprehension test provided children with different types of texts, such as short stories, newspaper articles, advertisements and instructional manuals (Weekers et al., 2011). 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 meant or choose an appropriate continuation of a story. The mathematics test had children solve 96 multiple-choice and open-ended problems that were presented either with or without context (Hop et al., 2017). Contextualized problems consisted of a short text in which the problem was outlined and a supporting picture. In problems without context, children would only be presented with the numerical operations.

The monitoring program provides raw scores as well as a proficiency score (I-V, with I being the highest level and V the lowest). The latter can be used to meaningfully compare scores across different versions of the monitoring program. Because all

participating schools used the same student monitoring program, but not all schools used the same version, these proficiency scores were used as predictor variables. As such, the association between children's scientific reasoning and their proficiency in reading comprehension and mathematical skilfulness could be assessed without burdening children with more tests. In order to improve interpretability, proficiency scores were recoded so that 1 represented the lowest proficiency and 5 the highest proficiency.

### *Worksheet scoring*

The worksheets served a dual purpose in this study. In addition to being a supportive device, they were used as a process measure of children's learning. To this end, the worksheets of all five lessons were made as comparable as possible, differing only with regard to subject content (i.e., names of variables and images directly related to the subject-specific content). The questions and scaffolds were identical throughout the lesson series.

Worksheets were coded for each component skill (hypothesizing, experimenting, interpreting data, drawing conclusions). For each skill a maximum of 3 points was awarded, resulting in a maximum of 12 points per worksheet (see Table 1 for the coding scheme). *Hypotheses* were classified according to their level of specificity using the criteria proposed by Lazonder et al. (2010). Given the young age group in the current study, the definition of a fully specified hypothesis was slightly altered: it included the variables involved and a prediction of the direction of effect. *Experimenting* was assessed from children's use of the control-of-variables strategy (CVS; Chen & Klahr, 1999). It is important to note that this was not an all-or-nothing evaluation: even if the CVS was not applied, some points could still be awarded depending on the severity of the misconceptions (Peteranderl, 2019). At the very least, children had to understand the need for contrast, so a confounded experiment still received one point, whereas an experiment in which no variables were changed received zero points. The worksheet assignment for *interpreting data* consisted of two parts. The first part was a yes/no question that asked children whether they had observed a difference in outcomes between the two values of the focal variable. If the inference matched their data, one point was awarded. This inference should ideally be made based on multiple iterations of the same experiment. However, data gathered by children can be complex and messy (Kanari & Millar, 2004) and if this was the case, the single comparison was evaluated as a check. In the second part, children were asked to justify their inference. Two more points were awarded if children stated that they used the data to make this inference (a verbal statement of (non)covariation; Moritz, in Ben-Zvi & Garfield,

2004) and/or explained what caused the result they found. *Conclusions* were, like hypotheses, evaluated in terms of their specificity (Lazonder et al., 2010). In addition to the criteria described above, the effect children mentioned in their conclusion had to match the data they gathered.

A set of 86 randomly selected worksheets was coded by a second independent rater; the intraclass correlation (ICC) was calculated as a measure of interrater reliability. The ICC was high for all component skills: hypothesizing (.91,  $p < .001$ ), experimenting (.82,  $p < .001$ ), interpreting data (.94,  $p < .001$ ), and drawing conclusions (.89,  $p < .001$ ). Differences in interrater agreement were resolved through discussion. If children were present during all lessons, nine worksheets would be available. In practice, some children missed one lesson and some worksheets got lost in the classroom. As a result, between six and nine worksheets were available per child.

### *Procedure*

The study was carried out over a period of six weeks according to the setup outlined in Figure 1. During the first week, all children made the pre-test in a whole-class test setting. In weeks 2-6, children participated in five one-hour lessons taught by the principal investigator. Due to time constraints, the final lesson included the post-test and, hence, contained only one small-group inquiry. As the study did not aim to compare different instructional treatments, all children received the exact same lessons.

**Table 1**  
*Coding scheme*

Skill	Evaluation criteria	Example
Hypothesizing	An <i>effect</i> was described The <i>direction</i> of the effect was described The <i>variables involved</i> were described	'I think it makes a difference' (1 point: effect described) 'I think the surface matters for the number of bounces (2 points: effect and variables described; no direction) 'I think there will be more bounces on a hard surface' (3 points)
Experimenting	<i>Comparison</i> is possible: at least one variable has been changed <i>Fair comparison</i> is possible: only one variable has been changed Experiment <i>aligns with the research question</i> : focal variable has been changed	Confounded experiment (1 point: comparison possible) Controlled experiment on non-focal variable (2 points) Controlled experiment on focal variable (3 points)
Interpreting data	Based on the gathered data, a correct <i>inference</i> was made  The <i>explanation of the inference</i> refers to the data or outcome variable The data on which the inference was based are <i>described</i> or the outcome is <i>explained</i>	Part 1: Do you see a difference in the table? <b>yes/no</b> 1 point if answer aligns with data; 0 if not Part 2: How do you know? 'the number of bounces is different' (1 point: refers to outcome variable) 'on a hard surface the ball makes 5 more bounces than on a soft surface' (2 points: describes data and refers to variable)
Drawing conclusions	The effect that was found was described The <i>direction</i> of the effect was described All <i>variables involved</i> were described	'It makes a difference' (1 point; only if this was really found) 'The surface matters for the number of bounces' (2 points) 'The ball made more bounces on a hard surface' (3 points)

## Results

Standardized progress monitoring data of reading comprehension and mathematical skilfulness were obtained from 138 children (see Table 2); their pre- and post-test scores on the SRI are shown in Table 3 and Figure 2. These data show that, overall, children improved in scientific reasoning, but improvement rates differed among component skills. In order to explore these differences in scores and establish their relations with reading comprehension and mathematical skilfulness, a repeated measures multivariate analysis of covariance (MANCOVA) was carried out with time and component skill as within-subject variables, and reading comprehension level and mathematics level as between-subject covariates.

**Table 2**

*Descriptive statistics on reading comprehension and mathematical skilfulness*

	Level				
	I	II	III	IV	V
Reading comprehension	30.4%	25.4%	19.6%	15.2%	9.4%
Mathematical skilfulness	26.1%	17.4%	25.4%	18.8%	12.3%

**Table 3**

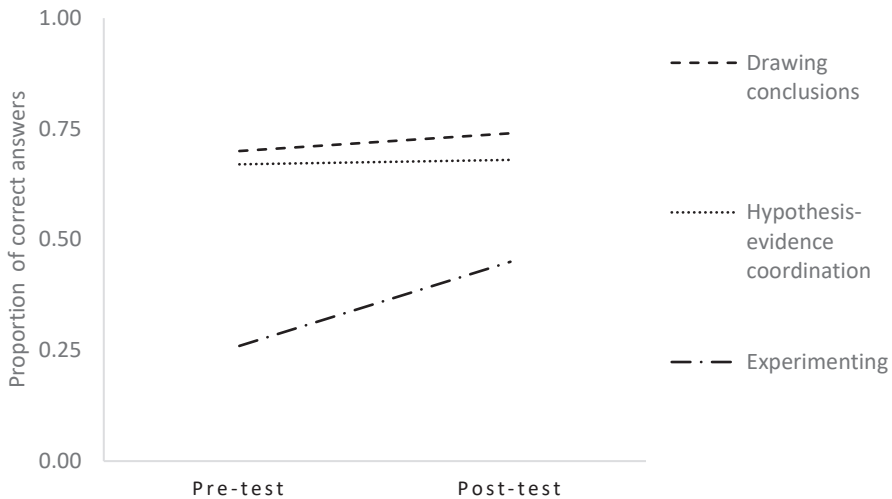
*Pre- and post-test scores on the Scientific Reasoning Inventory*

	Pre-test		Post-test		Gain	
	M	SD	M	SD	M	SD
Hypothesis-evidence coordination	0.67	0.23	0.68	0.27	0.01	0.21
Experimenting	0.26	0.22	0.45	0.35	0.20	0.38
Drawing conclusions	0.70	0.22	0.74	0.24	0.04	0.25
Overall	0.56	0.15	0.63	0.22	0.07	0.17

*Note.* Scores are reported as proportion of correct answers.

Figure 2

*Pre- and post-test scores per component scientific reasoning skill.*



### *Development and prediction*

Multivariate test results showed a main effect of time, Wilk's  $\lambda = .80$ ,  $F(1, 134) = 33.39$ ,  $p < .001$  and skill, Wilk's  $\lambda = .88$ ,  $F(1, 134) = 8.84$ ,  $p < .001$ . In addition to these main effects, an interaction was found between time and skill, Wilk's  $\lambda = .80$ ,  $F(2, 133) = 17.01$ ,  $p < .001$ , indicating asynchronous development of the component skills over time. Lastly, three-way interactions were found between time, skill and reading comprehension, Wilk's  $\lambda = .94$ ,  $F(2, 133) = 4.04$ ,  $p = .020$ , and time, skill and mathematical skilfulness, Wilk's  $\lambda = .91$ ,  $F(2, 133) = 6.84$ ,  $p = .001$ , indicating that both reading comprehension and mathematical skilfulness explain variation in development of the component skills throughout the lesson series.

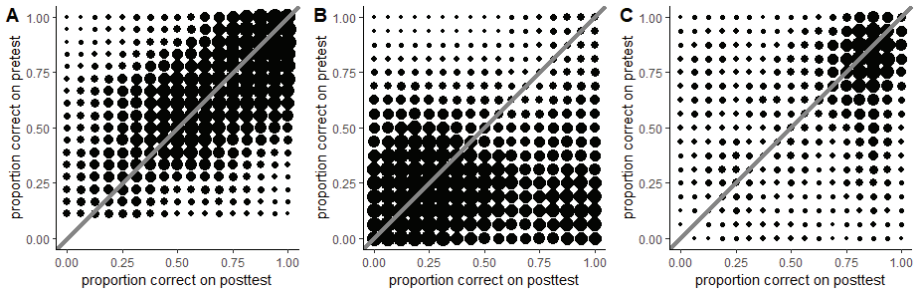
Both the data in Table 3 and the significant time  $\times$  skill interaction suggest that there may be subgroups of children who learned more than others. To examine this possibility, children's change in scores from pre- to post-test were visualized in density plots for each component skill (Figure 3). In these plots, the diagonal line stands for 'no development'; the area above the diagonal represents a decline in score, and the area below the diagonal indicates progress. For hypothesis-evidence coordination and drawing conclusions, most dots accumulate around the diagonal, meaning that children generally made little progress in these skills. A similar pattern was found for experimenting, except that there was an additional group of dots in the lower right corner. Thus, although the



majority of children hardly progressed in experimenting, a small group did. It is noteworthy that the two areas are horizontally aligned. This means that some children who scored very low on the pre-test still learned to experiment very well.

**Figure 3**

*Density plots for hypothesis-evidence coordination (A), experimenting (B) and drawing conclusions (C).*



*Note.* For all component skills, most scores cluster around the diagonal, indicating limited growth. For experimenting, a second cluster can be seen in the lower right corner, indicating a large improvement for a small group of children.

In order to further explore the three-way interactions, parameter estimates were obtained for pre- and post-test scores as well as for the gain scores (Table 4). These showed that for hypothesis-evidence coordination, both reading comprehension and mathematical skilfulness related to pre- and post-test scores. The predictors did not relate to gain scores on this skill, likely because there was very little progress. For experimenting, pre-test scores were not related to reading comprehension or mathematics, while post-test and gain scores were. Drawing conclusions was not related to children's reading comprehension or mathematical skilfulness at all.

**Table 4***Parameter estimates for interaction effects*

	Pre-test		Post-test		Gain scores	
	$\beta$	P	$\beta$	P	$\beta$	P
Hypothesis-evidence coordination						
Reading	.13	<.001	.19	<.001	.06	.179
Math	.06	.001	.09	.003	.00	.976
Experimenting						
Reading	.02	.591	.28	<.001	.25	<.001
Math	.04	.231	.24	<.001	.20	<.001
Drawing conclusions						
Reading	.02	.603	.06	.165	.04	.423
Math	.03	.324	.04	.235	.01	.808

*Note.* Previous analyses showed that gains in hypothesis-evidence coordination and drawing conclusions were not significant.

### *Key learning moments*

The third research question addressed children's learning process by identifying possible key learning moments during the lesson series. The worksheets children filled out during the lessons provided insight in this. A partial correlation between overall post-test scores (controlled for pre-test scores) and average worksheet scores was found, Spearman's  $\rho = .41$ ,  $p < .001$ , warranting further inspection of the process data summarized in Table 5. The partial correlation coefficients in this table show that the association between post-test and worksheet was consistent for some, but not all component skills. Specifically, hypothesizing and drawing conclusions (worksheets) were not related with any of the component skills measured by the Scientific Reasoning Inventory (SRI). Experimenting (worksheets) on the other hand did correlate with experimenting (SRI) as well as with hypothesis-evidence coordination (SRI). Interpreting data (worksheets) was associated with drawing conclusions (SRI).

**Table 5***Average worksheet scores and partial Spearman's rank correlations with post-test scores*

	Worksheets		Scientific Reasoning Inventory <sup>1</sup>		
	M	SD	H-E coordination $\rho$	Experimenting $\rho$	Drawing conclusions $\rho$
Hypothesizing	1.40	0.57	.12	.15	.14
Experimenting	1.96	0.63	.31**	.49**	.12
Interpreting data	1.59	0.55	.12	.10	.19*
Drawing conclusions	1.21	0.64	.08	.09	.03

*Note.* Worksheet scores ranged from 0 to 3 points.

<sup>1</sup>Post-test scores, controlled for pre-test scores.

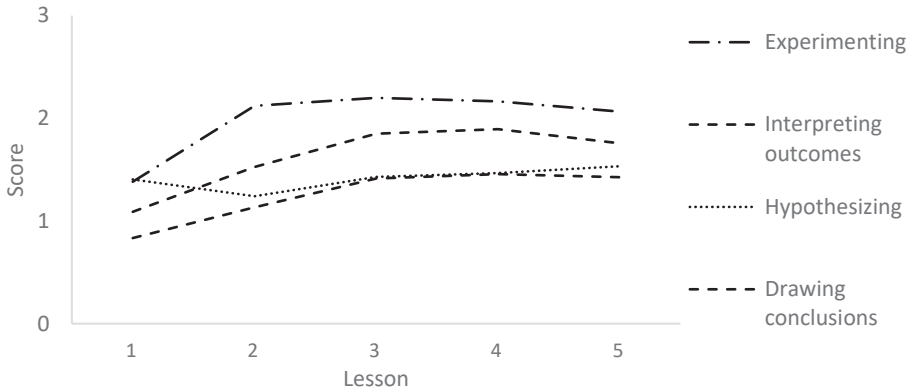
\* $p \leq .05$  \*\* $p \leq .001$

In addition to correlations between children's in-class performance and their achievements on the SRI, children's progress throughout the lessons was examined. First, visual inspection of the line graphs in Figure 4 helped determine whether progress was actually made, and if so, at which moment(s) during the lesson series this growth was most pronounced. For hypothesizing, the slope appears more or less level, indicating no or very moderate improvement. Progress in the other three component skills appears to be made between the first and third lesson, after which it levels off. The first, third and fifth lesson were therefore used as anchor points in children's developmental trajectories. Repeated measures ANOVAs were performed to compare the scores on each component skill at these three timepoints. As expected based on the line graph, no main effect was found for hypothesizing, Wilk's  $\lambda = 0.98$ ,  $F(2, 115) = 0.93$ ,  $p = .400$ , while significant within-subject differences were found for experimenting, Wilk's  $\lambda = 0.57$ ,  $F(2, 115) = 44.29$ ,  $p < .001$ , interpreting data, Wilk's  $\lambda = 0.66$ ,  $F(2, 115) = 29.71$ ,  $p < .001$ , and drawing conclusions, Wilk's  $\lambda = 0.77$ ,  $F(2, 115) = 17.15$ ,  $p < .001$ . Pairwise comparisons across the three timepoints were made to pinpoint when learning took place. The results in Table 6 show

that for experimenting, interpreting data and drawing conclusions, significant progress was made between lessons 1 and 3, but not between lessons 3 and 5.

**Figure 4**

*Worksheet scores per component scientific reasoning skill in each lesson*



To assess whether improvement of the worksheet scores could be explained by children's reading comprehension and mathematical skilfulness, a 3 (lessons)  $\times$  4 (skills) MANCOVA was performed, with reading comprehension and mathematical skilfulness as covariates. Multivariate test results showed no significant three-way interaction between lesson, skill and reading comprehension, Wilk's  $\lambda = .91$ ,  $F(6, 109) = 1.83$ ,  $p = .099$ . Between lesson, skill and mathematical skilfulness a three-way interaction was found, Wilk's  $\lambda = .88$ ,  $F(6, 109) = 2.46$ ,  $p = .029$ . However, further analysis of each component skill did not yield significant interactions between time and mathematical skilfulness. Thus, although mathematical skilfulness appears to predict progress in some component skills of scientific reasoning, this effect is not large enough to detect with more specific analyses.

**Table 6***Key learning moments in children's scientific reasoning skills inferred from their worksheet scores*

	Lesson	M	SD	Change <sup>1</sup>	<i>p</i> <sub>change</sub>
Hypothesizing	1	1.35	0.90		
	3	1.45	0.77	0.10	.574
	5	1.49	1.10	0.04	.978
Experimenting	1	1.37	0.80		
	3	2.19	0.88	0.82	<.001
	5	2.07	1.03	-0.12	.387
Interpreting data	1	1.09	0.97		
	3	1.85	0.74	0.76	<.001
	5	1.78	0.83	-0.07	.734
Drawing conclusions	1	0.84	1.00		
	3	1.44	0.84	0.60	<.001
	5	1.45	1.13	0.01	1.000

*Note.* Worksheet scores ranged from 0 to 4 points.<sup>1</sup>Compared to previous

## Discussion

The main purpose of this study was to investigate how children's scientific reasoning develops during an inquiry-based science lesson series, and which cognitive characteristics predict progress of its component skills. Process data gathered during the lessons was analysed to identify key moments during the lesson series when this progress was most pronounced. The findings, in short, point to a differential instructional effectiveness which should be considered in designing future adaptive learning arrangements.

Considerable diversity was observed in children's proficiency in and learning of scientific reasoning. Although there were significant overall gains on the SRI, this improvement did not apply equally to all component skills. Specifically, children advanced their experimenting skills, but not their ability to coordinate hypotheses with evidence and draw conclusions. Overall gains were explained by children's reading comprehension and mathematical skilfulness, as was their progress on experimenting skills and post-test performance on the hypothesis-evidence coordination items. However, both predictors

explained neither progress nor proficiency in drawing conclusions. Finally, children's worksheets evidenced progress over the lessons on experimenting, interpreting data and drawing conclusions, but not on hypothesizing. Most progress was made during the first half of the lesson series. These main outcomes of the study are discussed further below.

### *Predicting progress in scientific reasoning*

The first two research questions focused on children's progress on the component skills of scientific reasoning, with reading comprehension and mathematical skilfulness as predictors. Very little to no progress was expected to occur for hypothesis-evidence coordination and drawing conclusions, which indeed turned out to be the case. Although these component skills are often deemed more difficult than experimenting, pre-test scores were rather high in the current study. Still, the complete absence of progress is somewhat remarkable and suggests that both skills are not only hard to perform but also difficult to improve. No interactions were found between the predictor variables and progress on either hypothesis-evidence coordination or drawing conclusions, but children's *proficiency* in hypothesis-evidence coordination interacted with reading comprehension and mathematical skilfulness on both pre- and post-test. This result seems understandable because the scale combined items that tap into the ability to identify appropriate research questions and interpret data, which are component skills that were expected to interact with both predictor variables.

Our hypotheses regarding experimenting were twofold: we either expected to find high pre-test scores and little progress, or low pre-test scores and substantial growth. Evidence was found for the latter hypothesis, although on average post-test scores for experimenting were lower than those for hypothesis-evidence coordination and drawing conclusions. This is noteworthy because experimenting is often regarded as one of the least difficult scientific reasoning skills to learn. The large standard deviations on the post-test imply that some children had improved more than others, which was confirmed by the interactions of both the post-test scores and progress with reading comprehension and mathematical skilfulness. In combination with the density plots shown in Figure 3, it therefore seems plausible that some, but not all children developed adequate experimentation strategies through structured, repeated practice. Informal observations during the lessons further indicated that some children realized that the research question could not be answered based on a confounded experiment. As the worksheets did not explicitly link experimental design to drawing conclusions, conceptualizing this connection required unsupported data interpretation. The significant impact of children's language and

math skills suggests that only children with relatively high intellectual abilities were able to make this inference.

### *Progress on scientific reasoning during the lessons*

Children's entries on the worksheets were analysed to unveil key moments in the learning process where marked progress in scientific reasoning was made. Notable improvements in experimenting, interpreting data and drawing conclusions occurred between lesson 1 and lesson 3, whereas no progress in hypothesizing was made over the five lessons. The latter result may be due to the fact that, unlike the other component skills, children's hypotheses were rarely addressed during the whole-class discussions. Another possibility is that hypothesizing is easier if one has a theoretical basis in the topic of inquiry (Koslowski et al., 2008), which the children in our study had not or to an insufficient degree. The lack of growth in hypothesizing skills might be attributable to a combination of these factors.

Progress on the other component skills occurred between lesson 1 and lesson 3. Interestingly, children's performance stabilized after the third lesson, despite the absence of a ceiling effect. This raises the question why progress levelled off before mastery was reached. A possible answer lies in the design principles underlying the lesson series. Both the lessons and the worksheets were highly structured (guided inquiry, Bell et al., 2005) but contained few explicit directions and explanations. The available implicit guidance enabled children to improve their scientific reasoning to some extent, meaning that additional growth may require additional guidance, extended practice, or both.

Using a combination of instructional support measures might help sustain children's progress beyond the third lesson. Previous research comparing open and guided inquiry to direct instruction (e.g., Alfieri et al., 2011; Vorholzer et al., 2018; Wagensveld et al., 2014) indicated that open inquiry was often ineffective, whereas guided inquiry or direct instruction yielded higher learning outcomes. Using data from the 2015 Trends in International Mathematics and Science Study (TIMSS), Teig et al. (2018) also concluded that inquiry can be an effective approach, but only when combined with other, more explicit forms of guidance. Along these lines, more specific directions by the teacher or through the worksheets could have resulted in significant progress on hypothesizing and to full mastery of the skills experimenting, interpreting data and drawing conclusions. What these instructions should entail and how they are best combined with the scaffolding offered by the worksheets are interesting questions for future research.

### *Towards adaptive science instruction*

The present findings suggest that some children need little support to improve their scientific reasoning skills, whereas others seem to require more or more specific guidance. The worksheet data show that children improved in all scientific reasoning skills except hypothesizing; this progress was often modest and occurred in the first half of the lesson series. In order to help children further improve their scientific reasoning, we have three suggestions. First, guidance could be increased on component skills that are particularly difficult to learn, such as hypothesizing. Second, considering the relations found between scientific reasoning and children's reading comprehension and mathematical skilfulness, progress monitoring data of these school subjects can help teachers to adapt their science lessons in advance, for instance by planning more or more explicit guidance for children with lower levels of reading comprehension. Third, monitoring in-class performance can inform teachers when children need additional support. Using this information to make instant adjustments above and beyond the pre-planned adaptations could be a crucial next step in the improvement of primary science education.

### *Strengths and limitations*

On the positive side, this study examined multiple component skills of scientific reasoning under rather uniform conditions. As argued by Koerber and Osterhaus (2019), cross-study comparisons of proficiency and developmental growth in distinct scientific reasoning skills are likely confounded by differences in learner characteristics and task settings. Their plea for more comprehensive investigations of scientific reasoning was met here, and allows for more valid conclusions on the relative ease or difficulty with which individual scientific reasoning skills are acquired during primary science lessons.

Another asset of this study is the use of two complementary data sources: the SRI and the worksheets. The origin of an instrument (existing or made for the study) can affect the outcomes (Schwichow et al., 2016). So, in order to shed more light on children's science learning in regular classrooms, but without compromising experimental validity, we combined scores on the experimentally valid SRI, administered in a test setting, with more ecologically valid data from the worksheets children filled out during the lessons.

Although this approach yielded valuable insights in the development of some scientific reasoning skills, an unforeseen discrepancy between these two data sources arose. Although the SRI and the worksheets both targeted the same component skills (hypothesizing, experimenting, interpreting data and drawing conclusions), factor analysis



of the SRI-items in both the validation study and the current study required us to combine two of these skills in a single scale. This complicated the comparison of children's scores on the worksheets and the SRI.

This measurement inconsistency is inconvenient because different proficiency patterns emerged for the two test modalities, which are now difficult to explain. While the worksheets outcomes followed the hypothesized proficiency pattern, with highest scores for experimenting and lowest for hypothesizing, the SRI scores for hypothesis-evidence coordination and drawing conclusions were high and scores on experimenting were low. Strong claims about what accounted for these discrepancies cannot be made, but there are several possible explanations.

First, differences in test item format may have played a role. Previous studies showed that the type of data greatly influences the ease of interpretation (Kanari & Millar, 2004; Masnick & Morris, 2008). The hypothesis-evidence coordination items on the SRI featured unambiguous, dichotomous outcomes that were relatively easy to interpret, whereas the data children gathered during the lessons were continuous and messier. Although both called upon children's ability to interpret data, requirements on the SRI were relatively limited. The high scores on hypothesis-evidence coordination and drawing conclusions suggest that the SRI taps children's basic proficiency in these component skills, whereas the worksheet provides a more authentic assessment. Secondly, surface characteristics may have limited comparability as well (Stiller et al., 2016). For example, longer questions and data tables (as used on the SRI scale hypothesis-evidence coordination) can decrease difficulty, whereas the longer response options (which were used on the worksheets for hypothesizing and drawing conclusions) may increase difficulty.

Finally, reliability of the experimenting scale of the SRI pre-test was low. This was probably caused by the fact that children did not have much experience with experimenting, and because the item format was relatively difficult for them. As a result, the range of scores on the pre-test was small and this limited variability may have affected Cronbach's  $\alpha$ .

### *Implications and directions for further research*

Although the present study provides initial directions for adaptive science education, future research is needed to assess the effectiveness of these adaptations. This and other studies show that scientific reasoning can be taught to children of all cognitive levels (Zohar & Dori, 2003), yet less is known about how the needs of individual children in a class are best met. So, although our findings indicate that teachers can base instructional adaptations on

children's proficiency in reading and math, research should investigate additional ways to adapt instruction in scientific reasoning.

Although the relationship between reading comprehension and scientific reasoning is well-established and caused some to conclude that scientific reasoning is linguistic in nature (Van de Sande et al., 2019), the relation between mathematical skilfulness and scientific reasoning has only recently been shown (Koerber & Osterhaus, 2019). The current study confirms that such a relationship exists. Acknowledging the impact of mathematical skilfulness is important for the effective teaching of scientific reasoning, which can be more thoughtfully designed bearing this information in mind.

### *Conclusion*

Fifth graders generally improved in scientific reasoning during a five-week inquiry-based lesson series. They made progress in all component skills except hypothesizing, mainly during the first half of the lesson series, and consolidated their increased experimentation skills on the post-test. Reading comprehension and mathematical skilfulness accounted for part of the variance in children's progress and proficiency scores, and offer fertile grounds for adaptivity. However, more research is needed to fully grasp the individual variation in children's science learning and explore ways to accommodate these differences. The outcomes of these studies contribute to the design of effective primary science education for all.

## References

- Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*(1), 1-18. <https://doi.org/10.1037/a0021017>
- Bell, R. L., Smetana, L., & Binns, I. (2005). Simplifying inquiry instruction. *The Science Teacher, 72*(7), 30-33.
- Ben-Zvi, D., & Garfield, J. (2004). The challenge of developing statistical literacy, reasoning and thinking. Kluwer Academic Publishing.
- 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.
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098-1120. <https://doi.org/10.1111/1467-8624.00081>
- Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: spontaneous experiments in preschoolers' exploratory play. *Cognition, 120*(3), 341-349. <https://doi.org/10.1016/j.cognition.2011.03.003>
- Greven, J., & Letschert, J. (2006). *Kerndoelen primair onderwijs*. The Hague, the Netherlands
- Hop, M., Janssen, J., & Engelen, R. (2017). *Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 7*. [Scientific justification arithmetics and mathematics 3.0 for grade 5]. Cito.
- Inspectie van het Onderwijs. (2017). *Pijl.Natuur en Techniek* [Level of Science Education and Performance]. Inspectie van het Onderwijs.
- Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching, 41*(7), 748-769. <https://doi.org/10.1002/tea.20020>
- Kind, P. M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *Journal of Research in Science Teaching, 50*(5), 530-560. <https://doi.org/10.1002/tea.21086>
- Kind, P. M., & Osborne, J. (2017). Styles of scientific reasoning: A cultural rationale for science education? *Science Education, 101*(1), 8-31. <https://doi.org/10.1002/scs.21251>
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1-48. [https://doi.org/10.1207/s15516709cog1201\\_1](https://doi.org/10.1207/s15516709cog1201_1)
- Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development, 86*(1), 327-336. <https://doi.org/10.1111/cdev.12298>
- 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.org/10.1080/15248372.2019.1620232>
- Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss Journal of Psychology, 64*(3), 141-152. <https://doi.org/10.1024/1421-0185.64.3.141>
- Köksal-Tuncer, Ö., & Sodian, B. (2018). The development of scientific reasoning: Hypothesis testing and argumentation from evidence in young children. *Cognitive Development, 48*, 135-145. <https://doi.org/10.1016/j.cogdev.2018.06.011>
- Koslowski, B., Marasia, J., Chelenza, M., & Dublin, R. (2008). Information becomes evidence when an explanation can incorporate it into a causal framework. *Cognitive Development, 23*(4), 472-487. <https://doi.org/10.1016/j.cogdev.2008.09.007>

- Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *Blackwell Handbook of Childhood Cognitive Development* (pp. 371-393). Blackwell Publishers Ltd.
- Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, 16(11), 866-870. <https://doi.org/10.1111/j.1467-9280.2005.01628.x>
- Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cognition and Instruction*, 26(4), 512-559. <https://doi.org/10.1080/07370000802391745>
- Lazonder, A. W., Hagemans, M. G., & De Jong, T. (2010). Offering and discovering domain information in simulation-based inquiry learning. *Learning and Instruction*, 20(6), 511-520. <https://doi.org/10.1016/j.learninstruc.2009.08.001>
- 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>
- Lazonder, A. W., & Wiskerke-Drost, S. (2015). Advancing scientific reasoning in upper elementary classrooms: Direct instruction versus task structuring. *Journal of Science Education and Technology*, 24(1), 69-77. <https://doi.org/10.1007/s10956-014-9522-8>
- Lorch, R. F., Lorch, E. P., Freer, B., Calderhead, W. J., Dunlap, E., Reeder, E. C., Van Neste, J., & Chen, H.-T. (2017). Very long-term retention of the control of variables strategy following a brief intervention. *Contemporary Educational Psychology*, 51, 391-403. <https://doi.org/10.1016/j.cedpsych.2017.09.005>
- Lorch, R. F., Lorch, E. P., Freer, B. D., Dunlap, E. E., Hodell, E. C., & Calderhead, W. J. (2014). Using valid and invalid experimental designs to teach the control of variables strategy in higher and lower achieving classrooms. *Journal of Educational Psychology*, 106(1), 18-35. <https://doi.org/10.1037/a0034375>
- Masnick, A., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development*, 79(4), 1032-1048. <https://doi.org/10.1111/j.1467-8624.2008.01174.x>
- 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>
- Osterhaus, C., Koerber, S., & Sodian, B. (2017). Scientific thinking in elementary school: Children's social cognition and their epistemological understanding promote experimentation skills. *Developmental Psychology*, 53(3), 450-462. <https://doi.org/10.1037/dev0000260>
- 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. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, 14, 47-61. <https://doi.org/10.1016/j.edurev.2015.02.003>
- Peteranderl, S. (2019). Experimentation skills of primary school children
- Piekny, J., Grube, D., & Maehler, C. (2014). The development of experimentation and evidence evaluation skills at preschool age. *International Journal of Science Education*, 36(2), 334-354. <https://doi.org/10.1080/09500693.2013.776192>
- Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology*, 31(2), 153-179. <https://doi.org/10.1111/j.2044-835X.2012.02082.x>
- Schalk, L., Edelsbrunner, P. A., Deiglmayr, A., Schumacher, R., & Stern, E. (2019). Improved application of the control-of-variables strategy as a collateral benefit of inquiry-based physics education in elementary school. *Learning and Instruction*, 59, 34-45. <https://doi.org/10.1016/j.learninstruc.2018.09.006>

- Schlatter, E., Lazonder, A. W., Molenaar, I., & Janssen, N. (2021). Individual differences in children's scientific reasoning. *Education Sciences*, 11(9), Article 471. <https://doi.org/10.3390/educsci11090471>
- Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, 39, 37-63. <https://doi.org/10.1016/j.dr.2015.12.001>
- Siler, S. A., Klahr, D., Magaro, C., Willows, K., & Mowery, D. (2010, June). *Predictors of transfer of experimental design skills in elementary and middle school children*. 10th International Conference on Intelligent Tutoring Systems, Pittsburgh, PA.
- 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/10.1111/j.1467-8624.1991.tb01567.x>
- Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., & Upmeyer zu Belzen, A. (2016). Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education*, 41(5), 721-732. <https://doi.org/10.1080/02602938.2016.1164830>
- Tajudin, N. M., & Chinnappan, M. (2015). *Exploring relationship between scientific reasoning skills and mathematics problem solving*. 38th annual conference of the Mathematics Education Research Group of Australasia: Mathematics education in the margins, Sunshine Coast, Australia.
- Teig, N., Scherer, R., & Nilsen, T. (2018). More isn't always better: The curvilinear relationship between inquiry-based teaching and student achievement in science. *Learning and Instruction*, 56, 20-29. <https://doi.org/10.1016/j.learninstruc.2018.02.006>
- Van de Sande, E., Kleemans, M., Verhoeven, L., & Segers, E. (2019). The linguistic nature of children's scientific reasoning. *Learning and Instruction*, 62(1), 20-26. <https://doi.org/10.1016/j.learninstruc.2019.02.002>
- Van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instructional Science*, 43(3), 381-400. <https://doi.org/10.1007/s11251-015-9344-y>
- Van der Graaf, J., Segers, E., & Verhoeven, L. (2018). Individual differences in the development of scientific thinking in kindergarten. *Learning and Instruction*, 56, 1-9. <https://doi.org/10.1016/j.learninstruc.2018.03.005>
- Van Graft, M., Klein Tank, M., Beker, T., & Van der Laan, A. (2018). Wetenschap en technologie in het basis- en speciaal onderwijs: Richtinggevend leerplankader bij het leergebied Oriëntatie op jezelf en de wereld [Science and technology in primary and special education: directive educational framework for orientation on self and world]. SLO (nationaal expertisecentrum leerplanontwikkeling).
- Van Riesen, S., Gijlers, H., Anjewierden, A., & De Jong, T. (2018). Supporting learners' experiment design. *Educational Technology Research and Development*, 66(2), 475-491. <https://doi.org/10.1007/s11423-017-9568-4>
- Vorholzer, A., Von Aufschnaiter, C., & Boone, W. J. (2018). Fostering upper secondary students' ability to engage in practices of scientific investigation: A comparative analysis of an explicit and an implicit instructional approach. *Research in Science Education*, 50, 333-359. <https://doi.org/10.1007/s11165-018-9691-1>
- Wagensveld, B., Segers, E., Kleemans, T., & Verhoeven, L. (2014). Child predictors of learning to control variables via instruction or self-discovery. *Instructional Science*, 43(3), 365-379. <https://doi.org/10.1007/s11251-014-9334-5>

- Weekers, A., Groenen, I., Kleintjes, F., & Feenstra, H. (2011). *Wetenschappelijke verantwoording papieren toetsen begrijpend lezen voor groep 7 en 8* [Scientific justification paper-and-pencil tests reading comprehension grade 5 and 6]. Cito.
- Wong, V. (2019). 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. *International Journal of Science Education*, 41(13), 1805-1826.  
<https://doi.org/10.1080/09500693.2019.1641249>
- 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>
- Zohar, A., & Dori, Y. (2003). Higher order thinking skills and low-achieving students: Are they mutually exclusive? *Journal of the Learning Sciences*, 12(2), 145-181.  
[https://doi.org/10.1207/S15327809JLS1202\\_1](https://doi.org/10.1207/S15327809JLS1202_1)







# Chapter 4

**Learning scientific  
reasoning:**

**A latent transition analysis**

### Abstract

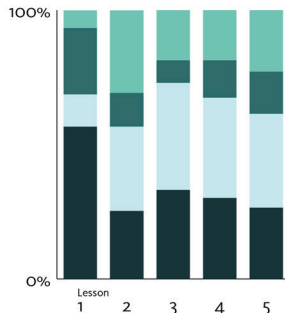
Primary education in many countries enables children to learn the scientific reasoning skills of hypothesizing, experimenting, interpreting data and drawing conclusions. Research has shown that these component skills develop at a different pace and with substantial variation among same-age children. How these differences play out in the short term is less well known. This study used Latent Transition Analysis of the worksheets filled out by 166 fifth graders during five science lessons to explore whether different proficiency profiles could be identified, and if so, how children transitioned between these profiles. The results distinguished four distinct profiles, which were labelled as high achievers, low achievers, experimenters and theorists. Children transitioned regularly among these profiles and two possible trajectories towards becoming a high achiever emerged: from low achievers via experimenters or via theorists. Awareness of these individual differences can help teachers differentiate their science lessons.

### Graphical abstract

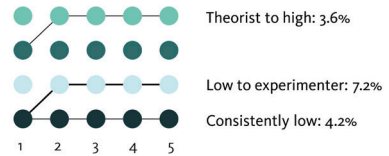
#### Four profiles were found:



#### Distribution of profiles changed over time



#### Transition paths taken by >2.5%:



102 transition paths were found in total

#### Conclusion

- Diversity in scientific reasoning is high
- Profiles can help teachers differentiate

\* Hypothesizing, Experimenting, Interpreting data & drawing Conclusions

This chapter is based on:

Schlatter, E., Molenaar, I., & Lazonder, A. W. (2021). Learning scientific reasoning: A latent transition analysis. *Learning and Individual Differences*, 92, Article 102043. <https://doi.org/10.1016/j.lindif.2021.102043>

## Introduction

Primary school science education enables children to learn scientific content knowledge as well as scientific reasoning skills. Scientific reasoning, which is the main focus of this chapter, includes the skills of hypothesizing, experimenting, and evaluating outcomes (Edelsbrunner & Dablander, 2018; Klahr & Dunbar, 1988). During a typical school science project, children have to perform these skills in consecutive order, starting with the generation of hypotheses to predict the outcomes of the upcoming investigation. To test their hypothesis, children set up and run one or more experiments in which only the variable of interest should be manipulated. When interpreting the gathered data, children have to interpret their raw measurements, for example by indicating which data column contains the highest values. Based on this inferred pattern in scores, children formulate a conclusion that indicates the truth value of their hypothesis in a clear and specific manner. Research suggests that these scientific reasoning skills neither emerge at the same age nor develop at the same pace. Experimenting is relatively easy to learn, and although a full understanding of this skill is generally reached in secondary education (Schwichow et al., 2020), pre-schoolers already show an emerging proficiency in designing systematic experiments (Koerber & Osterhaus, 2019; Piekny et al., 2014; Van der Graaf et al., 2018). Hypothesizing and drawing conclusions are more difficult for children and generally start developing toward the end of primary education (Piekny & Maehler, 2013). The difficulty of evaluating outcomes is greatly dependent on the data type, and develops throughout the primary school years (Kanari & Millar, 2004; Piekny & Maehler, 2013).

However, individual children tend to deviate from this general developmental trajectory (Schlatter et al., 2020 [Chapter 3]), causing some scholars to conclude that scientific reasoning develops by leaps and bounds (Lazonder et al., 2021). As children may be stagnant for one year and exhibit marked growth during the next, same-age children can differ substantially in scientific reasoning proficiency. In light of these individual differences, knowing what children at a certain age are generally capable of is informative but insufficient, in particular for teachers aspiring to differentiate their science lessons beyond the general proficiency level in their classrooms.

Extant developmental research gives these teachers few specific guidelines for differentiation. Most studies administered cross-sectional designs to infer how scientific reasoning develops over the years, often using different tasks and settings (Koerber & Osterhaus, 2019). Descriptive studies do allow for valid causal conclusions about how scientific reasoning develops in the short term as a result of some instructional intervention

(e.g., Edelsbrunner et al., 2018; Kuhn et al., 2015; Zohar & Dori, 2003) but are underrepresented in the research literature and generally disregard developmental differences across component skills and among children. The present study addressed this paucity in the research by examining how a sample of fifth graders improved in scientific reasoning during the course of a five week lesson series. Using Latent Transition Analysis, the study identified proficiency profiles for the four component skills, and examined the learning trajectories that emerged from the data.

*The case for examining proficiency profiles and developmental trajectories*

Previous research points to substantial differences in children's scientific reasoning (e.g., Koerber & Osterhaus, 2019; Mayer et al., 2014; Schlatter et al., 2020 [Chapter 3]; Zimmerman, 2007). These studies typically rely on variable-centred analyses to compare mean scores of groups defined by relatively stable characteristics such as age and cognitive capacities. In other words, they study the effect of one variable, such as age, on another, such as scientific reasoning (Howard & Hoffman, 2018). Although this deterministic approach is appropriate for analysing individual differences, it unlikely captures the full spectrum of variation in scientific reasoning. Useful complementary evidence can be provided by probabilistic, person-centred analysis techniques capable of uncovering latent proficiency profiles. Rather than relying on stable traits such as age and gender, person-centred analyses allow for the identification of subgroups based solely on the variables of interest (in the case of this chapter: children's scientific reasoning skills). As such, subgroups of children can be distinguished based on their similarity to one another (in this chapter: having a similar scientific reasoning proficiency profile) as well as their difference from children in other subgroups (Hickendorff et al., 2018). A person-centred approach appropriate for longitudinal data is Latent Transition Analysis (LTA). LTA comprises two steps. First, different proficiency profiles are identified, even if they cannot readily be observed. Then, it is determined if individuals remain in the same profile through time, or transition from one profile to another. The latter is particularly important when analysing learning (Reimann, 2009).

In the context of science education, LTA has mainly been applied to reveal mechanisms of conceptual change. For example, Schneider and Hardy (2013) examined children's conceptions about the topic of floating and sinking at three points in time: before an eight-lesson intervention, directly after this intervention and after one year. They found five profiles, some indicating overall understanding (many answers indicating scientific concepts and few indicating misconceptions or everyday concepts) or misunderstanding (many

answers indicating misconceptions and few indicating every day or scientific concepts), whereas the other profiles reflected an intermediate stage of understanding (e.g., some misconceptions, some everyday concepts and some scientific concepts). The study further showed that 63 percent of the children followed one of seven paths through these profiles. Most trajectories represented an upward trend (i.e., children progressed toward increasingly scientific profiles), although some children remained in the same profile at all three measurement points.

In another study of children's learning trajectories in early science instruction, Van der Graaf (2020) analysed children's strategy use in solving balance beam problems before and after a one-hour inquiry-based lesson. Previous research established that children use different types of strategies, ranging from reasoning about mass only to correctly taking both mass and distance into account (e.g., Boom & Ter Laak, 2007). These strategies become increasingly more sophisticated, and Van der Graaf (2020) demonstrated that this improvement is already apparent after one short lesson. Furthermore, advances in strategy use were more likely for children who received a lesson on the balance beam than for children in the control condition, who received a lesson on the working of gears. However, as only two measurement points were considered in this study, it is likely that the full complexity of learning trajectories was not captured.

Very few studies have applied person-centred analysis to capture the proficiency in and development of scientific reasoning. An exception is the study by Schwichow et al. (2020), who analysed profiles of experimenting skills in secondary school students. Experimenting was divided in four component skills: identifying controlled experiments, planning controlled experiments, interpreting outcomes from controlled experiments and interpreting outcomes from confounded experiments. Six proficiency profiles were found, some indicating overall proficiency or a lack thereof, and other profiles that reflected proficiency in some, but not all component skills of experimenting.

To conclude, person-centred analyses such as LTA offer unique possibilities to unravel individual differences in children's science learning, both in terms of proficiency profiles and learning trajectories. The method has recently gained momentum in science education research where it is predominantly employed to study the development of science content knowledge (Schneider & Hardy, 2013; Van der Graaf, 2020). The present study extends this pioneering work by using LTA to portray how scientific reasoning skills develop in upper primary education. Proficiency profiles and learning trajectories were established by analysing children's scientific reasoning on a lesson-by-lesson basis, which refines the coarse-grained approach of analysing pre- and post-test differences used in previous studies.

Another asset of the present study concerns the concurrent assessment and analysis of four component scientific reasoning skills. As scientific reasoning consists of multiple component skills that are often taught in coordination with one another, it is important to establish the developmental trajectories of each component skill under similar circumstances.

### *The present study*

The research reported here builds upon earlier work concerning learning outcomes in scientific reasoning (Schlatter et al., 2020 [Chapter 3]). In that study, variable-centred analyses on pre- and post-test data were used to examine whether children's scientific reasoning improved after five inquiry-based scientific reasoning lessons. Children did progress in some, but not all scientific reasoning skills, and this progress was predicted by standardized measures of reading comprehension and, to a lesser extent, mathematical skilfulness. Furthermore, although some children became rather proficient in scientific reasoning, others made no progress at all. Intriguingly, this pattern applied to some, but not all component scientific reasoning skills.

Prompted by this large variation in children's learning *outcomes*, we decided to do a fine-grained analysis of these children's learning *processes*. By using LTA, we aimed to identify latent proficiency profiles and learning trajectories based on similarity in scores on the worksheets children filled out during the five-week inquiry-based lesson series (rather than the pre- and post-test they took before and after the lesson series). The study sought to answer two research questions:

- (1) Which proficiency profiles can be identified from the process data of fifth graders engaged in a five-week inquiry-based lesson series that addresses the scientific reasoning skills of hypothesizing, experimenting, interpreting data and drawing conclusions?
- (2) How do children transition between these proficiency profiles during the course of the five lessons?

Because of the known asynchronous and nonlinear development of scientific reasoning in same-age children (e.g., Lazonder et al., 2021), different proficiency profiles were expected to be found. Some of these profiles were expected to be homogeneous, representing those children who are either good or bad at all component scientific reasoning skills. As some skills are known to be more difficult than others, it stands to reason that mixed profiles would emerge as well, with children being rather proficient in some component skills, in

particular the ones that mature early in life such as experimenting, but still less well-versed in the skills that are more difficult to develop.

Hypotheses regarding the second research question predicted that children would transition between proficiency profiles throughout the lesson series. Their transition paths were expected to show a general upward trend because children generally improve in scientific reasoning during a five-week lesson series (Schlatter et al., 2020 [Chapter 3]). However, scientific reasoning development is not linear and, thus, not all children will improve at the same time. Therefore, it is possible that some children stagnate in their progression or even experience a small setback.

## Method

### *Participants and procedure*

This study reports a detailed analysis of process data collected in the fall of 2018 as part of a comprehensive research project, approved by the institutional review board of the Behavioural Science Institute, Radboud University, under number 2018-074R1. Initial results pointed to considerable variation in children's development of scientific reasoning that could not be fleshed out by variable-centred analysis (Schlatter et al., 2020 [Chapter 3]). The present study aimed to scrutinize these individual differences by identifying developmental profiles from process data gathered during the lessons, and establish whether and how children transition between these profiles. The process data were collected in nine fifth-grade classes (in Dutch: 'groep 7') from seven schools in the central and northern part of the Netherlands. All children in these classrooms received five one-hour lessons as part of their regular science curriculum. Passive parental consent was acquired with the exception of one school, which preferred active parental permission.

The initial sample contained 168 children (52% boys) between the ages of 8 and 12, with the majority being age 10. We collected the worksheets these children filled out during the lessons. As the study spanned five weeks, it was not considered problematic if children missed one lesson. However, two children were excluded because they missed more than one lesson and as a result, complete data was available for 166 children.

### *Lessons*

Children engaged in five inquiry-based science lessons about the primary school physics topics displayed in Table 1. Previous studies showed that these topics are of similar

conceptual complexity to children (Chen & Klahr, 1999; Lazonder et al., 2021). All children received the exact same lessons that were taught by the principal investigator. The lessons encompassed a mixture of whole class discussions and small group inquiry. Specifically, the lessons started with a whole class introduction, followed by a 20-minute inquiry and a whole class discussion of the results. The second part of the lesson comprised another 20-minute inquiry and a whole class discussion of the results and the underlying physics principles. For practical reasons, the fifth lesson was shortened by skipping the second inquiry session. The current study analysed the learning processes during these small group inquiry sessions from children's entries on the worksheets.

In order to synchronize what children would investigate during the inquiry sessions, specific research questions were provided according to the following template: Does the *<input variable>* affect the *<outcome variable>*? For example, one worksheet for the pendulum swing experiment centred around the question 'Does the length of the rope affect the period of a swing?'. To answer this research question, children had to formulate a hypothesis, set up and run an experiment, interpret the outcomes, and draw a conclusion (see Figure 1). The hands on part of these investigations (i.e., the actual conduct of the experiment) was carried out in dyads for practical reasons. Children were paired by the principal investigator on an ad-hoc basis in each lesson, and served as each other's 'research assistant' during data collection. When children asked the principal investigator for help, only procedural support was given, for example by demonstrating how to handle the experimental equipment, reminding them to formulate a hypothesis before doing experiments, or explaining how to record their outcomes in the table. If children struggled content-wise, support was gradually increased from repeating the question, via supportive questions, to explicit instruction. In practice, very few children repeatedly asked for help.

### *Instruments*

#### *Worksheets*

Children's inquiry sessions were guided by worksheets that assisted them in formulating hypotheses, doing experiments, interpreting data, and drawing conclusions. The worksheets contained assignments and scaffolds that structured children's inquiry without explicitly instructing them what to do and why. Figure 1 gives an impression of the assignments and scaffolds; a sample worksheet is included in Appendix A. Although children examined a different topic each lesson, both the structural complexity of their inquiry (i.e., the number and type of variables; see Table 1) and the nature of the research questions, assignments



and scaffolds were identical. Children's worksheet entries could thus be meaningfully compared across lessons to identify proficiency profiles in scientific reasoning and learning trajectories through the lessons.

Children could complete nine worksheets in total: two for the first four lessons and one during the final lesson. Completed worksheets were coded for the scientific reasoning skills of hypothesizing, experimenting, interpreting data and drawing conclusions. A maximum of 3 points was awarded for each component skill, resulting in a maximum of 12 points per worksheet. Since the scientific reasoning skills were placed in an action sequence, the coding rules took children's performance of preceding skills into account so as to avoid violation of the local independence assumption (see Figure 1 for a summary of the coding scheme and Schlatter et al., 2020 [Chapter 3], for details). Interrater reliability of the worksheet coding was assessed by having an independent researcher score 86 randomly selected worksheets. The intraclass correlation (ICC) of all component skills was high (ICC > .82,  $p < .001$ ) and disagreements in scoring were resolved through discussion.

**Table 1**  
*Inquiry topics, materials and execution*







	Lesson 1: Pendulum swing	Lesson 2: Water absorption	Lesson 3: Inclined planes	Lesson 4: Bouncing balls	Lesson 5: Spring capacity
Research question	What affects the period of a pendulum?	What affects a fabric's water absorption?	What affects the distance a ball rolls?	What affects the number of bounces a ball makes?	What affects the spring capacity of a plank?
Materials	Stick; two lengths of rope; two washers to be used as weights; card to indicate amplitude; stopwatch	Fabric swatches (2 grams each); plastic cups; stopwatch; scale	Ramp with adjustable surface and starting gate; two same-sized balls (marble and ball bearing)	Ruler; piece of felt; two same-sized balls (polystyrene and ping pong)	Two same-sized blocks of wood; two wooden slats (narrow and wide); bean bag; measuring tape
Variables					
Independent:	Rope length	Fabric	Starting position	Type of ball	Length of plank
3 variables w.	Long, short	Lycra, towel	High, low	polystyrene, ping pong	Long, short
2 values	Bob weight	Time submerged	Weight of ball	Height of drop	Width of plank
	Heavy, light	Long, short	Heavy, light	High, low	Wide, narrow
	Amplitude	Folding status	Surface	Surface	Depth of push
	Far, near	Rolled, unfolded	Rough, smooth	Hard, soft	Deep, shallow
Dependent	Period (5 swings)	Weight of soaked swatch	Distance ball rolled	Number of bounces	Highest point of bean bag
Execution	Attach washer to rope Hang rope from stick Pull aside using amplitude indicator card for reference Simultaneously release and start stopwatch Count five swings; stop stopwatch	Fill one cup with water, place an empty cup on the scale Select dry swatch Set stopwatch Submerge swatch for set period of time Remove swatch from water and put on scale	Place surface insert Place starting gate Place ball Remove starting gate	Place or remove felt Hold ruler perpendicular to table Hold ball next to ruler (to assure height of drop) Drop ball	(If investigating deep push: stack two blocks. Otherwise, place one) Place slat on block at correct length Place bean bag at end of slat Hold block and slat while pushing down other end Hold measuring tape next to bean bag Release slat

### *Analyses*

Latent Transition Analysis (LTA) of the worksheet scores was performed to identify proficiency profiles and determine transition paths or learning trajectories through these profiles. LTA was performed using Mplus (Muthén & Muthén, 1998-2017). As Mplus was unable to process all nine worksheets due to computational limitations, we decided to analyse the first worksheet of each lesson. This was considered the best option because there was only one worksheet for the fifth lesson, and children's writings on the second worksheet could be influenced by their experiences during the first part of the lesson. The four component scientific reasoning skills, each measured at five time points, were used to determine the latent proficiency profiles. Profiles were constrained to be the same at each time point, but profile membership was not, allowing for comparisons over time. To ensure a stable solution, the analysis was run with 100 random starting values with a maximum of 20 iterations. The number of latent profiles was determined according to the guidelines proposed by Nylund et al. (2007), using the Bayesian information criterion (BIC), sample-size adjusted BIC, Akaike's information criterion (AIC), and theoretical considerations. Their simulation study showed that for a study with continuous latent profile indicators and less than 200 participants, the Bayesian information criterion (BIC) was most successful for selecting the optimal number of profiles. This information criterion was therefore leading in our model selection.

Figure 1

Overview of worksheet structure and coding. Assignments marked with a  were scored according to the evaluation criteria.

Structure	Example	Evaluation criteria												
	 <p><b>Pendulum swing</b> Rope length Weight Amplitude</p>													
 <p><b>Hypothesizing</b></p> <p>What do you think will happen to the <i>outcome variable</i> if you change the <i>input variable</i>?</p> <p>Sentence starter</p>	<p>What do you think will happen to the period of a swing if you change the length of the rope?</p> <p><input checked="" type="checkbox"/> I think....</p>	<ul style="list-style-type: none"> <li>- An <b>effect</b> was described<sup>1</sup></li> <li>- The <b>direction</b> of the effect was described<sup>1</sup></li> <li>- The <b>variables involved</b> were described<sup>1</sup></li> </ul>												
 <p><b>Experimenting</b></p> <p>Think up two experiments to check if the <i>input variable</i> affects the <i>outcome</i>. For each experiment you have to choose a <i>value for variable one, variable two and variable three</i>.</p> <p>Picture of focal variable</p> <p>Table for contrasting experiments</p>	<p>Think up two experiments to check if the length of the rope affects the period. For each experiment you have to choose the rope length, how far it's pulled aside and the pendulum's weight.</p>  <table border="1" data-bbox="515 937 734 1055"> <thead> <tr> <th><input checked="" type="checkbox"/> Experiment 1A</th> <th>Experiment 1B</th> </tr> </thead> <tbody> <tr> <td>Rope length: _____</td> <td>Rope length: _____</td> </tr> <tr> <td>How far pulled: _____</td> <td>How far pulled: _____</td> </tr> <tr> <td>Weight: _____</td> <td>Weight: _____</td> </tr> </tbody> </table>	<input checked="" type="checkbox"/> Experiment 1A	Experiment 1B	Rope length: _____	Rope length: _____	How far pulled: _____	How far pulled: _____	Weight: _____	Weight: _____	<ul style="list-style-type: none"> <li>- <b>Comparison</b> is possible: at least one variable has been changed<sup>2</sup></li> <li>- <b>Fair comparison</b> is possible: only one variable has been changed<sup>2</sup></li> <li>- Experiment <b>aligns with hypothesis</b>: focal variable has been changed<sup>2</sup></li> </ul>				
<input checked="" type="checkbox"/> Experiment 1A	Experiment 1B													
Rope length: _____	Rope length: _____													
How far pulled: _____	How far pulled: _____													
Weight: _____	Weight: _____													
 <p><b>Interpreting data</b></p> <p>Table for results</p> <p>Do you see a difference between <i>value one</i> and <i>value two</i> in the table?</p> <p>Explanation prompt</p>	<table border="1" data-bbox="512 1106 734 1219"> <thead> <tr> <th></th> <th>Experiment A</th> <th>Experiment B</th> </tr> </thead> <tbody> <tr> <td>Trial 1</td> <td></td> <td></td> </tr> <tr> <td>Trial 2</td> <td></td> <td></td> </tr> <tr> <td>Trial 3</td> <td></td> <td></td> </tr> </tbody> </table> <p><input checked="" type="checkbox"/> Do you see a difference between a short and a long rope in the table?</p> <p><input checked="" type="checkbox"/> How do you know?</p>		Experiment A	Experiment B	Trial 1			Trial 2			Trial 3			<ul style="list-style-type: none"> <li>- Based on the gathered data, a correct <b>inference</b> was made<sup>3</sup></li> <li>- The <b>explanation of the inference</b> refers to the data or outcome variable<sup>3</sup></li> <li>- The data on which the inference was based are <b>described</b> or the outcome is <b>explained</b><sup>3</sup></li> </ul>
	Experiment A	Experiment B												
Trial 1														
Trial 2														
Trial 3														
 <p><b>Drawing conclusions</b></p> <p>You can now answer the research question: Does the <i>input variable</i> affect the <i>outcome variable</i>?</p> <p>Sentence starter</p>	<p>You can now answer the research question: Does the length of the rope affect the period of a swing?</p> <p><input checked="" type="checkbox"/> My research shows ...</p>	<ul style="list-style-type: none"> <li>- The <b>effect that was found</b> was described<sup>1</sup></li> <li>- The <b>direction</b> of the effect was described<sup>1</sup></li> <li>- All <b>variables involved</b> were described<sup>1</sup></li> </ul>												

<sup>1</sup>Lazonder et al. (2010) <sup>2</sup>Chen and Klahr (1999); Peteranderl (2019) <sup>3</sup>Moritz, in Ben-Zvi and Garfield (2004); Kanari and Millar (2004)

## Results

### *Descriptive statistics*

Table 2 shows means and standard deviations for the four component skills at all five time points. These descriptive statistics show different patterns of development for the component skills. However, high standard deviations for most component skills and at most time points indicate large variation, warranting a further look at the underlying latent profiles.

**Table 2**

*Means and standard deviations per lesson for each component skill*

	Lesson 1		Lesson 2		Lesson 3		Lesson 4		Lesson 5	
	M	SD	M	SD	M	SD	M	SD	M	SD
Hypothesizing	1.65	1.11	1.23	1.15	1.48	0.93	1.43	0.83	1.58	1.10
Experimenting	1.33	0.83	2.14	1.01	2.07	0.95	2.08	0.97	2.02	0.98
Interpreting data	2.12	0.63	2.07	0.74	2.01	0.75	1.91	0.64	1.93	0.69
Drawing conclusions	1.58	1.10	1.69	1.21	1.48	0.92	1.48	0.94	1.65	1.06

### *Number of latent profiles*

As a first step, the number of latent proficiency profiles was determined following the recommendations by Nylund et al. (2007). Table 3 shows that all information criteria consistently improved when the number of profiles was increased from two to four. The four-profile solution also yielded the lowest value for entropy, indicating relatively low uncertainty. When the number of profiles was further increased to five, the values of Akaike's information criterion (AIC) and the sample-size adjusted Bayesian Information Criterion (BIC) improved slightly – but not as much as in previous model iterations – whereas the BIC, our leading information criterion, did not. Content wise, this five-profile solution resembled the four-profile solution (see Figure 2) and added a profile with extremely low scores for experimenting. As this additional profile applied to less than 2% of the sample, and because children with low and extremely low scores need similar support, this profile was not considered a meaningful addition. Further extension from five to six

profiles caused none of the fit criteria to improve – it even worsened the value of the BIC – and provided no meaningful theoretical addition. Therefore, the model with four profiles was selected.

**Table 3**  
*Information criteria of LTA models with two to six proficiency profiles*

	Number of profiles				
	2	3	4	5	6
Number of free parameters	37	58	87	124	169
Log-likelihood	-3298.070	-3239.175	-3111.33	-3031.358	-2978.558
Entropy	.914	.858	.845	.904	.906
AIC	6670.14	6594.351	6396.664	6310.715	6295.115
BIC	6785.283	6774.846	6667.407	6696.602	6821.041
Sample-size adjusted BIC	6668.138	6591.214	6391.958	6304.008	6285.974

*Note.* AIC = Akaike’s information criterion BIC = Bayesian information criterion

**Figure 2**  
*Latent proficiency profiles in the four-, five- and six-profile models*

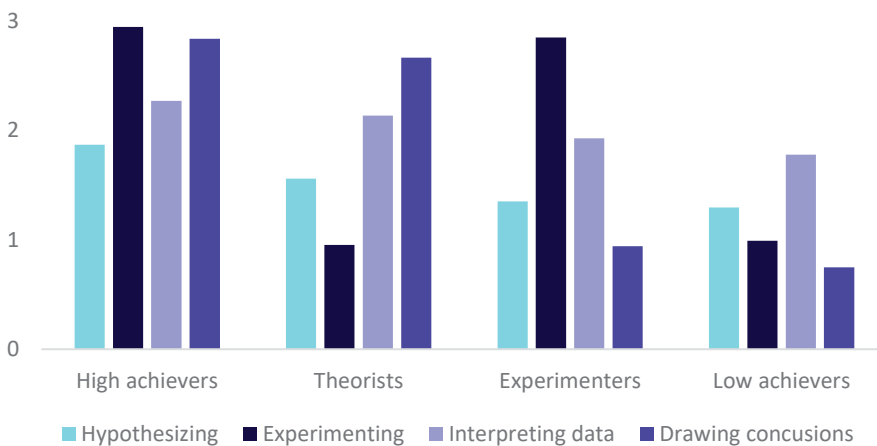


*Note:* The four bars in each cluster represent the component skills of hypothesizing, experimenting, interpreting data and drawing conclusions, respectively.

### Profile characteristics

Mean scores of the component scientific reasoning skills in each profile can be found in Table 4 and are graphically presented in Figure 3. Two profiles were readily interpretable: the *high achievers*, who scored high on all component skills, and the *low achievers*, who scored low on all component skills. The other two profiles were mixed. In one profile, scores were slightly below those of the high achievers except for experimenting: for this skill, scores were more in line with the low achievers. As such, one could say children in this profile are good at the theoretical aspects of an inquiry, but mediocre in practical experimentation. We therefore call this profile *theorists*. In the other mixed profile, the opposite happens: scores are in general slightly higher than those of the low achievers except for experimenting: for this skill, scores are more in line with the high achievers. Thus, one could say that although children in this profile have grasped the practical experimentation aspect of an inquiry, they fall short on the theoretical aspects. We therefore call this profile *experimenters*. It is important to note that, although four distinct profiles were found, differences across all four profiles were relatively small for hypothesizing and interpreting data, and large for experimenting and drawing conclusions. Nonetheless, scores for hypothesizing and interpreting data were in line with the general proficiency level of the profiles – that is, lowest for low achievers, slightly higher for experimenters and theorists, and highest for high achievers.

Figure 3  
Mean scientific reasoning scores in the four latent profiles



**Table 4***Latent profiles and sample proportions at each measurement point*

	Hypothesizing	Experimenting	Interpreting data	Drawing conclusions	Proportion of the sample (%)				
	M (SE)	M (SE)	M (SE)	M (SE)	L1	L2	L3	L4	L5
High achievers	1.87 (0.10)	2.95 (0.02)	2.27 (0.07)	2.84 (0.04)	6.6	30.7	18.7	18.7	22.9
Theorists	1.56 (0.11)	0.95 (0.03)	2.14 (0.09)	2.66 (0.06)	24.7	12.6	8.4	13.9	15.7
Experimenters	1.35 (0.08)	2.85 (0.04)	1.93 (0.05)	0.94 (0.05)	12.0	31.3	39.8	37.3	34.9
Low achievers	1.29 (0.08)	1.00 (0.02)	1.78 (0.07)	0.75 (0.04)	56.6	25.3	33.1	30.1	26.5

*Note.* Measurement points (i.e., lessons) are indicated by L1 to L5



*Transition probabilities*

The right section of Table 4 shows the percentage of children in each profile during each lesson. Most children were classified as either low achievers (56.6%) or theorists (24.7%) in Lesson 1. As both profiles are characterized by poor experimenting skills, many children initially struggled with the design and conduct of an inquiry. This ability improved over the lessons, as can be seen from the increasing percentages of experimenters (up 19 percentage points in Lesson 2) and high achievers (up 24 percentage points in lesson 2), both of which have high average experimenting scores. This result suggest that part of the low achievers and theorists became experimenters or high achievers at some point, mostly in the first half of the lesson series. In the second half, the distribution of children over the profiles appeared to stabilize.

However, these accumulated change patterns do not show among which profiles the majority of transitions took place. We therefore calculated latent transition probabilities between distinct proficiency profiles over lessons; these statistics can be found in Table 5 and an interactive visual representation is offered in the online supplementary materials<sup>1</sup>. It is important to note that there were many transitions between profiles, which only stabilized towards the end of the lessons series: the probability to remain in the same profile (found on the downward diagonals of Table 5) increased as the lessons progressed, and in the last two lessons approximately half of the children remained in the same profile. As an exception, low achievers were most likely to remain in the same profile throughout the lesson series. High achievers and experimenters tended to iterate among profiles, but remained mostly stable after Lesson 2. Theorist are a remarkable profile as children were highly unlikely to remain in this profile over multiple lessons. This indicates a rather unstable transition path over multiple lesson from this profile.

---

<sup>1</sup> See <http://www.erikaschlatter.nl/dissertation/onlinesupplements>

**Table 5**  
*Latent transition probabilities*

Profile	Transition probability towards next lesson			
	High Achievers	Theorists	Experimenters	Low Achievers
Lesson 1				
High Achievers	.317	.000	.542	.140
Theorists	.350	.119	.264	.266
Experimenters	.663	.093	.244	.000
Low Achievers	.196	.193	.302	.309
Lesson 2				
High Achievers	.363	.109	.344	.184
Theorists	.124	0	.405	.471
Experimenters	.029	.026	.59	.355
Low Achievers	.212	.211	.174	.402
Lesson 3				
High Achievers	.619	.086	.238	.058
Theorists	.344	.000	.172	.485
Experimenters	.079	.041	.638	.241
Low Achievers	.024	.311	.196	.469
Lesson 4				
High Achievers	.584	.052	.364	.000
Theorists	.224	.484	.049	.242
Experimenters	.247	.049	.533	.171
Low Achievers	.000	.276	.191	.533

The probability of extreme transitions (found on the upward diagonals of Table 5: between high and low achievers, and between experimenters and theorists) was relatively low and decreased toward the end of the lesson series. The results point to two possible upward

paths: from low to high achieving via the experimenters profile, or from low to high achieving via the theorists profile.

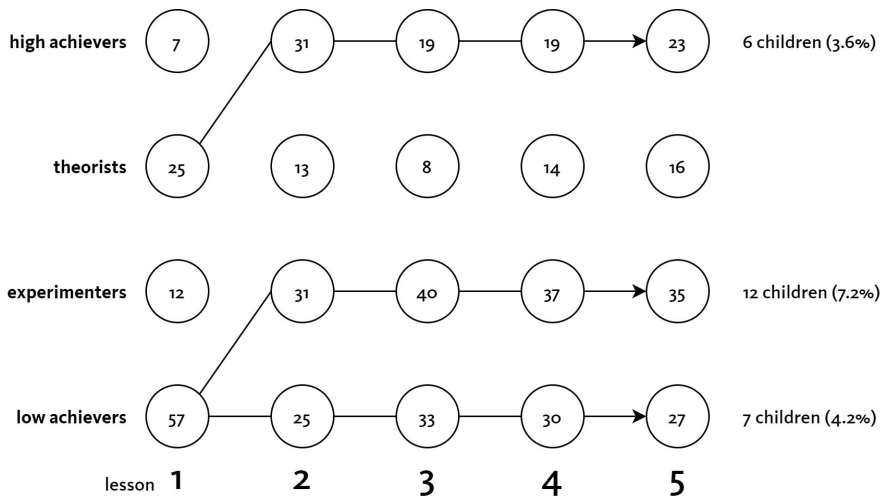
Finally, the probability to transition from a profile with high experimenting skills (high achievers or experimenters) to a profile in which this skill is less well mastered (low achievers or theorists) was low, indicating that once experimenting is learned, children are unlikely to fall back.

### *Transition paths*

In addition to the transition probabilities, transition paths can give more specific information about the most likely learning trajectories by identifying the specific path each child is most likely to take over the five lessons. Consistent with the dynamic picture that emerged from the transition probabilities, 102 transition paths were found, the majority of which were taken by less than 2.5% of the children.

**Figure 4**

*Number of children in each profile during the five lessons and their prevalent transition paths*



Only three specific transition paths were taken by a larger share of the sample (see Figure 4). The first of these paths is a flat line: seven children (4.2%) started out in the low achievers profile and never transitioned to another profile. The other two paths show some improvement: twelve children (7.2%) move from low achievers to experimenters and six

children (3.6%) move from theorists to high achievers within the first two lessons. These paths confirm that, once children have learned to experiment, they are unlikely to transition to a profile associated with low experimenting scores. It must be noted that although these most prevalent paths are relatively stable and show improvement in the early stages of the lesson series, the remaining 99 transition paths, each taken by less than 2.5% of the children, are less stable and show transitions in later lessons as well.

## Discussion

This study used process data from a fifth grade lesson series to identify scientific reasoning proficiency profiles and examine transitions between these profiles over time. Four profiles were distinguished and although many different transition paths were found, only three paths were taken by more than 2.5% of children. These results indicate that the component scientific reasoning skills develop differently in a group of same-age children, thus confirming the importance of treating scientific reasoning as a multidimensional construct (Edelsbrunner & Dablander, 2018). Furthermore, the high number of transitions confirm that learning scientific reasoning, like learning in many other subjects, is not linear (Flynn et al., 2007).

### *Profiles and transition paths*

Consistent with hypotheses, different proficiency profiles emerged from the data, two of which were homogeneous (low achievers and high achievers) and the other two were mixed (theorists and experimenters). Scores in the homogeneous profiles were either low or high on all scientific reasoning skills, whereas the mixed profiles deviated from these uniform profiles mainly because of scores on experimenting. That is, theorists resembled the high achievers except for poor experimenting scores, and experimenters scored low on all skills except experimenting. Although the latter profile was expected to be found because experimenting skills develop from an early age, the former was not, which underlines the added value of person-centred analysis methods such as LTA.

Hypotheses for the second research question predicted that most children would improve in scientific reasoning and, hence, transition to a higher-esteemed profile. Worksheet scores of the first lesson indicate that most children had room for improvement: more than half of the sample was classified as low achievers and only 6.6% as high achievers. These mediocre scores confirm that scientific reasoning is difficult to fifth graders, with experimenting being the initially least well-mastered skill (Lorch et al., 2014;

Peteranderl, 2019). Another explanation for the low starting level is that our sample had little experience in conducting a self-directed investigation. However, children's worksheet scores increased during the first three lessons and then stabilized. This growth was predicted to occur, but the stagnation was not and could point to a possible novelty effect—that is, performing an isomorphic inquiry task every week can become boring. Despite this slight stagnation, the number of transition paths found was high. Considering the large diversity in children's scientific reasoning found in earlier studies (Koerber & Osterhaus, 2019; Piekny & Maehler, 2013; Van der Graaf, 2020), the absence of a clear-cut path might be another instance of this diversity.

The theorists and experimenters profiles merit some further discussion because they represent the intermediate stages of skill development. These profiles appear to be each other's opposites: experimenters are good at experimenting but not at drawing conclusions whereas theorists are good at drawing conclusions and poor in experimenting. It is theoretically unlikely that children would pass through both profiles in their process toward becoming proficient in scientific reasoning, and modest transition probabilities support this idea. Furthermore, the current study shows few transitions from profiles in which experimenting is mastered (experimenters or high achievers) to profiles in which it is not (theorists or low achievers), which suggests that experimenting is not easily unlearned (cf. Lorch et al., 2017). So, although no consolidated transition paths were found for large groups of children, there is reason to suspect two main learning trajectories from low to high achievers exist: one via the theorists profile and one via the experimenters profile.

The pathway through the theorists profile seems the least stable because transition probabilities of this profile deviated from those of the other profiles, which stabilized halfway through the lesson series. Two possible explanations can be given for this outcome, which both relate to the discerning features of the theorists profile, namely poor experimenting skills and high scores on drawing conclusions. First, it is plausible that some theorists learned to design systematic experiments in one of the first lessons, and because this skill is unlikely to be unlearned (e.g., Lorch et al., 2017), the odds of returning to the theorists profile was negligible. Secondly, drawing conclusions hinges on understanding the outcomes of an investigation as well as being able to write precisely. The latter is crucial for a good conclusion, yet it is possible that children who understood their outcomes did not always write down a clear and precise conclusion (e.g., Salac & Franklin, 2020). We do not consider such instances to be false negatives, precisely because a conclusion should clearly communicate the outcomes of an investigation. Still, this phenomenon could have increased instability in the theorists profile.

Along the same line, experimenting appears to be a major barrier for some children (those passing through the theorists profile), whereas it seems less problematic for others (those passing through the experimenters profile). As such, some children with an intermediate understanding of scientific reasoning need experimenting support, whereas others need support for drawing conclusions. This implication supports the claim by Kuhn and Dean (2005) that developing scientific reasoning involves more than learning to design and run experiments. Nonetheless, past research has overwhelmingly focused on this skill (Koerber & Osterhaus, 2019) and consider it a key requirement for the development of scientific reasoning (Schwichow et al., 2016; Zacharia et al., 2015; Zimmerman, 2007). Our data suggest that this view is too restrictive and that more research is needed to understand the development of other scientific reasoning skills.

### *Limitations*

A pitiful limitation was that the Mplus software could not handle all available data. Running an LTA with nine worksheets per child exceeded the program's processing capacities – but fortunately the analysis of the first worksheet of each lesson proved doable. Although five measurement points still compares favourably to most learning research using LTA, it remains unknown whether different profiles would have emerged if all worksheets were included, and whether there was a difference in transition probabilities within and between lessons. Another limitation is the use of worksheets per se as a data source. The tacit assumption underlying the LTA was that children's entries on the worksheets reflect their true abilities. However, related research showed that inferring children's understanding from their own writings can lead to false negatives or at the very least underestimate what they are actually capable of (Salac & Franklin, 2020). A final limitation concerned the topics children investigated during the lessons. As using a single topic throughout the lesson series was undesirable for obvious reasons, we designed five structurally equivalent inquiries (see Table 1) which were uniformly supported by worksheets. Still, differences in the topic of inquiry and the equipment at hand could have caused spurious variation in children's scientific reasoning and might have inflated the number of transition paths found.

### *Practical implications*

The outcomes of this study can help teachers optimize their lessons through in-class differentiation (Van Geel et al., 2018) and deployment of targeted interventions (Bray & Dziak, 2018). In preparing a lesson period, effective teachers base the instructional support

they intend to give on the developmental level of the children in their classroom (Van Geel et al., 2018). Some of these decisions apply to the entire class. In the case of scientific reasoning in upper primary education, the component skills of hypothesizing and interpreting data warrant such a whole class approach because scores for these skills are consistently low across profiles. Research on differentiation indicates that when skills are similarly developed in students across the class, adjustment of instruction is less necessary (Tomlinson et al., 2003). Experimenting, on the other hand, requires a more differentiated approach because the initial proficiency and developmental pace differs among children. Our results suggest that support for experimenting should be considered for individual children on a lesson-by-lesson basis. When children show problems with experimenting, more support should be given to help them distinguish the variable of interest from the control variables (e.g., Chen & Klahr, 1999; Lorch et al., 2017). Once children have shown successful experiments this skill is generally maintained so the support can be withdrawn. Here, differences in prior knowledge and learning pace should drive adjustment of instruction and support, which could potentially lead to more effective and efficient skill acquisition as is shown in other domains (Vanbecelaere et al. 2020). Support for drawing conclusions should be delivered similarly, in particular to low achievers and children who merely excel in experimenting. The focus of this support should be on the connection between outcomes and conclusion, and on precise formulation of the conclusion. Questions remain about the nature and specificity of effective teacher support for all component scientific reasoning skills. Future research should address these issues as well as the preparation of (prospective) primary science teachers to monitor children's progress and support needs.

### *Conclusion*

With its focus on quantitative analysis of short-term development of multiple component skills of scientific reasoning, the current study expands existing research. Previous studies predominantly examined single component skills (as discussed by Koerber & Osterhaus, 2019), often in single-sitting interventions (e.g., Chen & Klahr, 1999; Lorch et al., 2017), or on interventions spanning multiple lessons and multiple skills where the focus lied mostly on qualitative analysis of small groups of students (e.g., Kuhn & Dean, 2005; Zohar & Dori, 2003). The current study confirmed that same-age children differ considerably in their command of the component scientific reasoning skills, and through the latent profiles offers a new and accessible way to interpret these differences. Wide variation exists also in how that proficiency is reached – so much so that the transition paths taken by quite many

children proved unique. This confirms that the common practice of analysing pre- and post-tests, even if these address multiple component skills (e.g., Kruit et al., 2018; Van der Graaf, 2020), does not paint a complete picture of the development of scientific reasoning in the classroom. These insights further our understanding of how scientific reasoning develops in the upper primary grades and can help teachers adapt their science lessons to children's individual needs. The current study is a starting point for future quantitative and qualitative explorations of how scientific reasoning is learned and should be taught during short instruction episodes. We think LTA is well suited for this type of research.



## References

- Ben-Zvi, D., & Garfield, J. (2004). *The challenge of developing statistical literacy, reasoning and thinking* (D. Ben-Zvi & J. Garfield, Eds.). Kluwer Academic Publishing.
- Boom, J., & ter Laak, J. (2007). Classes in the balance: Latent class analysis and the balance scale task. *Developmental Review, 27*(1), 127-149. <https://doi.org/10.1016/j.dr.2006.06.001>
- Bray, B. C., & Dziak, J. J. (2018). Commentary on latent class, latent profile, and latent transition analysis for characterizing individual differences in learning. *Learning and Individual Differences, 66*, 105-110. <https://doi.org/10.1016/j.lindif.2018.06.001>
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098-1120. <https://doi.org/10.1111/1467-8624.00081>
- Edelsbrunner, P. A., & Dablander, F. (2018). The psychometric modeling of scientific reasoning: A review and recommendations for future avenues. *Educational Psychology Review, 31*(1), 1-34. <https://doi.org/10.1007/s10648-018-9455-5>
- Edelsbrunner, P. A., Schalk, L., Schumacher, R., & Stern, E. (2018). Variable control and conceptual change: A large-scale quantitative study in elementary school. *Learning and Individual Differences, 66*, 38-53. <https://doi.org/10.1016/j.lindif.2018.02.003>
- Flynn, E., Pine, K., & Lewis, C. (2007). Using the microgenetic method to investigate cognitive development: An introduction. *Infant and Child Development, 16*(1), 1-6. <https://doi.org/10.1002/icd.503>
- Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2018). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning and Individual Differences, 66*, 4-15. <https://doi.org/10.1016/j.lindif.2017.11.001>
- Howard, M. C., & Hoffman, M. E. (2018). Variable-centered, person-centered, and person-specific approaches: Where theory meets the method. *Organizational Research Methods, 21*(4), 846-876. <https://doi.org/10.1177/1094428117744021>
- Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching, 41*(7), 748-769. <https://doi.org/10.1002/tea.20020>
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1-48. [https://doi.org/10.1207/s15516709cog1201\\_1](https://doi.org/10.1207/s15516709cog1201_1)
- 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.org/10.1080/15248372.2019.1620232>
- Kruit, P. M., Oostdam, R. J., Van den Berg, E., & Schuitema, J. A. (2018). Effects of explicit instruction on the acquisition of students' science inquiry skills in grades 5 and 6 of primary education. *International Journal of Science Education, 40*(4), 421-441. <https://doi.org/10.1080/09500693.2018.1428777>
- Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science, 16*(11), 866-870. <https://doi.org/10.1111/j.1467-9280.2005.01628.x>
- Kuhn, D., Ramsey, S., & Arvidsson, T. S. (2015). Developing multivariable thinkers. *Cognitive Development, 35*, 92-110. <https://doi.org/10.1016/j.cogdev.2014.11.003>
- Lazonder, A. W., Hagemans, M. G., & De Jong, T. (2010). Offering and discovering domain information in simulation-based inquiry learning. *Learning and Instruction, 20*(6), 511-520. <https://doi.org/10.1016/j.learninstruc.2009.08.001>

- 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>
- Lorch, R. F., Lorch, E. P., Freer, B., Calderhead, W. J., Dunlap, E., Reeder, E. C., Van Neste, J., & Chen, H. T. (2017). Very long-term retention of the control of variables strategy following a brief intervention. *Contemporary Educational Psychology*, 51, 391-403. <https://doi.org/10.1016/j.cedpsych.2017.09.005>
- Lorch, R. F., Lorch, E. P., Freer, B. D., Dunlap, E. E., Hodell, E. C., & Calderhead, W. J. (2014). Using valid and invalid experimental designs to teach the control of variables strategy in higher and lower achieving classrooms. *Journal of Educational Psychology*, 106(1), 18-35. <https://doi.org/10.1037/a0034375>
- 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>
- Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide*. (Vol. 8). Muthén & Muthén.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4), 535-569. <https://doi.org/10.1080/10705510701575396>
- Peteranderl, S. (2019). *Experimentation skills of primary school children*. [Doctoral dissertation, ETH Zürich]. ETH Zürich Research Collection. <https://doi.org/10.3929/ethz-b-000370663>
- Piekny, J., Grube, D., & Maehler, C. (2014). The development of experimentation and evidence evaluation skills at preschool age. *International Journal of Science Education*, 36(2), 334-354. <https://doi.org/10.1080/09500693.2013.776192>
- Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology*, 31(2), 153-179. <https://doi.org/10.1111/j.2044-835X.2012.02082.x>
- Reimann, P. (2009). Time is precious: Variable- and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning*, 4(3), 239-257. <https://doi.org/10.1007/s11412-009-9070-z>
- Salac, J., & Franklin, D. (2020, June 15-19). If they build it, will they understand it? Exploring the relationship between student code and performance. ITiCSE '20, Trondheim, Norway.
- Schlatter, E., Molenaar, I., & Lazonder, A. W. (2020). Individual differences in children's development of scientific reasoning through inquiry-based instruction: Who needs additional guidance? *Frontiers in Psychology*, 11, Article 904. <https://doi.org/10.3389/fpsyg.2020.00904>
- Schneider, M., & Hardy, I. (2013). Profiles of inconsistent knowledge in children's pathways of conceptual change. *Developmental Psychology*, 49(9), 1639-1649. <https://doi.org/10.1037/a0030976>
- 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, 101923. <https://doi.org/10.1016/j.cedpsych.2020.101923>
- Schwichow, M., Zimmerman, C., Croker, S., & Härtig, H. (2016). What students learn from hands-on activities. *Journal of Research in Science Teaching*, 53(7), 980-1002. <https://doi.org/10.1002/tea.21320>
- Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., Conover, L. A., & Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted*, 27(2), 119-145. <https://doi.org/10.1177/016235320302700203>

- Van der Graaf, J. (2020). Inquiry-based learning and conceptual change in balance beam understanding. *Frontiers in Psychology, 11*, Article 1621. <https://doi.org/10.3389/fpsyg.2020.01621>
- Van der Graaf, J., Segers, E., & Verhoeven, L. (2018). Individual differences in the development of scientific thinking in kindergarten. *Learning and Instruction, 56*, 1-9. <https://doi.org/10.1016/j.learninstruc.2018.03.005>
- Van Geel, M., Keuning, T., Frèrejean, J., Dolmans, D., Van Merriënboer, J., & Visscher, A. J. (2018). Capturing the complexity of differentiated instruction. *School Effectiveness and School Improvement, 30*(1), 51-67. <https://doi.org/10.1080/09243453.2018.1539013>
- Zacharia, Z. C., Manoli, C., Xenofontos, N., De Jong, T., Pedaste, M., Van Riesen, S. A. N., Kamp, E. T., Mäeots, M., Siiman, L., & Tsourlidaki, E. (2015). Identifying potential types guidance for supporting student inquiry when using virtual and remote labs in science: A literature review. *Educational Technology Research and Development, 63*(2), 257-302.
- 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>
- Zohar, A., & Dori, Y. (2003). Higher order thinking skills and low-achieving students: Are they mutually exclusive? *Journal of the Learning Sciences, 12*(2), 145-181. [https://doi.org/10.1207/S15327809JLS1202\\_1](https://doi.org/10.1207/S15327809JLS1202_1)



# Chapter 5

**Adapting scientific  
reasoning instruction to  
children's needs:**

**Effects on learning  
processes and learning  
outcomes**

### **Abstract**

Scientific reasoning is an important skill that helps children understand the world around them. Teaching scientific reasoning starts in primary school and can be challenging because not all component scientific reasoning skills develop at the same age and not all children learn these skills at the same pace. A differentiated teaching approach thus seems called for. The current study compared two types of adaptive instruction to a non-adaptive control condition. Over the course of four lessons, children in the control condition ( $n = 49$ ) practiced scientific reasoning skills with the help of worksheets that offered medium support. Children in the two adaptive conditions received worksheets with tailor-made support that was either based on their standardized test scores (macro-adaptive condition;  $n = 58$ ) or their performance in the previous lesson (micro-adaptive condition;  $n = 46$ ). Thus, the two adaptive conditions differed regarding the learner variables used to assign children to a level of support and the frequency with which these learner variables were obtained to adapt the worksheets. Analysis of children's pre- and post-test scores showed comparable improvements in scientific reasoning in all three conditions. Because many children in both adaptive conditions received medium support, additional analyses were done on children in the macro-adaptive condition who received high or low-support worksheets, and their control group counterparts who would have qualified for high or low support had they been in the macro-adaptive condition. Learning gains for these groups were similar. Scores on the worksheets also improved, and this improvement did interact with condition.

This chapter is based on:

Schlatter, E., Molenaar, I., & Lazonder, A. W. (resubmitted).

Adapting scientific reasoning instruction to children's needs: Effects on learning processes and learning outcomes.

## Introduction

Scientific reasoning refers to a skillset that helps primary school children understand science content and the world at large. Scientific reasoning comprises multiple skills, which can be subsumed under the three core components of scientific inquiry: hypothesizing, experimenting and evaluating outcomes (Klahr & Dunbar, 1988; Zimmerman, 2007). The teaching of these skills is both important and challenging because children in the same classroom differ considerably in terms of their entry levels (Koerber et al., 2015; Koerber & Osterhaus, 2019; Piekny & Maehler, 2013) and learning trajectories (Schlatter et al., 2021 [Chapter 4]; Schneider & Hardy, 2013; Van der Graaf, 2020). These individual differences call for adaptive teaching in which information on children's performance or general aptitude is used to adjust instruction or teaching materials (Alevén et al., 2016). For adaptivity in scientific reasoning instruction, few guidelines exist. This study aimed to contribute to the development of such guidelines by comparing the instructional effectiveness of two adaptive approaches to a non-adaptive control condition.

### *Learning and teaching scientific reasoning in the primary school years*

Scientific reasoning starts developing at preschool age: young children demonstrate an emergent understanding of experimenting (Piekny et al., 2014; Van der Graaf et al., 2015) and evaluation of outcomes (Köksal-Tuncer & Sodian, 2018; Piekny et al., 2014). Both cross-sectional and longitudinal data have shown that children improve in scientific reasoning over the course of their primary education (Koerber et al., 2015; Lazonder et al., 2021; Piekny & Maehler, 2013). Longitudinal research has shown that not all children develop at the same pace: some progress slowly whereas others make big leaps in scientific reasoning, which do not always occur at the same grade level (Lazonder et al., 2021). The natural development of scientific reasoning can be supported with both direct and inquiry-based instruction. Direct instruction for scientific reasoning has mainly been studied for teaching the control-of-variables strategy (CVS; an important aspect of experimenting). In their meta-analysis, Schwichow et al. (2016) found that demonstrating experiments in front of the class and inducing cognitive conflict are particularly effective for learning the CVS. Such interventions generally involve the teacher showing an uncontrolled experiment, which then invokes classroom discussion on the experiment's flaws, and can be as short as 20 minutes (Chen & Klahr, 1999; Lorch et al., 2014). If the demonstration experiments are not over-simplified (Lorch et al., 2019), the effect of these short interventions can be maintained over many years (Lorch et al., 2017). However, as

these studies assessed a single aspect of scientific reasoning in isolation, it is not clear whether direct instruction is as effective for teaching other component skills.

Inquiry-based instruction, by contrast, has been used in more holistic studies of scientific reasoning, where multiple component skills were taught. This mode of instruction can range from confirmation inquiries in which the research question, methods and solution are given, to open inquiry in which children are free to choose their own questions, methods and solutions. Guided inquiry, an intermediate stage, provides children with some but not all of these elements (Bell et al., 2005). In their meta-analysis of guidance in inquiry-based learning, Lazonder and Harmsen (2016) found that guided inquiry yielded better outcomes in learning activities, performance success and learning outcomes than open inquiry.

Guided inquiry provides structure, for example by limiting the number of independent variables and their possible values, and is characterized by repeated practice. Kuhn and Dean (2005) found that sixth graders who practiced inquiry in a constrained computer simulation and were prompted to pay attention to specific variables drew more correct inferences from their data than their peers who did receive this kind of guidance.

It is important to note here that not all teaching strategies may be effective for all children. For example, Lazonder and Harmsen (2016) found differential effects due to students' age: more specific support appears to be more effective for children compared to adolescents and adults. Although differences between children of the same age group are also well-documented (e.g., Koerber & Osterhaus, 2019; Piekny & Maehler, 2013; Schlatter et al., 2020 [Chapter 3]), it is not yet clear how support can be adapted within the classroom to account for these differences.

### *Towards adaptivity in scientific reasoning instruction*

One way to address differences within science classrooms is through adaptivity: the adjustments of learning materials and teaching strategies based on information about learners (Aleven et al., 2016). Plass and Pawar (2020) distinguished three dimensions of adaptivity: the learner variables used by the adaptive mechanism, the way these variables are measured, and the type of adjustments made to the learning materials and teaching strategies. Learner variables can be cognitive, such as prior knowledge or task performance, as well as motivational, affective and socio-cultural. Measurements can range from one-off, unidimensional observations to continuous measurement of multiple variables. The adjustments can be embedded in the core learning activity, for example in the form of scaffolds, or be applied at an overarching level of preparation, cross-course progression and



learning assessment. Overall, good adaptive teaching is learner-centred, pro-active, and adapts teaching materials to individual students' needs (Tomlinson et al., 2003).

In the realm of science education, research on adaptivity is rather scarce and mostly geared towards the acquisition of content knowledge. Two studies in inclusive classrooms showed that when science learning materials were adapted based on children's ability level as perceived by the teacher, learning outcomes improved for children with and without mild (learning) disabilities (Mastropieri et al., 2006; McCrea Simpkins et al., 2009). Other studies showed that teacher training can be effective as well: children whose teachers were trained in adaptive teaching using either formal learner variables such as standardized tests (e.g., Eysink et al., 2017) or informal learner information from observations and questioning in the classroom (e.g., Vogt & Rogalla, 2009) acquired more content knowledge than children from untrained teachers. Although these studies did not address scientific reasoning in particular, the substantial variation in children's scientific reasoning makes it likely that this is another area where adaptive teaching would be beneficial.

As outlined above, the right amount of guidance is an important aspect of effective inquiry-based instruction in scientific reasoning. As such, adaptive scientific reasoning instruction requires informed decisions on how much guidance should be offered to whom. In order to determine what information can be used as learner variables to adapt guidance (Plass & Pawar, 2020), it is important to understand the nature of individual differences in scientific reasoning. On the one hand, research has shown that individual differences in scientific reasoning among same-age children are attributable to cognitive learner characteristics such as intelligence and spatial reasoning (Mayer et al., 2014), reading comprehension (Lazonder et al., 2021; Schlatter et al., 2020 [Chapter 3]; Schlatter et al., 2021 [Chapter 2]; Van de Sande et al., 2019) and numerical ability (Koerber & Osterhaus, 2019; Schlatter et al., 2020 [Chapter 3]; Schlatter et al., 2021 [Chapter 2]). As most schools have data on children's reading comprehension and mathematical skilfulness readily available, these predictors may be particularly useful learner variables to adapt for.

On the other hand, studies that employed latent profile and latent transition analysis revealed a highly diverse and dynamic learning process. In a cross-sectional study of secondary school students' experimenting skills, Schwichow et al. (2020) identified six proficiency profiles that differed with regard to students' ability to identify, plan, interpret and understand controlled experiments. Some of these profiles showed high, intermediate or low performance on all component skills while other profiles were mixed. For example, students in the 'planning proficiency' profile were highly skilful in planning experiments, but intermediate in the identification, interpretation and understanding of experiments.

Schlatter et al. (2021 [Chapter 4]) found four proficiency profiles as well as a highly dynamic development during the course of five lessons. These differences in the process of learning scientific reasoning suggests that close monitoring of children's performance during a lesson series, resulting in frequent measurement of learner variables directly related to the content, could be important for adaptive scientific reasoning instruction as well. Thus, task performance could be another important learner variable to adapt for.

### *Current study*

Our snapshot of the literature showed that both stable traits and task performance can be useful learner variables to inform the tailoring of the level of guidance in scientific reasoning instruction. As the relative effectiveness of these approaches has not been investigated, the current study compared two adaptive scientific reasoning instruction formats – one based on traits and one on task performance – to a non-adaptive control condition in order to answer the following research questions:

- (1) What is the effect of adaptive scientific reasoning instruction on children's learning outcomes?
- (2) What characterizes children's learning processes, and do these learning processes differ across conditions?

The first experimental condition used a coarse-grained adaptivity mechanism based on stable traits, namely children's reading comprehension and mathematical skilfulness, and will hereafter be called the *macro-adaptive* condition. In this condition, the amount of support was determined prior to the lesson series and remained invariant in all four lessons. The second experimental condition used a fine-grained adaptivity mechanism based on children's task performance, and will hereafter be called the *micro-adaptive* condition. In this condition, children's performance was evaluated after each lesson in order to adapt the teaching materials for the upcoming lesson when necessary.

The adaptive support was provided via the worksheets children used during the lessons, which were similar in all conditions except for the amount of support. As unguided inquiry is ineffective for most children (Alfieri et al., 2011), worksheets in the non-adaptive control condition provided children with medium support through prompts reminding them of what to consider when formulating a hypothesis, setting up an experiment, interpreting data and drawing a conclusion. In both adaptive conditions, this baseline support could be scaled up or down.

As previous studies have shown that adaptive instruction is generally more effective than non-adaptive instruction (Deunk et al., 2018), both adaptive conditions were expected to yield higher learning outcomes than the non-adaptive control condition. Previous analyses of process data (Schlatter et al., 2020 [Chapter 3]) has shown that children's worksheet entries improve over the course of a lesson series. Therefore, it was expected that the quality of children's entries on the worksheets would improve in all conditions – but more rapidly in the adaptive conditions than in the non-adaptive control condition. It seems plausible that the micro-adaptive condition would be more effective than the macro-adaptive condition for a number of reasons. First, the learner variables used in the micro-adaptive condition were closely related to the learning goal, whereas the macro-adaptive condition used more stable, but less closely related, learner variables. Secondly, the frequent measurement of task performance in the micro-adaptive condition allowed for more dynamic adjustment of the teaching materials. It was therefore hypothesized that the micro-adaptive condition would be more effective than the macro-adaptive condition in terms of learning processes and learning outcomes.

## Method

### *Participants*

Seven fifth-grade classrooms from six schools participated in this study. Two of these classrooms were randomly assigned to the control condition, three to the macro-adaptive condition and two to the micro-adaptive condition. The 182 children who attended these classrooms (54% boys) were ages 9–12, with the majority being 10 years old. Parental consent was obtained for all but one child. Data of another 28 children were excluded from analysis because they missed either the pre- or post-test (18 children), more than one lesson (9 children) and/or their standardized progress monitoring test scores were unavailable (8 children). The final sample thus contained 153 children: 49 in the control condition, 58 in the macro-adaptive condition and 46 in the micro-adaptive condition.

### *Intervention*

#### *Lessons*

All children participated in four inquiry-based science lessons, taught by the principal investigator, about primary-school physics topics. All lessons were structured similarly,

starting with a short whole-class discussion of the inquiry cycle and the topic of investigation. Children then formed dyads and did their first 20-minute inquiry. Dyads were formed on an ad-hoc basis, and children in the adaptive conditions could only choose a partner who received the same amount of support. The outcomes of this first round of inquiry were discussed with the whole class. This discussion was guided by questions such as ‘who found an answer to the research question?’ and ‘who found a different result than they expected?’. After this discussion, a second small-group inquiry session (20 minutes) took place in the same dyads, followed by a whole-class discussion of the outcomes and underlying physics principles.

The lessons addressed the scientific reasoning skills of hypothesizing, experimenting, interpreting data and drawing conclusions. These skills were introduced in the first lesson through a demonstration experiment and then practiced in children’s small-group investigations. Each lesson revolved around a different physics topic: the bouncing of balls, rolling on inclined planes, pendulum swing time and the spring capacity of a plank. To investigate these topics, children could manipulate three dichotomous input variables and observe the effect on a continuous output variable. For example, in the bouncing balls experiment the type of ball (a hollow ping pong ball or a filled polystyrene ball), the surface on which it bounced (hard or soft) and the drop height (high or low) could be changed. If an input variable could be interpreted as continuous, such as the drop height, it was dichotomized through the materials. For example, in the bouncing balls experiment children received a ruler with a sticker with the letter H (high) at 30 centimetres and a sticker with the letter L (low) at 10 centimetres. The outcome measure was always a continuous variable (e.g., number of bounces the ball made), and in order to practice the interpretation of messy data children were asked to repeat each experiment three times.

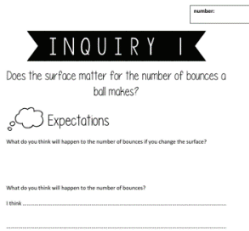
### *Worksheets*

Children’s investigations during each lesson were guided by worksheets that offered either a low, medium or high amount of support (see, for an example of a high-support worksheet, Appendix C), depending on the children’s condition. As illustrated in Figure 1, the low support worksheets provided children with questions that could be answered by filling out templates (i.e., sentence starters and tables). The medium support worksheets contained additional prompts to remind children of what they should consider in formulating a hypothesis, setting up an experiment, interpreting their data and drawing a conclusion. On the high support worksheets, procedural information was added to explain how a particular

action should be performed. Children used one worksheet per lesson. To prevent them from starting the second inquiry cycle before the first was reviewed during the whole-class discussion, each worksheet contained a filler assignment (e.g., a puzzle).

**Figure 1**  
*Worksheet support levels*

**LOW SUPPORT**



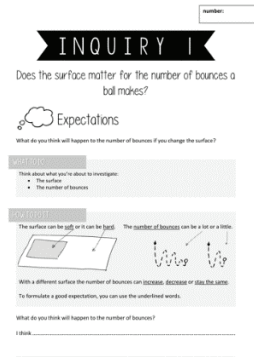
- Questions
- Templates

**MEDIUM SUPPORT**



- Questions
- Prompts
- Templates

**HIGH SUPPORT**



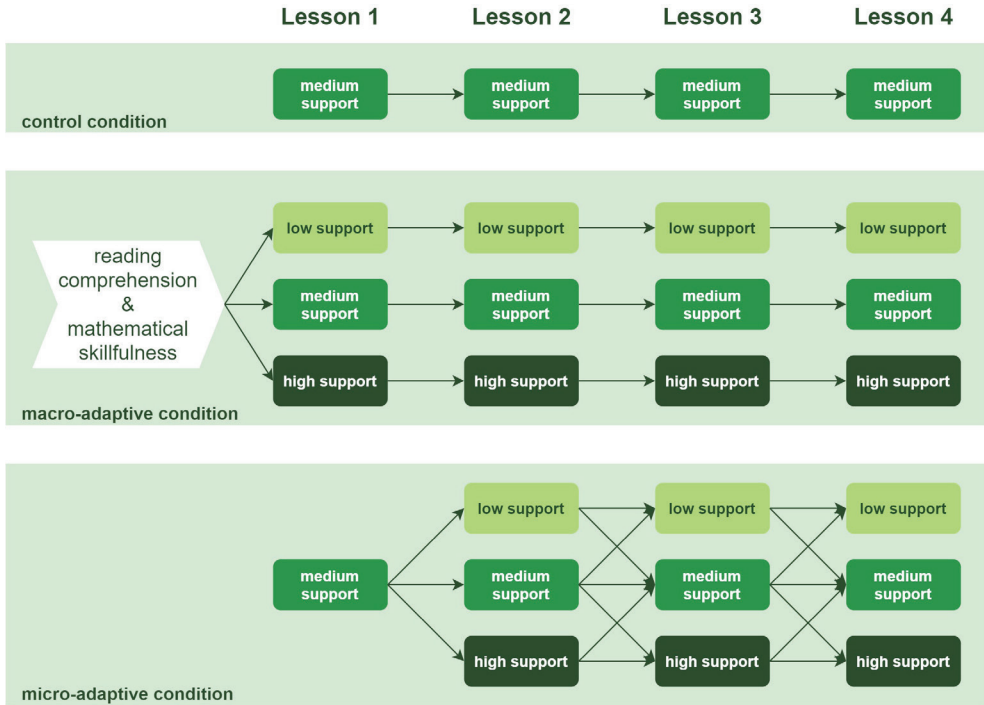
- Questions
- Prompts
- Procedural information
- Templates

*Note.* An example of the high support worksheet can be found in Appendix C.

*Adaptivity mechanisms*

Classrooms were randomly assigned to either the control condition or one of the two adaptive conditions. The adaptivity structure of each condition is visualized in Figure 2. In the non-adaptive control condition, all children received worksheets with medium support during each lesson. The adaptive conditions differed with regard to the frequency of adaptation, which occurred once in the macro-adaptive condition and after each lesson in the micro-adaptive condition. The adaptive conditions also differed with regard to the source of adaptation: standardized test scores in the macro-adaptive condition and task performance during the previous lesson in the micro-adaptive condition.

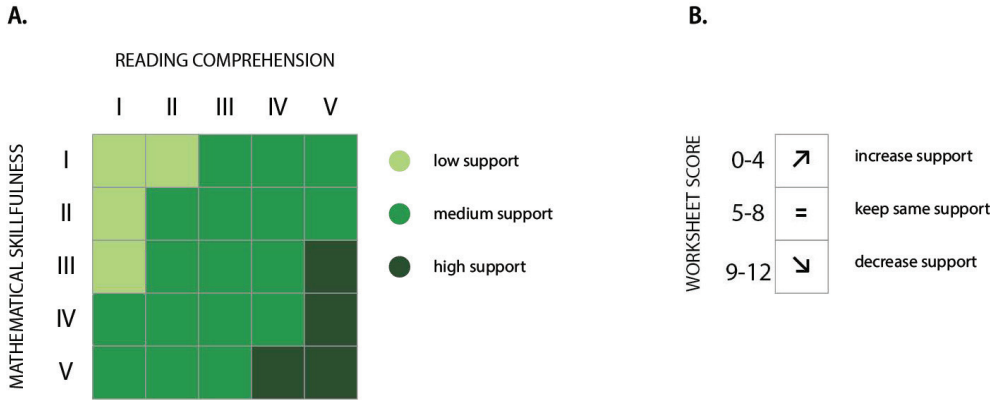
Figure 2  
Adaptivity structures



In the macro-adaptive condition, standardized test scores of reading comprehension and mathematical skillfulness were used to decide the amount of support offered to a child. The standardized progress monitoring tests used by the participating schools offers proficiency levels, ranging from I (highest) to V (lowest). Worksheets were assigned based on these proficiency levels, using the matrix in Figure 3, prior to the first lesson. Because earlier research showed reading comprehension to be a stronger predictor of scientific reasoning than mathematics (Schlatter et al., 2020 [Chapter 3]), reading comprehension level played a more prominent role in worksheet assignment than mathematics level. Previous research (Schlatter et al., 2020 [Chapter 3]) as well as data collected from 195 children in Spring 2020 (in a study that was ultimately aborted due to school closures during the COVID-19 pandemic) suggested that with this assignment scheme about 50% of children would be assigned a medium support worksheet, 20% a high-support worksheet and 30% a low-support worksheet.

Figure 3

Worksheet assignment rules for the macro-adaptive condition (A) and micro-adaptive condition (B).



All children in the micro-adaptive condition received a medium-support worksheet in the first lesson, and after each lesson their task performance was used to determine which worksheet would be offered in the subsequent lesson. To this end, worksheets were scored using a coding scheme developed in previous research (Schlatter et al., 2020 [Chapter 3]). Up to three points could be earned for each component skill (hypothesizing, experimenting, interpreting data and drawing conclusions), leading to a total of 12 points per inquiry cycle. Each lesson comprised two inquiry cycles and children could have different scores per session. In a performance-based assessment such as this, there is a fair chance that children do not show their full understanding at all times (i.e., false negatives; Salac & Franklin, 2020). Therefore, the highest of the two scores was chosen to inform worksheet assignment (see Figure 3). If the highest score was between 0 and 4 points, the support level would be raised in the next lesson (from low to medium support, or from medium to high support). If the highest score was between 5 and 8, the child would receive the same amount of support in the next lesson. If the highest score was between 9 and 12, the support level would be lowered in the next lesson (from high to medium support, or from medium to low support). If a child was absent during the first lesson, they would receive the medium support worksheet in the second lesson. If a child was absent during any of the subsequent lessons, they would be assigned a worksheet based on their performance during the last lesson they attended. Previous research (Schlatter et al., 2020 [Chapter 3]) as well as data collected in Spring 2020 suggested that with this assignment scheme about 50% of children

would be assigned a medium support worksheet, 30% a high-support worksheet and 20% a low-support worksheet.

### *Instruments*

#### *Scientific Reasoning Inventory*

Children's scientific reasoning skills were assessed at pre- and post-test using the Scientific Reasoning Inventory (SRI; Van de Sande et al., 2019), a pencil-and-paper test consisting of 24 multiple-choice items with three or four answer options each. Items on this test are thematically embedded in one of five cover stories, which are meaningful and appealing to children, such as the living conditions of wildlife and sports activities. The test items spanned three scales.

The *hypothesis-evidence coordination* scale (9 items), consisted of two types of items. Five items presented children with four research questions, and asked them to select the question that best matched the research purpose described in the cover story. The nature of these items closely resembled the way in which the skill of hypothesizing was addressed during the lessons. Four additional items measured children's ability to interpret a table with research data. These questions related to the skill of interpreting data as was addressed during the lessons. Although these nine items loaded on the same scale during the initial validation of the SRI, they were practiced separately during the intervention because they represent a different stage of the inquiry cycle.

The second scale, *experimenting* (7 items), required children to select the best experiment based on a cover story. Each item presented children with three experimental designs with either two variables (2 items) or three variables (5 items). For each experiment only one experimental setup allowed for valid causal conclusions. The other experiments were either confounded, did not change any variables, or were controlled but did not manipulate the target variable.

Items on the third scale, *drawing conclusions* (8 items), contained two premises and a question about those premises children could answer with 'yes', 'no' or 'maybe'. These syllogisms were embedded in the overarching cover story. For example, one of the syllogisms in the sports storyline was: 'All children who will go rowing, are wearing shorts. Anna will go rowing. Is she wearing shorts?'

The SRI was administered by the classroom teachers. Even though they checked all pre- and post-tests for missing values, 37 children had skipped one or more questions. To maintain as much data as possible, the mean proportion of correctly completed items was



calculated. No difference in proportional score was found between children with and without missing data. Therefore, proportional scores were used as the dependent variable in the analyses.

### *Standardized test scores*

In order to keep the test burden for this study to a minimum, the results of standardized progress monitoring tests were requested from the schools. All participating schools used the student monitoring program of the National Institute for Educational Testing and Assessment (Cito) in the Netherlands, in which children's cognitive abilities are assessed twice a year. The most recent scores available for all classrooms were those from halfway through Grade 4, because a number of schools opted out of the assessment at the end of Grade 4 due COVID-19 related school closures. The reading comprehension test consisted of 55 multiple-choice items around different types of texts, such as short stories, newspaper articles, advertisements and instruction manuals (Weekers et al., 2011). The mathematics test consisted of 96 multiple-choice and open-ended items, some of which were formulated in a real-world context whereas other items were presented as plain numerical operations (Hop et al., 2017). Results of both tests are given in the form of a continuous score as well as a proficiency level from I (highest) through V (lowest). The continuous scores were used in the analyses, whereas the proficiency levels were used to assign worksheets in the macro-adaptive condition.

### *Procedure*

Children were assigned to one of three conditions on a whole-class basis. To prevent classrooms within the same school to be assigned to the same condition, randomization took place in clusters of three classrooms. In each cluster, one classroom was randomly assigned to the control condition, one to the macro-adaptive condition, and one to the micro-adaptive condition.

Parents were informed at least two weeks before data collection started. Data collection spanned two weeks. In the first week, the teacher sent in the standardized progress monitoring data and administered the pre-test in a whole-class setting. The teacher received a testing protocol including an introduction, information on the type of procedural help that could be provided, and instructions on the classroom setup. Teachers were also instructed to check whether all questions were filled out. In the second week, the principal investigator administered the intervention on four consecutive days. The

classroom teacher was always present during the lessons, but was instructed not to support children during the inquiry sessions. On Fridays, the teacher administered the post-test in a whole- class setting.

## Results

### *Preliminary analyses*

The summary statistics in Table 1 show minimal a priori differences between the three conditions. Although children in the micro-adaptive condition scored slightly lower on all three pre-instructional measures, this difference was not statistically significant for reading comprehension,  $F(2, 150) = 1.22, p = .297$ , mathematical skilfulness,  $F(2, 150) = 0.58, p = .562$ , and the scientific reasoning pre-test,  $F(2, 150) = 1.93, p = .149$ .

By design, worksheets distribution differed per condition (see Figure 4). Children in the non-adaptive control condition received medium-support worksheets in all lessons. Figure 4 shows reasonable distribution schemes in both experimental conditions, a clear difference between conditions, and a clear difference between the lessons in the micro-adaptive condition, indicating good treatment fidelity. In comparison to the macro-adaptive condition, learners in the micro-adaptive condition more often received a low-support worksheet and less often a high-support worksheet. Furthermore, the percentage of children in the micro-adaptive condition who received a high-support worksheet dropped over the course of the lesson series, whereas the percentage of children who received a low-support worksheet increased.

Table 1

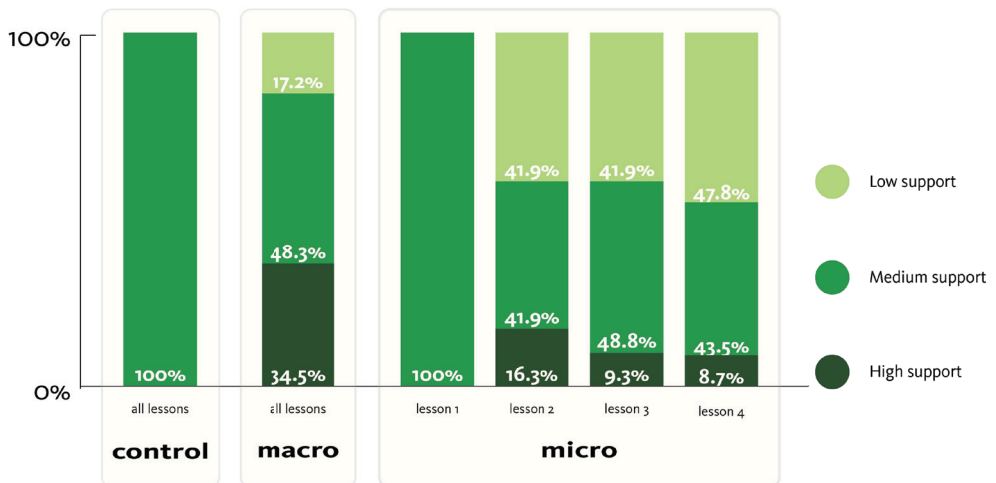
*Descriptive statistics of children's test scores*

	n	Scientific Reasoning Inventory <sup>1</sup>							
		Reading		Mathematics		Pre-test		Post-tests	
		M	SD	M	SD	M	SD	M	SD
All conditions	153	163.82	25.28	227.99	27.09	0.51	0.15	0.58	0.21
Control	49	166.12	24.97	229.45	26.09	0.53	0.16	0.59	0.24
Macro	58	165.72	26.66	229.62	27.06	0.52	0.14	0.59	0.21
Micro	46	158.96	23.65	224.39	28.40	0.48	0.15	0.56	0.19

<sup>1</sup>Scores are reported as proportion of correctly completed items

Figure 4

*Worksheet distribution*



*Learning outcomes*

A repeated measures MANOVA was used to assess children's learning from pre- to post-test on the three component scientific reasoning skills measured with the SRI; descriptive statistics can be found in Table 2. Although a main effect was found for learning over time, Wilk's  $\lambda = .80$ ,  $F(1, 150) = 37.51$ ,  $p < .001$ , learning over time did not interact with condition, Wilk's  $\lambda = .99$ ,  $F(2, 150) = 112.55$ ,  $p = .632$ . The between-subject effect of condition was not significant either,  $F(2, 150) = 0.77$ ,  $p = .467$ . Thus, neither condition was more effective than the other two.

**Table 2***Scores on the SRI per component skill*

	HEC <sup>1</sup>		Experimenting		Conclusions	
	M	SD	M	SD	M	SD
Control						
Pre-test	0.63	0.25	0.23	0.23	0.69	0.23
Post-test	0.58	0.32	0.48	0.39	0.68	0.25
Macro-adaptive						
Pre-test	0.59	0.25	0.26	0.19	0.68	0.25
Post-test	0.61	0.30	0.42	0.31	0.71	0.23
Micro-adaptive						
Pre-test	0.55	0.24	0.23	0.15	0.61	0.24
Post-test	0.57	0.25	0.42	0.34	0.69	0.24

<sup>1</sup>Hypothesis-evidence coordination

A large proportion of children in the adaptive conditions received the same support as children in the control condition, which may have concealed the effect of the adaptive support. Additional analyses were therefore carried out to compare only those children whose support was adapted to their counterparts in the non-adaptive control condition. This comparison was only made between the macro-adaptive and control condition; the high variability in support in the micro-adaptive condition (see also Figure 5) did not allow for such comparison.

**Table 3***Gain scores on the SRI of children eligible for high and low support worksheets*

	n	HEC <sup>1</sup>		Experimenting		Conclusions	
		M	SD	M	SD	M	SD
Low support							
Macro	10	0.03	0.12	0.50	0.36	-0.01	0.12
Control <sup>2</sup>	10	0.00	0.13	0.41	0.38	0.01	0.07
High support							
Macro	20	-0.03	0.19	-0.01	0.21	0.02	0.18
Control <sup>2</sup>	16	-0.12	0.23	0.11	0.29	-0.05	0.31

*Note.* Gain scores were computed as post-test – pre-test.

<sup>1</sup>Hypothesis-evidence coordination

<sup>2</sup>Subsample of children in the control condition who would have received this worksheet, had they been in the macro-adaptive condition

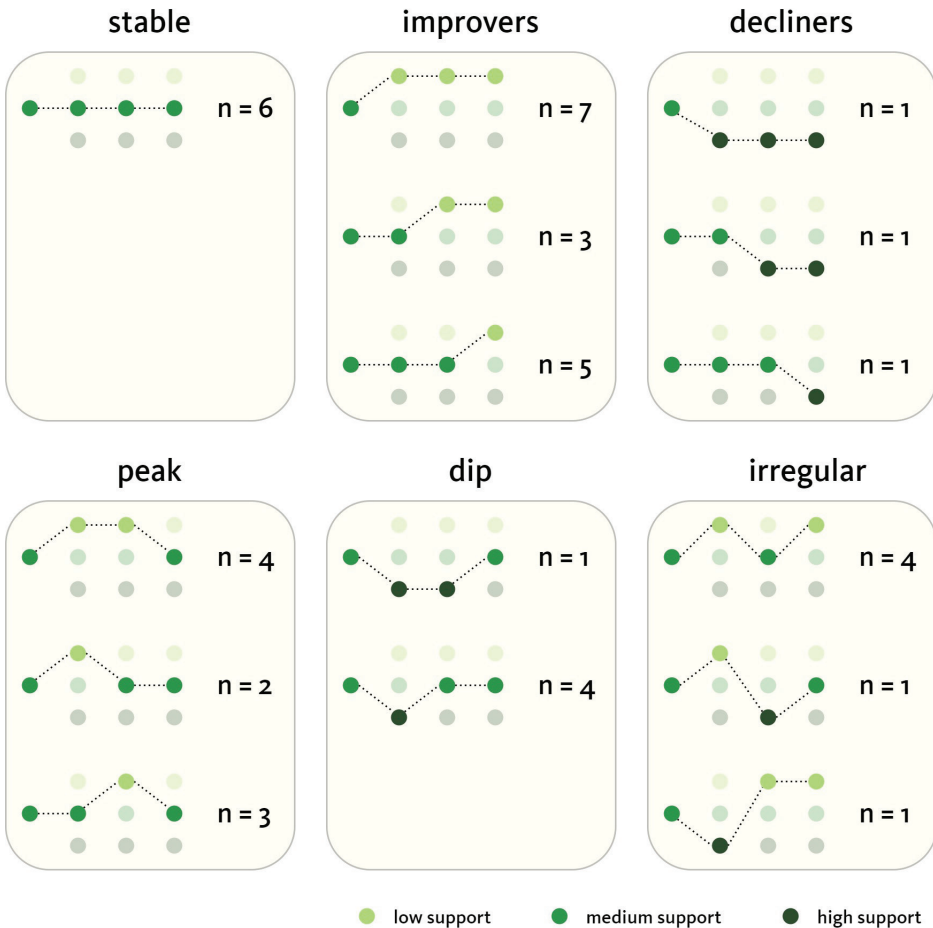
For children whose standardized test scores qualified them for either the high or low-support worksheets, the difference between pre- and post-test scores was calculated. These gain scores were then compared between the macro-adaptive and control condition (see Table 3). Mann-Whitney U tests targeting children who received low-support worksheets showed no significant differences between conditions regarding hypothesis-evidence coordination,  $z = -1.39$ ,  $p = .164$ , experimenting,  $z = -0.54$ ,  $p = .590$ , and drawing conclusions,  $z = -0.92$ ,  $p = .357$ . Similar analyses of children who received high-support worksheets yielded comparable results (hypothesis-evidence coordination:  $z = -1.20$ ,  $p = .231$ ; experimenting:  $z = -1.41$ ,  $p = .160$ ; drawing conclusions:  $z = -0.88$ ,  $p = .377$ ). Thus, the learning gains of children whose support was not adapted improved as much in scientific reasoning as children from the macro-adaptive condition who were provided with high or low support.

### *Learning processes*

With regard to learning processes, worksheet distribution patterns in the micro-adaptive condition were inspected first. Figure 5 visualizes the support levels assigned to the 44 children in the micro-adaptive condition who completed all worksheets. Fifteen patterns

emerged from the data, which were classified into six groups. The group ‘stable learners’ ( $n = 6$ ) received medium-support worksheets in all four lessons. Fifteen children could be classified as ‘improvers’ as they went from medium to low support. Only three children declined without recovery, while five children had a performance dip they managed to overcome. The opposite pattern was also found: nine children showed a temporary peak in performance, which resulted in low-support worksheets, but returned to the medium-support level in subsequent lessons. The remaining six children exhibited an irregular pattern and most of them switched back and forth between medium and low support worksheets.

Figure 5  
Worksheet assignment patterns across the four lessons in the micro-adaptive condition



Changes in worksheet distribution in the micro-adaptive condition (see Figures 4 and 5) showed a decline in high-support worksheets over the lessons, and an increase in low-support worksheets, indicating an overall improvement in task performance in this group. In order to assess whether this improvement was statistically significant and occurred in all conditions, children's scores on the worksheets were analysed by repeated measures ANOVA. There was a main effect of time, Wilk's  $\lambda = .872$ ,  $F(3, 128) = 6.29$ ,  $p = .001$ , a nonsignificant between-subject effect of condition,  $F(2, 130) = 2.41$ ,  $p = .094$ , and a significant time  $\times$  condition interaction, Wilk's  $\lambda = .90$ ,  $F(6, 256) = 2.36$ ,  $p = .031$ . The interaction effect indicated that learning over time differed across conditions, and the worksheet scores in Table 4 suggest that these differences primarily occurred in the second half of the lesson series. Pairwise comparisons between conditions on a lesson-by-lesson basis (see Table 5) confirmed this idea. Differences were most apparent between the control condition and macro-adaptive condition in lesson 3 and in lesson 4, and between the micro-adaptive and macro-adaptive conditions in lesson 3. In all these cases, children in the macro-adaptive condition had lower scores.

**Table 4**

*Descriptive statistics of the learning process data*

	n	Lesson 1		Lesson 2		Lesson 3		Lesson 4	
		M	SD	M	SD	M	SD	M	SD
Control	45	7.04	2.18	8.53	2.09	8.36	2.25	8.33	2.06
Macro-adaptive	47	7.26	2.97	7.68	2.58	6.89	2.68	6.87	3.10
Micro-adaptive	41	7.37	2.65	8.00	2.06	8.12	1.99	7.85	2.29

**Table 5***Pairwise comparisons of the learning process data*

	Control vs. Macro-adaptive			Control vs. Micro-adaptive			Macro-adaptive vs. Micro adaptive		
	$\Delta_M$	SE	p	$\Delta_M$	SE	p	$\Delta_M$	SE	p
Lesson 1	-0.21	.55	.973	-0.32	.57	.922	-0.11	.56	.996
Lesson 2	0.85	.47	.205	0.53	.49	.623	-0.32	.48	.883
Lesson 3	1.46	.49	.010	0.23	.51	.955	-1.23	.50	.045
Lesson 4	1.46	.53	.020	0.48	.55	.765	-0.98	.54	.202

*Note.* The Sidak adjustment for multiple comparison was applied to correct  $p$ -values.

## Discussion

This study investigated adaptivity in scientific reasoning instruction, an area that to our knowledge has not been studied extensively before. Learning outcomes and learning processes were studied in two adaptive conditions and a control condition. Although it was hypothesized that children in the adaptive conditions would have higher learning outcomes, no differences were found at the group level. Additional analyses showed no differences in learning outcomes for the subgroup of children from the macro-adaptive condition who received low or high-support worksheets, compared to their counterparts in the control condition who qualified for the same amount of support had they been in the macro-adaptive condition. With regard to learning processes, the worksheet distribution in the micro-adaptive condition showed several patterns of improvement over time. Contrary to expectations, task performance in both adaptive conditions did not exceed that in the control condition. Thus, although previous research on adaptive education did show



improved learning (Aleven et al., 2016; Deunk et al., 2018), the current study showed a marginal effect on learning processes and an overall lack of effect on learning outcomes. Critics might use these findings to argue that adaptivity is of no avail for scientific reasoning – but since adaptive scientific reasoning instruction is clearly understudied, it is worthwhile to explore other explanations as well. One way in which this study distinguishes itself from other adaptivity studies is its unplugged, hands-on approach. Adaptive education is usually organized through, or at least accompanied by, computer algorithms that interpret student data and provide either personalised support to children or progress information in the form of teacher dashboards (Aleven et al., 2016; Shute & Towle, 2003). Although our approach aligns with everyday classroom practice, especially for primary school science where working with physical materials is still highly prevalent (Evangelou & Kotsis, 2018), it has to our knowledge not been attempted before. When unplugged adaptivity is applied, it often relies on within-class grouping or assignment of learning materials based on general cognitive learner characteristics, similar to our macro-adaptive condition. Studies on such unplugged adaptivity found smaller effects than studies using computer-supported adaptivity (Deunk et al., 2018). This suggests that the sensitivity and precision of the adaptivity matters, and might explain the limited effects found in the current study.

Both adaptivity mechanisms used in this study relied on relatively little information, both in terms of learner variables and measurement frequency. Although this made the mechanisms easy to apply, they may not have been sensitive enough. For the macro-adaptive condition, learner variables were limited to two predictors of scientific reasoning. A relatively easy addition to these learner variables would be children's current proficiency in scientific reasoning, as measured by the pre-test. Although the micro-adaptive condition relied on a finer-grained adaptivity mechanism with four learner variables (i.e., the component skills of scientific reasoning), children's performance was only assessed between lessons and scores for component skills were taken together in order to manage the principal investigator's workload, possibly resulting in a too coarse adaptivity mechanism as well. Thus, the adaptive conditions could be improved by increasing the number of learner variables and the number of measurement points.

An alternative explanation for the limited effects in this study points in the opposite direction: a finer grained adaptivity mechanism is susceptible to more noise. In this study, such noise could have occurred in the micro-adaptive condition, where children's answers to open-ended questions and assignments were used as learner variables to set the support level for the next lesson. On such open assignments, children often do not show their full potential (Salac & Franklin, 2020). Although we attempted to limit the influence of these

false negatives by using the highest worksheet score of each lesson, the large number of patterns found in the micro-adaptive condition could point to an overly sensitive adaptivity mechanism.

It is in this context also important to note that for most children in the micro-adaptive condition, support was decreased over time. Despite this reduction, their worksheet scores were similar to those of children in the control group. It is possible that increased learning in the micro-adaptive condition was offset by the decreased support. As this explanation could not be tested here, future research should examine whether the fading of support curbed learning progress in the micro-adaptive condition. Such studies could use the current micro-adaptive data to determine the rate of fading, and apply the same rate to a new control condition in order to straighten this out.

In addition to learner variables and measurement characteristics, Plass and Pawar (2020) defined a third dimension of adaptivity: the adjustments made to learning materials or teaching strategies. On this dimension, both adaptive conditions were similar in that all component skills were adapted at once. Although this made the adaptivity mechanisms user-friendly, it ignored the fact that not all children are equally proficient in each component skill. This is particularly important for intermediate scientific reasoners, who can be adept at some but not all component skills (Schlatter et al., 2021 [Chapter 4]). Making more specific adjustments to the learning materials could result in higher learning outcomes for this group. However, a more precise adaptivity mechanism is needed to make these more specific adjustments – which was deemed unfeasible in the unplugged approach used in the current study.

An obvious solution would be to move away from working with physical materials when attempting to implement fine-grained adaptivity, and other studies indeed found that intelligent tutoring systems and other forms of computer-supported adaptivity were particularly effective (Deunk et al., 2018). However, such computer-based approaches cannot readily be implemented in primary science classrooms where physical materials prevail (Evangelou & Kotsis, 2018). A hybrid form, where assignments are presented on a tablet or laptop but the experiments are performed with physical materials, could offer more accurate adaptivity (Aleven et al., 2016) while maintaining the hands-on nature of science education. The design of such a hybrid approach to science education could draw upon extant literature on mobile learning (Crompton et al., 2017) and be even more versatile than the usual computer-supported science instruction. Usually, computer-supported adaptive science education involves simulations that might need to be made or at least

adjusted for each experiment (e.g., De Jong et al., 2021). In a hybrid system, it would suffice to adjust the information on the variables, thus making it easy to use for teachers.

### *Strengths and limitations*

The manner in which adaptivity was implemented in the current study, namely through the worksheets, was a strength with respect to experimental rigor but a limitation with respect to the effectiveness of the teaching materials. Although the static support given on the worksheets would in regular classroom situations have been combined with other strategies such as extended instruction or ad-hoc support (Martin et al., 2019), the use of multiple strategies would have jeopardized the study's experimental validity. Another limitation was that children who received most support had to read the most. Because reading comprehension and scientific reasoning are closely related (e.g., Van de Sande et al., 2019) this could have limited the effectiveness of the support, and although the principal investigator and classroom teacher were allowed to read out the instruction given on the worksheets if asked for help, they could not do so for each child in each instance. This intricate balance between experimental validity and effective education is a challenge for every education researcher, and we believe more than one type of study is needed: our experimentally sound classroom study could and should be complemented by both more experimental lab studies and ecologically valid classroom studies, which together paint a complete picture.

### *Conclusion*

Although children in all three conditions improved their scientific reasoning, the current study shows little effect of micro- or macro-adaptivity on the learning process and no effect on learning outcomes in scientific reasoning. Still, the limited learning gains for groups of children indicate a clear need for support. Further research should concentrate on more fine-grained adaptivity and non-written support in order to ultimately develop guidelines that help teachers support children and adapt their teaching of scientific reasoning.

## References

- Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016). Instruction based on adaptive learning technologies. In R. E. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 522-560). Routledge.
- Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*(1), 1-18. <https://doi.org/10.1037/a0021017>
- Bell, R. L., Smetana, L., & Binns, I. (2005). Simplifying inquiry instruction. *The Science Teacher, 72*(7), 30-33.
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098-1120. <https://doi.org/10.1111/1467-8624.00081>
- Crompton, H., Burke, D., & Gregory, K. H. (2017). The use of mobile learning in PK-12 education: A systematic review. *Computers & Education, 110*, 51-63. <https://doi.org/10.1016/j.compedu.2017.03.013>
- De Jong, T., Gillet, D., Rodriguez-Triana, M. J., Hovardas, T., Dikke, D., Doran, R., Dziabenko, O., Koslowsky, J., Korventausta, M., Law, E., Pedaste, M., Tasiopoulou, E., Vidal, G., & Zacharia, Z. C. (2021). Understanding teacher design practices for digital inquiry-based science learning: the case of Go-Lab. *Educational Technology Research and Development, 69*, 417-444. <https://doi.org/10.1007/s11423-020-09904-z>
- Deunk, M. I., Smale-Jacobse, A. E., De Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education. *Educational Research Review, 24*, 31-54. <https://doi.org/10.1016/j.edurev.2018.02.002>
- Evangelou, F., & Kotsis, K. (2018). Real vs virtual physics experiments: Comparison of learning outcomes among fifth grade primary school students. A case on the concept of frictional force. *International Journal of Science Education, 41*(3), 330-348. <https://doi.org/10.1080/09500693.2018.1549760>
- Eysink, T. H. S., Hulsbeek, M., & Gijlers, H. (2017). Supporting primary school teachers in differentiating in the regular classroom. *Teaching and Teacher Education, 66*, 107-116. <https://doi.org/10.1016/j.tate.2017.04.002>
- Hop, M., Janssen, J., & Engelen, R. (2017). *Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 7*. [Scientific justification arithmetics and mathematics 3.0 for grade 5]. Cito.
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1-48. [https://doi.org/10.1207/s15516709cog1201\\_1](https://doi.org/10.1207/s15516709cog1201_1)
- Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development, 86*(1), 327-336. <https://doi.org/10.1111/cdev.12298>
- 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.org/10.1080/15248372.2019.1620232>
- Köksal-Tuncer, Ö., & Sodian, B. (2018). The development of scientific reasoning: Hypothesis testing and argumentation from evidence in young children. *Cognitive Development, 48*, 135-145. <https://doi.org/10.1016/j.cogdev.2018.06.011>
- Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science, 16*(11), 866-870. <https://doi.org/10.1111/j.1467-9280.2005.01628.x>
- Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning. *Review of Educational Research, 86*(3), 681-718. <https://doi.org/10.3102/0034654315627366>

- 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>
- Lorch, R. F., Lorch, E. P., Freer, B., Calderhead, W. J., Dunlap, E., Reeder, E. C., Van Neste, J., & Chen, H. T. (2017). Very long-term retention of the control of variables strategy following a brief intervention. *Contemporary Educational Psychology, 51*, 391-403. <https://doi.org/10.1016/j.cedpsych.2017.09.005>
- Lorch, R. F., Lorch, E. P., Freer, B. D., Dunlap, E. E., Hodell, E. C., & Calderhead, W. J. (2014). Using valid and invalid experimental designs to teach the control of variables strategy in higher and lower achieving classrooms. *Journal of Educational Psychology, 106*(1), 18-35. <https://doi.org/10.1037/a0034375>
- Lorch, R. F., Lorch, E. P., Wheeler, S. L., Freer, B. D., Dunlap, E., Reeder, E. C., Calderhead, W., Van Neste, J., & Chen, H.-T. (2019). Oversimplifying teaching of the control of variables strategy. *Psicología Educativa, 26*(1), 7-16. <https://doi.org/10.5093/psed2019a13>
- Martin, N. D., Dornfeld Tissenbaum, C., Gnesdilow, D., & Puntambekar, S. (2019). Fading distributed scaffolds: The importance of complementarity between teacher and material scaffolds. *Instructional Science, 47*(1), 69-98. <https://doi.org/10.1007/s11251-018-9474-0>
- Mastropieri, M. A., Scruggs, T. E., Norland, J. J., Berkeley, S., McDuffie, K., Halloran Tonquist, E., & Connors, N. (2006). Differentiated curriculum enhancement in inclusive middle school science: effects on classroom and high-stakes tests. *Journal Of Special Education, 40*(3), 130-137. <https://doi.org/10.1177/00224669060400030101>
- 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>
- McCrea Simpkins, P., Mastropieri, M. A., & Scruggs, T. E. (2009). Differentiated curriculum enhancements in inclusive fifth-grade science classes. *Remedial and Special Education, 30*(5), 300-308. <https://doi.org/10.1177/0741932508321011>
- Piekny, J., Grube, D., & Maehler, C. (2014). The development of experimentation and evidence evaluation skills at preschool age. *International Journal of Science Education, 36*(2), 334-354. <https://doi.org/10.1080/09500693.2013.776192>
- Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology, 31*(2), 153-179. <https://doi.org/10.1111/j.2044-835X.2012.02082.x>
- Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education, 52*(3), 275-300. <https://doi.org/10.1080/15391523.2020.1719943>
- Salac, J., & Franklin, D. (2020, June 15–19). If they build it, will they understand it? Exploring the relationship between student code and performance. ITiCSE '20, Trondheim, Norway.
- Schlatter, E., Molenaar, I., & Lazonder, A. W. (2020). Individual differences in children's development of scientific reasoning through inquiry-based instruction: Who needs additional guidance? . *Frontiers in Psychology, 11*, Article 904. <https://doi.org/10.3389/fpsyg.2020.00904>
- Schlatter, E., Molenaar, I., & Lazonder, A. W. (2021). Learning scientific reasoning: A latent transition analysis. *Learning and Individual Differences, 92*, Article 102043. <https://doi.org/10.1016/j.lindif.2021.102043>
- Schneider, M., & Hardy, I. (2013). Profiles of inconsistent knowledge in children's pathways of conceptual change. *Developmental Psychology, 49*(9), 1639-1649. <https://doi.org/10.1037/a0030976>

- Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review, 39*, 37-63. <https://doi.org/10.1016/j.dr.2015.12.001>
- 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*, Article 101923. <https://doi.org/10.1016/j.cedpsych.2020.101923>
- Shute, V., & Towle, B. (2003). Adaptive E-Learning. *Educational Psychologist, 38*(2), 105-114. [https://doi.org/10.1207/s15326985ep3802\\_5](https://doi.org/10.1207/s15326985ep3802_5)
- Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., Conover, L. A., & Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted, 27*(2), 119-145. <https://doi.org/10.1177/016235320302700203>
- Van de Sande, E., Kleemans, M., Verhoeven, L., & Segers, E. (2019). The linguistic nature of children's scientific reasoning. *Learning and Instruction, 62*(1), 20-26. <https://doi.org/10.1016/j.learninstruc.2019.02.002>
- Van der Graaf, J. (2020). Inquiry-based learning and conceptual change in balance beam understanding. *Frontiers in Psychology, 11*, Article 1621. <https://doi.org/10.3389/fpsyg.2020.01621>
- Van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instructional Science, 43*(3), 381-400. <https://doi.org/10.1007/s11251-015-9344-y>
- Vogt, F., & Rogalla, M. (2009). Developing adaptive teaching competency through coaching. *Teaching and Teacher Education, 25*(8), 1051-1060. <https://doi.org/10.1016/j.tate.2009.04.002>
- Weekers, A., Groenen, I., Kleintjes, F., & Feenstra, H. (2011). *Wetenschappelijke verantwoording papieren toetsen begrijpend lezen voor groep 7 en 8* [Scientific validation paper-and-pencil tests reading comprehension for grade 5 and 6]. Cito.
- 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>







# Chapter 6

## **General discussion**

## Introduction

Scientific reasoning is an important skill that helps people function in modern knowledge societies (Trilling & Fadel, 2009). As a result, it has found its way into school curricula around the globe (Abd-El-Khalick et al., 2004), including the Netherlands (Greven & Letschert, 2006; Techniekpact, 2013). Scientific reasoning comprises multiple component skills, which are broadly categorized as hypothesizing, experimenting and evaluating outcomes (Klahr & Dunbar, 1988; Pedaste et al., 2015). Because of this multidimensionality, scientific reasoning can be a challenging topic to teach, in particular in primary schools where classmates differ considerably in terms of cognitive capacities. Therefore, this thesis aimed to answer two main research questions:

- (1) How can differences in upper primary school children's scientific reasoning be characterized and predicted?
- (2) How can these differences be addressed in upper primary classrooms?

These research questions were studied at 11 schools throughout the Netherlands. This final chapter discusses the main outcomes of the four studies presented in this thesis in light of the research questions mentioned above. Next, the limitations of the presented research are considered and implications for both theory and practice are given. Alongside the limitations and implications, some suggestions for future research are proposed.

## Characterizing and predicting differences in scientific reasoning

The first aim of the research described in this dissertation was to characterize and predict differences in children's scientific reasoning. Previous research points to considerable variation in scientific reasoning proficiency and learning, both between children and across skills. Therefore, three different types of assessments and a variety of analysis techniques were used to characterize and predict differences in scientific reasoning.

Considering the multidimensionality of scientific reasoning, the relative difficulty of the component skills will be discussed first. Previous research suggests that not all component skills are equally difficult to perform and learn, and a tentative rank order from least difficult (experimenting) to most difficult (skills associated with evaluating outcomes, such as drawing conclusions) could be established from this earlier work (Zimmerman, 2007). However, the circumstances in which this order was established were not ideal: many studies assessed either a single skill (often experimenting; c.f. Koerber & Osterhaus,

2019), or multiple skills through separate tests or questions (e.g., Piekny & Maehler, 2013), often using paper-and-pencil items (Opitz et al., 2017).

Furthermore, scientific reasoning is not only multidimensional but also an inherently whole-task endeavour: one needs carefully formulated research questions and hypotheses in order to design an informative experiment and collect meaningful data, which in turn require accurate interpretation to come to valid conclusions (Lazonder & Janssen, 2021). Therefore, it is important to assess the relative difficulty of the component scientific reasoning skills in a whole-task setting that resembles the learning context and mimics authentic scientific practice.

The study reported in Chapter 2 therefore employed a performance-based test that enabled children to complete four head-to-tail inquiry cycles of increasing difficulty. Using multivariate multiple regression analysis, this study confirmed the rank order established in earlier research: experimenting was found to be the least difficult skill for children to perform, followed by hypothesizing, interpreting data, drawing conclusions and finally evaluating data characteristics as the most difficult skill.

The findings from Chapter 2, which were limited to a single time point, were supplemented by analyses of the learning of scientific reasoning over time in Chapters 3 through 5. Although the aim of these chapters was not to rank the component skills from least to most difficult, the learning of most component skills mirrored the general pattern found in Chapter 2. The one exception was the component skill ‘experimenting’, which had particularly low pre-test scores – on average below chance level. This result contradicts the conclusions of most international studies (e.g., Zimmerman, 2007) but matches the outcomes of a recent national benchmarking study (Inspectie van het Onderwijs, 2017), indicating that Dutch children are somehow less proficient in experimenting than their peers in other countries. For this reason, and because an interaction effect between learning over time and component skill indicated that experimenting was learned very easily by a small subset of children, it seems unlikely that the low pre-test scores were entirely attributable to differences in test modality. The small group of children who did learn to experiment during the lesson series scored remarkably high on the post-test – between 90 and 100% correct. Thus, although experimenting clearly does not develop naturally in all children, once learned it is relatively easy to apply (Chen & Klahr, 1999; Lorch et al., 2017; Peteranderl, 2019).

The finding that, indeed, some scientific reasoning skills are more difficult than others can help determine which component skills require more attention in educational research and practice. Still, as primary school classrooms are inherently diverse, this general

pattern may not apply to all children in a classroom. Therefore, Chapter 4 employed person-centred analysis of the worksheets children filled out during the lessons. Through these analyses, subgroups of children were distinguished based on their similarity to one another as well as their difference from children in other subgroups. Several proficiency profiles were found, confirming that not all component skills are equally difficult for all children.

One of the proficiency profiles aligned with the pattern found in Chapter 2. Children in this profile, which was labelled 'experimenters', scored low on drawing conclusions and high on experimenting. Another profile was more in line with the pattern found in Chapter 3. Children in the 'theorists' profile were better at drawing conclusions than at experimenting. Two more homogenous profiles were found as well: low achievers and high achievers. A subsequent latent transition analysis suggested that the experimenter and theorist profiles might both be intermediate stages between overall low performance and overall high performance, suggesting that there is more than one route to learning scientific reasoning.

The more traditional variable-centred analyses in Chapters 2 and 3 sought to explain these individual differences by three cognitive learner characteristics: reading comprehension, numerical ability or mathematical skilfulness, and problem solving skill. While the former two are part of the progress monitoring tests used in most Dutch schools, problem solving is not. As previous studies suggested that problem solving might account for differences in children's scientific reasoning (e.g., Mayer et al., 2014), it was used as a predictor in Chapter 2 – where it did not explain differences in scientific reasoning. As discussed in Chapter 2, this result was possibly due to a lack of congruence with the nature of scientific problem solving (Jonassen, 2000): scientific problems are open-ended and ill-defined, whereas the solution to the Tower of Hanoi task is known beforehand and can be reached by applying a constrained set of rules. As the latter is a much simpler form of problem solving, the Tower of Hanoi may not have been sensitive enough to distinguish differences in children's problem solving ability. Therefore, and because it was not part of progress monitoring in schools, problem solving was dropped as a predictor in subsequent studies.

Reading comprehension, on the other hand, had already been established as a robust predictor of scientific reasoning proficiency in paper-and-pencil test settings (Koerber, Mayer, et al., 2015; Mayer et al., 2014; Siler et al., 2010; Van de Sande et al., 2019). It was therefore used as a predictor in Chapters 2 and 3. Because the paper-and-pencil tests used in previous research present a possible confound with reading comprehension, Chapter 2

sought to establish this relation using a performance-based test of scientific reasoning. The results show that reading comprehension still predicts scientific reasoning if reading is not a requirement of the test. This suggests that a more general reasoning ability underlies both reading and scientific reasoning. Indeed, relations have been established between scientific reasoning and verbal reasoning (Siler et al., 2010) as well as nonverbal reasoning (Van de Sande et al., 2019). In Chapter 3 it was found that reading comprehension not only predicts *proficiency in*, but also the *learning of* scientific reasoning over a five week lesson series. However, as this study used a paper-and-pencil test to assess scientific reasoning, this finding might be subject to the confound discussed above.

In contrast to reading comprehension, numerical ability has not been at the forefront in research attempting to explain scientific reasoning, which is particularly remarkable because scientific reasoning does involve reasoning about numerical outcomes (Krummenauer & Kuntze, 2019; Makar et al., 2011; Mayer et al., 2014; Piekny & Maehler, 2013). The relatively short task used in Chapter 2 measured basic numerical skills (addition, subtraction, multiplication and division), which correlated with no other component skill than ‘evaluating data characteristics’ and did not predict any component skill. Chapter 3 used a more elaborate mathematics measure that involved more complex, often contextualized problems. This measure did predict children’s learning of experimenting over the five-week lesson series, as well as their proficiency in hypothesis-evidence coordination. These findings suggest that while basic numerical skills do not play a role in children’s scientific reasoning, more advanced mathematical skills do.

Through the use of both variable-centred and person-centred analyses, Chapters 2 through 4 produced a rich and comprehensive picture of upper primary school children’s scientific reasoning. The studies underline the importance of unraveling scientific reasoning in its component skills and demonstrate once again that same-age children can differ considerably in their proficiency in and learning of scientific reasoning skills.

### **Addressing differences in scientific reasoning in the classroom**

The second aim of the research described in this dissertation was to address the differences in scientific reasoning described in Chapters 2 through 4 in upper primary science classrooms. The outcomes of these studies informed the development of an intervention that delivered adaptive scientific reasoning support. Adaptive support, in short, uses information about learners, also known as learner variables, to adjust teaching materials or instructional strategies (Aleven et al., 2016; Plass & Pawar, 2020) and is generally a very

effective means to improve learning outcomes (Deunk et al., 2018).

Although most adaptivity research relies on computers for fast and accurate adaptation of support (Aleven et al., 2016), primary school science is often taught in non-digital, or unplugged, settings (Evangelou & Kotsis, 2018). To accommodate this practice, the adaptivity mechanisms and support materials used in Chapter 5 were designed to be used without a computer. Therefore, the adaptivity mechanisms were based on a few simple rules that could be applied by hand after each lesson. These mechanisms were used to assign children to low-, medium- or high-support worksheets, either at the beginning of the lesson series based on their scores on standardized progress monitoring tests of reading comprehension and mathematics (macro-adaptive condition) or at the start of each lesson based on their worksheet score of the previous lesson (micro-adaptive condition). In addition to the two adaptive conditions, a control condition in which all children received an intermediate level of support was also included in the study described in Chapter 5.

Similar to the outcomes of Chapter 3 and 4, children in all conditions improved their scientific reasoning over time. However, the adaptive support did not lead to better learning outcomes compared to the control condition, and performance of children in the macro-adaptive condition even showed a small decline. As adaptivity has been found effective in many domains (Deunk et al., 2018), including the learning of science content knowledge (e.g., McCrea Simpkins et al., 2009), it is unlikely that it is by definition ineffective in the realm of scientific reasoning.

Rather, it is likely that the unplugged approach – regardless of its careful design – limited the effectiveness of both the micro-adaptive and macro-adaptive condition. In order to make adaptivity work in an unplugged setting, some precision was sacrificed: additional learner variables such as prior content knowledge (as measured by a pre-test) were omitted and in the micro-adaptive condition, scores for the component skills were aggregated to make the mechanism easy to apply – even though Chapter 4 showed intermediate scientific reasoners are often either proficient in experimenting or in drawing conclusions. Thus, treating component scientific reasoning skills as separate learner variables might be one way to fine-tune the adaptivity mechanisms.

A further disadvantage of the unplugged approach was that adaptive support was solely delivered via the worksheets; in order to maintain experimental validity, no extended instruction or ad-hoc guidance was provided. Although the worksheet-based support was designed to be as language-lean as possible, reading load was highest for children who qualified for the highest support level. The first author and the classroom teacher mitigated this glitch by reading out the instructions on the worksheets, but this was not possible for

each child in every instance. Considering the close relation between reading comprehension and scientific reasoning found in earlier chapters of this thesis, this written delivery mode could have limited the effectiveness of the adaptive support.

Thus, although the unplugged approach to adaptivity was a key feature in Chapter 5, it is possible that inherent limitations to the adaptivity mechanisms and the support delivery mode reduced the effectiveness of the adaptive teaching materials. Rather than taking this outcome as a justification to abandon physical investigation of science materials in favour of on-screen science education, the findings reported in this thesis should be used to explore different avenues for adaptivity without radically changing classroom settings.

One such avenue could be the hybrid science instruction described in the discussion of Chapter 5, which combines physical science material with adaptive computerized scaffolding. Recent research suggests that children learn more in such hybrid environments than when learning on-screen only (Yannier et al., 2020). Furthermore, interactive and adaptive guidance offered by such hybrid systems yields higher learning outcomes than unsupported discovery. Yannier et al. (2020) studied children's learning when using a science station consisting of a large screen, a Kinect motion detector and an earthquake simulation table. Dutch schools, on the other hand, often have a large number of tablets available, making the context suitable for smaller-scale solutions such as mobile adaptive learning environments. This is particularly interesting because mobile devices can be used in a variety of environments, and their size and portability make them suitable to supplement physical science materials in a regular classroom.

## Limitations

As in all education research, choices had to be made to strike an optimal balance between experimental and ecological validity while maintaining practical feasibility. These decisions inevitably incurred some limitations for the studies presented in this dissertation, most of which concerned a single study and were therefore addressed in the discussion sections of their respective chapter. Limitations concerning this thesis as a whole are discussed below.

The studies presented in this thesis focus on scientific reasoning, an important aspect of the broader skillset involved in doing a scientific investigation. In particular, the interplay of the component scientific reasoning skills was studied. This left little room to also consider other aspects of scientific thinking, such as understanding of the nature of science and science content knowledge. Because these aspects have been found to interact with children's scientific reasoning (Koerber, Osterhaus, et al., 2015; Osterhaus et al., 2017;

Schalk et al., 2019; Schwichow et al., 2020), future research should take these aspects into account.

An important methodological asset of this dissertation is the choice of instruments, which all measured multiple components of scientific reasoning. Still, these instruments had some limitations of their own. Although the performance-based test used in Chapter 2 did not require reading, it was labour-intensive to administer and therefore consisted of only 15 items (3 per component skill). These items had good content validity and acceptable internal consistency, but convergent validity was difficult to assess (Lazonder & Janssen, 2021). Furthermore, because this test was administered individually, it was practically unfeasible to use in Chapters 3 through 5.

The studies reported in Chapters 3 and 5 used the Scientific Reasoning Inventory, a paper-and-pencil test that can be administered on a whole-class basis (Van de Sande et al., 2019). Like the performance-based test used in Chapter 2 and the worksheets analysed in Chapters 3 through 5, the Scientific Reasoning Inventory distinguished several component skills. Still, it was difficult to compare the outcomes of both tests because the measured skills did not entirely match, both in name and in the way they were operationalised. For example, the data interpretation items on the Scientific Reasoning Inventory featured unambiguous, dichotomous outcomes, while the other instruments used in this thesis required interpretation of continuous and sometimes messy data. Because the type of data greatly influences the ease with which results are interpreted (Kanari & Millar, 2004; Piekny & Maehler, 2013), it is likely that the data interpretation items on the Scientific Reasoning Inventory were easier than those on the performance-based test and the worksheets. With regard to drawing conclusions, where the performance-based test used data related to earlier items, the items on the Scientific Reasoning Inventory were relatively abstract and detached from the data sets provided in the other items. Lastly, hypothesizing and interpreting data loaded on the same scale in the paper-and-pencil test (Van de Sande et al., 2019) – a methodological challenge that could not be overcome. Together, these differences in operationalisation made it difficult to make one-on-one comparisons across instruments.

Another methodological challenge was to safeguard both experimental and ecological validity in the classroom intervention studies presented in Chapters 3 through 5. One means to simulate a regular classroom situation is to spread out the lesson series over multiple weeks, with one lesson being given every week – as was done in the study presented in Chapters 3 and 4. Data collection for Chapter 5, however, took place during the coronavirus pandemic. Although it was possible to collect data in classrooms in the fall



of 2020, it was decided to condense the lesson series into a single week to minimize the risk of data loss caused by a larger-than-usual number of children calling in sick, drop out of entire classrooms due to confirmed cases, or overall termination of data collection due to universal school closure. Practically, this meant teachers administered the pre-test in the week preceding the intervention, and the lessons were taught on four consecutive days – usually from Monday through Thursday to leave room for unexpected events during the week. At the end of the lesson week, the post-test was administered by the teachers.

Because of these implementation differences, the outcomes of Chapter 5 may not compare well with those of Chapters 3 and 4, even though the basis of the lesson series remained the same. And although intensive courses in higher education appear to yield similar or better learning outcomes on immediate and delayed tests compared to semester-long courses (Anastasi, 2007; Deichert et al., 2015), it is unknown how condensing a lesson series affects learning in primary school.

Finally, this thesis studied diversity in scientific reasoning using context-rich tasks. Specifically, children worked on physics experiments during all lessons and most test items addressed physics inquiries. Although the inquiries involved different physics topics, it is unknown whether the learned scientific reasoning skills transferred to other areas of science – or, in other words, whether children acquired domain-general skills during the lessons. Research has shown that basic-level skills can transfer to other subject-specific topics (e.g., Masnick et al., 2017; Wagenveld et al., 2014). Yet different areas of science require different research methods: in biology, for example, investigations often involve observation and classification instead of experimentation (Osborne, 2018). These differences point towards the existence of domain-specific modes of scientific reasoning (Chinn & Golan Duncan, 2018; Schauble, 2018). As the current dissertation focused on physics only, it cannot be determined whether the skills learned by children are domain-general or domain-specific, thus limiting the findings of this dissertation to a single area of science.

## Implications

### *Theoretical implications*

The large diversity found at the component skill level confirms that scientific reasoning is a multidimensional construct, and reinforces the importance of studying scientific reasoning in a whole-task setting. Yet, this does not mean that all component skills should receive the same amount of attention in future research. In past research, experimenting has been

overrepresented (Koerber & Osterhaus, 2019; Rönnebeck et al., 2016; Zacharia et al., 2015), and effective interventions were developed for teaching experimenting skills (e.g., Chen & Klahr, 1999; Lorch et al., 2014). However, the studies presented in this dissertation confirmed that experimenting is not necessarily the hardest skill to learn. Skills that have received less attention in past research, hypothesizing and evaluating outcomes, were found to be particularly difficult and should thus receive more attention in future research

A few studies have provided insight in the development of and pitfalls in evaluating outcomes (Kanari & Millar, 2004; Masnick et al., 2017; Masnick & Morris, 2008; Piekny & Maehler, 2013), and some work has been done to improve children's ability to interpret existing data in the field of statistical literacy (Ben-Zvi & Garfield, 2004; Makar, 2016). However, little has been done to develop interventions to improve children's ability to evaluate evidence they gathered themselves. Promising work by Grimm et al. (2021) uses structuring scaffolds to help children focus their attention to the information at hand, and problematizing scaffolds to increase awareness of certain data aspects that might be at odds with their intuitive interpretation of the data. These scaffolds decrease the gap between children with varying levels of inhibition, prior knowledge, and logical reasoning. Further development of such interventions, to go alongside existing interventions aimed at experimenting, has the potential to bring scientific reasoning instruction to a new level.

The multidimensionality of scientific reasoning has implications for the instruments and analysis techniques used in educational research as well. First of all, it is crucial that both instruments and analyses distinguish the component skills of scientific reasoning. Chapter 4 has shown the power of person-centred analysis for disentangling individual differences in scientific reasoning during a classroom intervention, complementing the variable-centred analysis of pre- and post-test data in Chapter 3. The use of these person-centred analyses is relatively new in scientific reasoning research (Edelsbrunner & Dablander, 2018), and when applied to a range of measurements, such as whole-class or standardized tests, could be the next step in understanding scientific reasoning. Furthermore, when these measurements are taken over a longer period of time (e.g., multiple years) person-centred analysis could give insight in the long-term development of scientific reasoning proficiency profiles as well.

In order to capture the short-term development of scientific reasoning, the study in Chapter 4 made use of task performance measurements. Task performance or process measures make it possible to look closer at the development of scientific reasoning in a classroom setting, and can therefore be a valuable complement to whole-class testing.

Although such performance measurements are frequently used in adaptive teaching materials (Alevan et al., 2016; Plass & Pawar, 2020) and can be used by teachers themselves as formative evaluation (Kruit et al., 2020), the use of paper-and-pencil task performance measurements in research is less common. This is at least partially due to the tension that arises when developing assignments that should both be reliable as an assessment and valuable as a learning opportunity. Additionally, performance measures with an open character (like the ones used in this thesis) might not always motivate children to show their full potential, leading to many ‘false negative’ scores (Salac & Franklin, 2020). Thus, although performance measurements can and should be used as an added source of information in research, they should be developed carefully and always be accompanied by more formal measurements.

Chapters 2 through 4 contribute to the acknowledgement of diversity in learning of scientific reasoning and thereby justify the need for adaptivity in teaching scientific reasoning skills. Although the entirely unplugged approach tested in Chapter 5 did not yield the expected results, it did provide important information on and practical experience with the implementation possibilities of adaptivity in learning situations where physical materials are used. In future research, the development of adaptive learning environments that align with the regular classroom situation should be considered, and theory on the use of adaptivity in different types of classroom situations should be further developed.

### *Practical implications*

The research presented in this dissertation shows great diversity in children’s learning of scientific reasoning. Some skills prove more difficult to learn and perform than others, while differences also exist between children. This result has important implications for classroom practice. First, although the modest overall learning gains implicate that more structural attention for scientific reasoning throughout the year would benefit all children, the differences between children suggest that adaptation of support is warranted as well. Consistent monitoring and adaptation of support to a child’s current proficiency level tends to be slightly more effective in this regard than ability grouping based on other cognitive learner characteristics such as reading comprehension and mathematical skilfulness.

A second implication arises from the finding that the component skills are not equally difficult to learn, so more teaching and practice time should be allocated to more difficult skills. A skill that is particularly difficult for many children is interpreting data – although its difficulty can depend on the characteristics of the data set at hand (Masnick & Morris, 2008; Piekny & Maehler, 2013). Therefore, it is not only important to allocate

ample time to this particular skill, but also that teachers select data (or experiments that generate data) that match the entry levels of the children in their classrooms. Experimenting appears to be a skill that, once it has been learned, is relatively easy to apply. Because experimenting is fundamental to scientific reasoning (Schwchow et al., 2016) and errors made in the experimenting phase propagate throughout the rest of an inquiry (Lazonder & Wiskerke-Drost, 2015), it is important to teach it well – and Chapter 3 shows that repeated practice is not effective for most children. Previous research has led to short interventions with proven long-term effects on experimenting abilities of most children (Chen & Klahr, 1999; Lorch et al., 2017). If such interventions, which involve demonstration and whole-class discussion of a confounded experiment, are used relatively little whole-class attention has to be spent on experimenting. The remaining time can then be used to further instruct children who do not learn experimenting during the whole-class intervention. Hypothesizing and drawing conclusions are similar both in nature and in difficulty. These skills require children to be specific and precise about their expectations or findings, and are neither difficult nor hard for most children. As such, baseline support for the entire classroom is warranted for these component skills. This support should be more rigorous for drawing conclusions, which is more difficult than hypothesizing as it should reflect the collected data, possibly combined with background information, rather than personal views (Koerber, Osterhaus, et al., 2015; Koslowski et al., 2008).

Lastly, monitoring and adaptation of support should take place on the component skill level. This recommendation is based on the results presented in Chapter 4, which showed that some children do well on experimenting but less well on drawing conclusions. These children should therefore receive extra support for the latter skill. Other children were proficient in drawing conclusions but less skilful in experimenting. These children would benefit from support for experimenting, but not from support on drawing conclusions. Adapting support on the component skill level is not feasible with worksheet-based support alone, and as digital adaptive teaching materials for scientific reasoning are not yet developed, teachers could monitor these developments more informally in order to provide just-in-time support or extended instruction for specific skills to those children who need this extra support.

### **Concluding remarks**

The studies presented in this dissertation aimed to describe and improve the learning of scientific reasoning of upper primary school children. The science classes children of this

age group participate in often include lively, hands-on activities in which experiments are set up and carried out using physical materials. It was in this context that two aspects of scientific reasoning in the primary school classroom were studied.

First, with regard to the learning of scientific reasoning at the component skill level it was found that some skills are more difficult than others, and that children differed in their scientific reasoning proficiency. Zooming in further, this variation was found to be not uniform: different proficiency profiles were found for different groups of children. The second aspect explored in this dissertation was adaptivity in the unplugged science classroom. Although this specific intervention did not improve children's scientific reasoning, it did provide practical experience with regard to the implementation of adaptivity in unplugged learning environments.

Together, the studies in this dissertation emphasize once again the importance of treating scientific reasoning as a multidimensional construct in both research and practice, and provides important implications for teaching scientific reasoning in the primary school classroom. The latter is particularly important for the Dutch educational context, where schools and teachers have a large curricular responsibility and science education still very much under development. In this development, it is important to help schools attend to scientific reasoning: in the modern knowledge society we live in, having a thorough command of scientific reasoning helps children learn from inquiry and to practice intentional knowledge seeking throughout their lives.

## References

- Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., Niaz, M., Treagust, D., & Tuan, H. L. (2004). Inquiry in science education: International perspectives. *Science Education, 88*(3), 397-419. <https://doi.org/10.1002/sce.10118>
- Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016). Instruction based on adaptive learning technologies. In R. E. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 522-560). Routledge.
- Anastasi, J. S. (2007). Full-semester and abbreviated summer courses: An evaluation of student performance. *Teaching of Psychology, 34*(1), 19-22. <https://doi.org/10.1080/00986280709336643>
- Ben-Zvi, D., & Garfield, J. (2004). The challenge of developing statistical literacy, reasoning and thinking. Kluwer Academic Publishing.
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098-1120. <https://doi.org/10.1111/1467-8624.00081>
- Chinn, C. A., & Golan Duncan, R. (2018). What is the value of general knowledge of scientific reasoning? In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific Reasoning and Argumentation* (pp. 77-101). Routledge.
- Deichert, N. T., Maxwell, S. J., & Klotz, J. (2015). Retention of information taught in introductory psychology courses across different accelerated course formats. *Teaching of Psychology, 43*(1), 4-9. <https://doi.org/10.1177/0098628315619725>
- Deunk, M. I., Smale-Jacobse, A. E., De Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education. *Educational Research Review, 24*, 31-54. <https://doi.org/10.1016/j.edurev.2018.02.002>
- Edelsbrunner, P. A., & Dablander, F. (2018). The psychometric modeling of scientific reasoning: A review and recommendations for future avenues. *Educational Psychology Review, 31*(1), 1-34. <https://doi.org/10.1007/s10648-018-9455-5>
- Evangelou, F., & Kotsis, K. (2018). Real vs virtual physics experiments: Comparison of learning outcomes among fifth grade primary school students. A case on the concept of frictional force. *International Journal of Science Education, 41*(3), 330-348. <https://doi.org/10.1080/09500693.2018.1549760>
- Greven, J., & Letschert, J. (2006). *Kerndoelen primair onderwijs* [Curricular goals for primary education]. Ministerie van Onderwijs, Cultuur en Wetenschap.
- Grimm, H., Edelsbrunner, P. A., & Möller, K. (2021). Accommodating heterogeneity: The interaction of instructional scaffolding with student preconditions in the learning of hypothesis-based reasoning. *PsyArXiv*. <https://doi.org/10.31234/osf.io/sn9c3>
- Inspectie van het Onderwijs. (2017). *Pijl.Natuur en Techniek* [Level of Science Education and Performance]. Inspectie van het Onderwijs.
- Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*(4), 63-85. <https://doi.org/10.1007/BF02300500>
- Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching, 41*(7), 748-769. <https://doi.org/10.1002/tea.20020>
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1-48. [https://doi.org/10.1207/s15516709cog1201\\_1](https://doi.org/10.1207/s15516709cog1201_1)

- Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development, 86*(1), 327-336. <https://doi.org/10.1111/cdev.12298>
- 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.org/10.1080/15248372.2019.1620232>
- 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>
- Koslowski, B., Marasia, J., Chelenza, M., & Dublin, R. (2008). Information becomes evidence when an explanation can incorporate it into a causal framework. *Cognitive Development, 23*(4), 472-487. <https://doi.org/10.1016/j.cogdev.2008.09.007>
- Kruit, P., Oostdam, R., Van den Berg, E., & Schuitema, J. (2020). Performance assessment as a diagnostic tool for science teachers. *Research in Science Education, 50*, 1093-1117. <https://doi.org/10.1007/s11165-018-9724-9>
- Krummenauer, J., & Kuntze, S. (2019, February). *Primary students' reasoning and argumentation based on statistical data*. Eleventh Congress of the European Society for Research in Mathematics Education, Utrecht, the Netherlands. <https://hal.archives-ouvertes.fr/hal-02398118/>
- Lazonder, A. W., & Janssen, N. (2021). Development and initial validation of a performance-based scientific reasoning test for children. *Studies in Educational Evaluation, 68*, Article 100951. <https://doi.org/10.1016/j.stueduc.2020.100951>
- Lazonder, A. W., & Wiskerke-Drost, S. (2015). Advancing scientific reasoning in upper elementary classrooms: Direct instruction versus task structuring. *Journal of Science Education and Technology, 24*(1), 69-77. <https://doi.org/10.1007/s10956-014-9522-8>
- Lorch, R. F., Lorch, E. P., Freer, B., Calderhead, W. J., Dunlap, E., Reeder, E. C., Van Neste, J., & Chen, H. T. (2017). Very long-term retention of the control of variables strategy following a brief intervention. *Contemporary Educational Psychology, 51*, 391-403. <https://doi.org/10.1016/j.cedpsych.2017.09.005>
- Lorch, R. F., Lorch, E. P., Freer, B. D., Dunlap, E. E., Hodell, E. C., & Calderhead, W. J. (2014). Using valid and invalid experimental designs to teach the control of variables strategy in higher and lower achieving classrooms. *Journal of Educational Psychology, 106*(1), 18-35. <https://doi.org/10.1037/a0034375>
- Makar, K. (2016). Developing young children's emergent inferential practices in statistics. *Mathematical Thinking and Learning, 18*(1), 1-24. <https://doi.org/10.1080/10986065.2016.1107820>
- Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning, 13*(1-2), 152-173. <https://doi.org/10.1080/10986065.2011.538301>
- Masnick, A., Klahr, D., & Knowles, E. R. (2017). Data-driven belief revision in children and adults. *Journal of Cognition and Development, 18*(1), 87-109. <https://doi.org/10.1080/15248372.2016.1168824>
- Masnick, A., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development, 79*(4), 1032-1048. <https://doi.org/10.1111/j.1467-8624.2008.01174.x>
- 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>

- McCrea Simpkins, P., Mastropieri, M. A., & Scruggs, T. E. (2009). Differentiated curriculum enhancements in inclusive fifth-grade science classes. *Remedial and Special Education, 30*(5), 300-308. <https://doi.org/10.1177/0741932508321011>
- Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning – a review of test instruments. *Educational Research and Evaluation, 23*(3-4), 78-101. <https://doi.org/10.1080/13803611.2017.1338586>
- Osborne, J. (2018). Styles of scientific reasoning: What can we learn from looking at the product, not the process, of scientific reasoning? In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation* (pp. 162-186). Routledge.
- Osterhaus, C., Koerber, S., & Sodian, B. (2017). Scientific thinking in elementary school: Children's social cognition and their epistemological understanding promote experimentation skills. *Developmental Psychology, 53*(3), 450-462. <https://doi.org/10.1037/dev0000260>
- 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. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review, 14*, 47-61. <https://doi.org/10.1016/j.edurev.2015.02.003>
- Peteranderl, S. (2019). *Experimentation skills of primary school children*. [Doctoral dissertation, ETH Zürich]. ETH Zürich Research Collection. <https://doi.org/10.3929/ethz-b-000370663>
- Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology, 31*(2), 153-179. <https://doi.org/10.1111/j.2044-835X.2012.02082.x>
- Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education, 52*(3), 275-300. <https://doi.org/10.1080/15391523.2020.1719943>
- Rönnebeck, S., Bernholt, S., & Ropohl, M. (2016). Searching for a common ground – A literature review of empirical research on scientific inquiry activities. *Studies in Science Education, 52*(2), 161-197. <https://doi.org/10.1080/03057267.2016.1206351>
- Salac, J., & Franklin, D. (2020, June 15–19). If they build it, will they understand it? Exploring the relationship between student code and performance. ITiCSE '20, Trondheim, Norway.
- Schalk, L., Edelsbrunner, P. A., Deiglmayr, A., Schumacher, R., & Stern, E. (2019). Improved application of the control-of-variables strategy as a collateral benefit of inquiry-based physics education in elementary school. *Learning and Instruction, 59*, 34-45. <https://doi.org/10.1016/j.learninstruc.2018.09.006>
- Schauble, L. (2018). In the eye of the beholder: Domain-general and domain-specific reasoning in science. In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific Reasoning and Argumentation* (pp. 11-33). Routledge.
- Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review, 39*, 37-63. <https://doi.org/10.1016/j.dr.2015.12.001>
- 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*, Article 101923. <https://doi.org/10.1016/j.cedpsych.2020.101923>
- Siler, S. A., Klahr, D., Magaro, C., Willows, K., & Mowery, D. (2010, June). *Predictors of transfer of experimental design skills in elementary and middle school children*. 10th International Conference on Intelligent Tutoring Systems, Pittsburgh, PA.
- Techniecpact. (2013). *Nationaal Techniecpact 2020*. Retrieved from <https://www.rijksoverheid.nl/documenten/convenanten/2013/05/13/nationaal-techniecpact-2020>



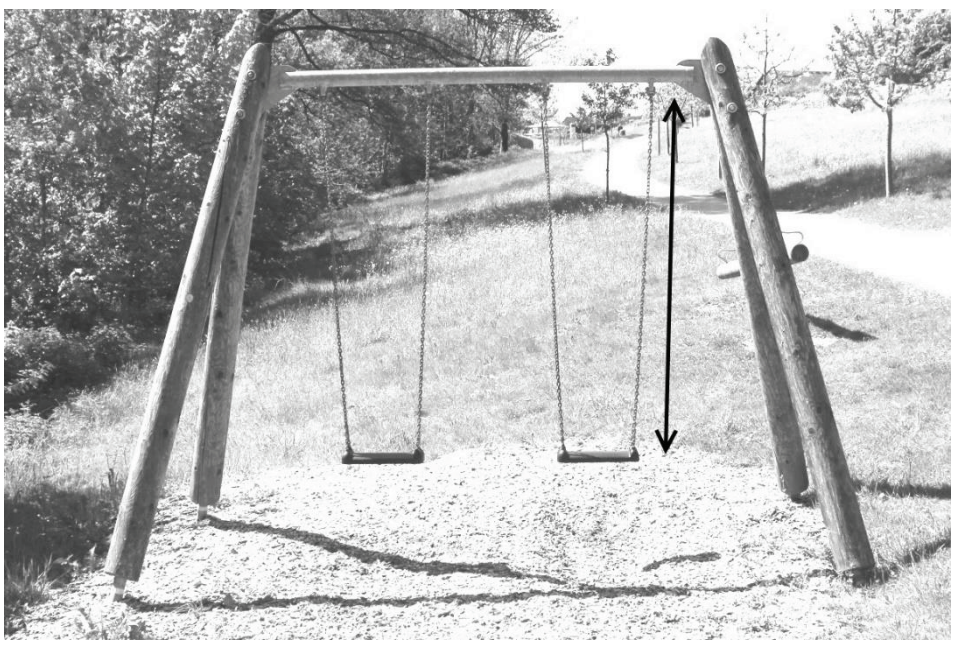
- Trilling, B., & Fadel, C. (2009). 21st century skills - learning for life in our times. Jossey-Bass.
- Van de Sande, E., Kleemans, M., Verhoeven, L., & Segers, E. (2019). The linguistic nature of children's scientific reasoning. *Learning and Instruction*, 62(1), 20-26. <https://doi.org/10.1016/j.learninstruc.2019.02.002>
- Wagensveld, B., Segers, E., Kleemans, T., & Verhoeven, L. (2014). Child predictors of learning to control variables via instruction or self-discovery. *Instructional Science*, 43(3), 365-379. <https://doi.org/10.1007/s11251-014-9334-5>
- Yannier, N., Hudson, S. E., & Koedinger, K. R. (2020). Active learning is about more than hands-on: A mixed-reality AI system to support STEM education. *International Journal of Artificial Intelligence in Education*, 30(1), 74-96. <https://doi.org/10.1007/s40593-020-00194-3>
- Zacharia, Z. C., Manoli, C., Xenofontos, N., de Jong, T., Pedaste, M., van Riesen, S. A. N., Kamp, E. T., Mäeots, M., Siiman, L., & Tsourlidaki, E. (2015). Identifying potential types guidance for supporting student inquiry when using virtual and remote labs in science: A literature review. *Educational Technology Research and Development*, 63(2), 257-302. <https://doi.org/10.1007/s11423-015-9370-0>
- 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>

# **Appendix A**

## **Worksheet example for Chapters 3 and 4**

# RESEARCH

Does the length of the rope affect the period of a swing?



 Expectations

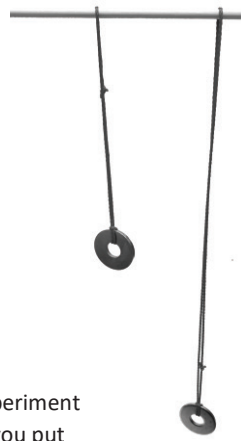
What do you think will happen to the period of a swing if you change the length of the rope?

I think .....

.....



# Experiment



Think up two experiments to check if the length of the rope affects the period. For each experiment you have to choose the length of the rope, how far you pull it aside and how much weight you put on the pendulum

**Experiment 1A**

Rope length: .....

How far pulled aside: .....

Weighth: .....

**Experiment 1B**

Rope length: .....

How far pulled aside: .....

Weighth: .....

Perform each of the experiments together with an assistant. Take care to measure the same thing each time: how long does it take for the pendulum to swing 5 times?

**Experiment 1A**

Outcome: .....

**Experiment 1B**

Outcome: .....

Do you see a difference between a short and a long rope?

**yes/no**

To be more sure, you can repeat your experiments:

**Experiment 1A**

Outcome 2nd trial: .....

Outcome 3rd trial: .....

**Experiment 1B**

Outcome 2nd trial: .....

Outcome 3rd trial: .....



Copy all outcomes of all trials in the table below (do not forget your first trial!). That way, all outcomes are conveniently grouped together.

	<b>Experiment 1A</b>	<b>Experiment 1B</b>
1st trial	.....	.....
2nd trial	.....	.....
3rd trial	.....	.....

Do you see a difference between a short and a long rope in the table?

**yes/no**

How do you know?

.....



## Conclusion

You can now answer the research question:

Does the length of the rope affect the period of a swing?

My research shows .....

.....



# **Appendix B**

## **Support levels during inquiry sessions**



## Hypothesizing

- (1) What are you going to change? (identify the variable of interest)
  - (a) You can change **A**, **B** or **C**
  - (b) Give answer
- (2) How can you change **A**? (determine the levels of the variable of interest)
  - (a) Which different **A**'s can you use in this inquiry?  
(e.g. if rope length is the variable of interest, you can choose long or short)
  - (b) Give answer
- (3) What are you going to measure? (identify output variable)
  - (a) Will you measure **X**, or **Y**, or something else?
  - (b) Give answer
- (4) Repeat original question: what do you think will happen with **X** if you change **A**?

## Experimenting

- (1) What are you investigating? (identify the variable of interest)
  - (a) You can investigate **A**, **B** or **C**, which one are you investigating on this worksheet?
  - (b) Give answer
- (2) What should you do with **A** to find out whether **A** makes a difference for **outcome X**?
  - (a) You can change **A** or keep it the same, what do you think is best?
  - (b) Give answer
- (3) What should you do with **B** and **C** to investigate fairly whether **A** makes a difference for **outcome X**?
  - (a) You can change **B** and **C** or keep them the same, what should you do for a fair investigation?
  - (b) Give answer

### Inferencing 1 (single comparison)

- (1) What did you measure?
  - (a) Did you measure **X**, or **Y**, or something else?
  - (b) Give answer
- (2) What was the outcome of **experiment 1**? and of **experiment 2**?
  - (a) Point out what the child wrote down
  - (b) Give answer
- (3) Where the outcomes the same or different?
  - (a) Is **outcome 1** the same as **outcome 2**?
  - (b) Give answer

### Inferencing 2 (multiple comparison)

- (1) For every time you performed the experiment, draw a circle around the largest outcome
  - (a) The first time you did the experiment, was the outcome of **experiment 1** the largest or the outcome of **experiment 2**? Draw a circle around the largest one.
    - (i) Ask child to do the same for each replication
    - (ii) Repeat question 1a for each of the replication
  - (b) Give answer
- (2) Did you draw more circles for **experiment 1** or for **experiment 2**?
  - (a) How many circles did you draw for **experiment 1**? And how many for **experiment 2**?
  - (b) Give answer
- (3) Was the difference very large, or not so large?

## Conclusion

- (1) What did you investigate? (identify the variable of interest)
  - (a) Did you investigate **A**, **B** or **C** on this worksheet?
  - (b) Give answer
- (2) What was the **value** of **A** in experiment 1? And in **experiment 2**? (determine the levels of the variable of interest & how they were set in the experiments this child performed)
  - (a) Stimulate to look it up
  - (b) Point out/give answer
- (3) Were the outcomes for **experiment 1** different than those for **experiment 2**?
  - (a) Stimulate to look it up
  - (b) Point out/give answer

# **Appendix C**

## **Worksheet example for Chapter 5**

# INQUIRY I

Does the surface matter for the number of bounces a ball makes?



What do you think will happen to the number of bounces if you change the surface?

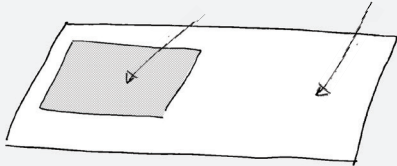
## WHAT TO DO

Think about what you're about to investigate:

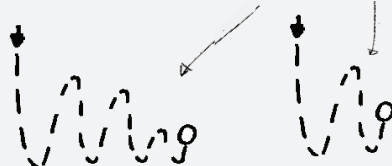
- The surface
- The number of bounces

## HOW TO DO IT

The surface can be soft or it can be hard.



The number of bounces can be a lot or a little.



With a different surface the number of bounces can increase, decrease or stay the same.

To formulate a good expectation, you can use the underlined words.

What do you think will happen to the number of bounces?

I think .....

.....



# Experiment

Think up two experiments to check if the surface affects the number of bounces.

## WHAT TO DO

Two things are important when designing your experiment:

- You have to know what you are going to change
- You have to know what you are going to keep the same

## HOW TO DO IT

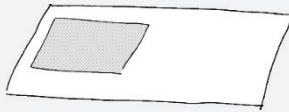
For each experiment you have to make three choices:

The **type of ball**



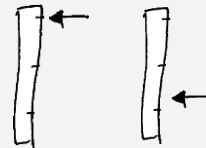
Pingpong or styrofoam

The **surface**



Soft or hard

The **starting height**



High or low

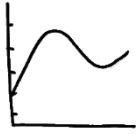
You're going to investigate one of these. That's the one you change.

You're not going to investigate the other two. You keep those two the same in both experiments.

Write your experiments down:

Experiment A
Type of ball:
Surface:
Starting height:

Experiment B
Type of ball:
Surface:
Starting height:



Outcomes

Perform each of the experiments three times and write your outcomes down in this table:

	Experiment A Surface:	Experiment B Surface:
1st trial		
2nd trial		
3rd trial		

## WHAT TO DO

Take notice:

- Is the outcome of one experiment always bigger, or does it change each time?
- Is the difference consistent?

## HOW TO DO IT

Record the largest outcome for each trial. Which pattern do you

see?

Experiment A is always the largest 	Experiment B is always the largest 	Experiment A and B are about the same 	Sometimes experiment A is largest, sometimes B 
--	--	---	--

Also check whether the difference are large or small. Can you be sure there is a difference?

Is there a difference between a hard and a soft surface?

yes/no

Explain how you used the table to figure that out:

.....



# Conclusion

You can now answer the research question:

## Does the surface matter for the number of bounces a ball makes?

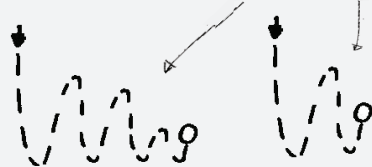
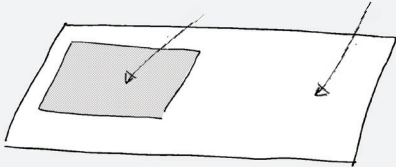
### WHAT TO DO

Think about what you investigated:

- The surface
- The number of bounces

### HOW TO DO IT

The surface could be soft or it could be hard. The number of bounces could be a lot or a little.



With a different surface the number of bounces could increase, decrease or stay the same.

To formulate a good conclusion, you can use the underlined words.

What happened to the number of bounces?

My research shows .....

.....



# Word search

Did you finish inquiry 1? Don't continue with inquiry 2. Below, you find a word search puzzle about doing research. You can work on the puzzle as long as the others are working on the inquiry.

J	M	A	M	O	P	R	O	E	F	J	E	U	B	A	L	L	P
T	A	E	S	T	U	I	T	E	R	L	A	A	G	E	E	I	O
O	T	N	Y	V	G	U	I	T	K	O	M	S	T	E	N	N	E
N	E	C	O	Y	L	E	L	W	P	R	T	W	R	D	H	I	X
D	R	Y	Y	V	M	I	V	O	M	T	K	F	W	W	O	A	P
E	I	R	S	I	K	E	R	U	R	C	N	L	O	T	O	A	E
R	A	Z	H	L	W	H	O	L	L	L	H	Z	N	I	G	L	R
Z	A	C	J	T	S	A	N	C	K	D	A	A	S	F	D	Y	I
O	L	X	K	K	Z	A	C	H	T	T	E	L	L	E	N	Q	M
E	D	Z	W	A	A	R	T	E	K	R	A	C	H	T	Z	L	E
K	O	N	D	E	R	Z	O	E	K	S	C	Y	C	L	U	S	N
E	D	A	X	H	G	X	S	C	O	N	C	L	U	S	I	E	T
R	M	Q	G	F	E	E	T	A	B	E	L	V	M	H	J	L	E
W	R	P	V	M	N	O	B	T	A	F	E	L	G	M	V	K	H
G	E	O	N	D	E	R	G	R	O	N	D	B	K	E	W	T	A
J	U	A	D	P	I	N	G	P	O	N	G	M	R	T	Z	W	R
V	E	R	W	A	C	H	T	I	N	G	G	S	J	E	J	S	D
B	C	D	Y	P	I	E	P	S	C	H	U	I	M	N	O	D	D

BAL  
CONCLUSIE  
EXPERIMENT  
GEVULD  
HARD  
HOL  
HOOG  
LAAG  
LINIAAL

MATERIAAL  
METEN  
ONDERGROND  
ONDERZOEKER  
ONDERZOEKSCYCLUS  
PIEPSCHUIM  
PINGPONG  
PROEFJE  
STUITER

TABEL  
TAFEL  
TELLEN  
UITKOMSTEN  
VERWACHTING  
VILT  
ZACHT

# INQUIRY 2

Does the starting height matter for the number of bounces the ball makes?



What do you think will happen to the number of bounces if you change the starting height?

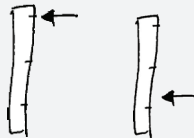
## WHAT TO DO

Think about what you're about to investigate:

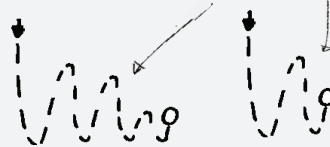
- The starting height
- The number of bounces

## HOW TO DO IT

The starting height can be high or low.



The number of bounces can be a lot or a little.



With a different starting height the number of bounces can increase, decrease or stay the same.

To formulate a good expectation, you can use the underlined words.

What do you think will happen to the number of bounces?

I think .....

.....



# Experiment

Think up two experiments to check if the starting height affects the number of bounces.

## WHAT TO DO

Two things are important when designing your experiment:

- You have to know what you are going to change
- You have to know what you are going to keep the same

## HOW TO DO IT

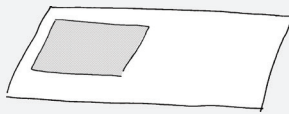
For each experiment you have to make three choices:

The **type of ball**



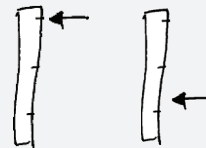
Pingpong or styrofoam

The **surface**



Soft or hard

The **starting height**



High or low

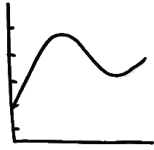
You're going to investigate one of these. That's the one you change.

You're not going to investigate the other two. You keep those two the same in both experiments.

Write your experiments down:

Experiment A
Type of ball:
Surface:
Starting height:

Experiment B
Type of ball:
Surface:
Starting height:



Outcomes

Perform each of the experiments three times and write your outcomes down in this table:

	Experiment A Starting height:	Experiment B Starting height:
1st trial		
2nd trial		
3rd trial		

## WHAT TO DO

Take notice:

- Is the outcome of one experiment always bigger, or does it change each time?
- Is the difference consistent?

## HOW TO DO IT

In the table above, draw a circle around the largest outcome for each trial. Which pattern do you see?

Experiment A is always the largest <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Experiment B is always the largest <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Experiment A and B are about the same <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>	Sometimes experiment A is largest, sometimes B <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
---	---	--	---

Also check whether the difference are large or small. Can you be sure there is a difference?

Is there a difference between a high and a low start?

yes/no

Explain how you used the table to figure that out:

.....



## Conclusion

You can now answer the research question:

Does the starting height matter for the number of bounces the ball makes?

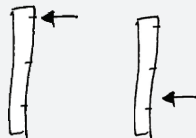
### WHAT TO DO

Think about what you investigated:

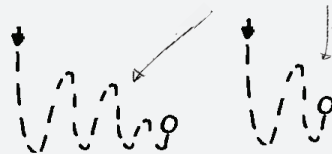
- The starting height
- The number of bounces

### HOW TO DO IT

The starting height could be high or low.



The number of bounces could be a lot or a little.



With a different starting height the number of bounces can increase, decrease or stay the same.

To draw a good conclusion, you can use the underlined words.

What happened to the number of bounces?

My research shows .....

.....

# Nederlandstalige samenvatting

Wetenschap en Technologie (ook wel bekend als W&T) is al enige tijd opgenomen in de kerndoelen voor het Nederlandse basisonderwijs, en is daar inmiddels ook behoorlijk goed ingeburgerd. Als je tijdens zo'n W&T-les de klas binnenloopt, is de kans groot dat je kinderen praktisch bezig ziet met het uitvoeren van hun eigen onderzoek, bijvoorbeeld door informatie op te zoeken of experimenten uit te voeren. Volgens kerndoel 42 (Greven & Letschert, 2006) moeten kinderen namelijk leren *'onderzoek doen aan materialen en natuurkundige verschijnselen, zoals licht, geluid, elektriciteit, kracht, magnetisme en temperatuur'*. In de tussendoelen en leerlijnen (TULE) van dit kerndoel<sup>1</sup> staan hierbij vooral vakinhoudelijke doelen gedefinieerd. Bovenbouwleerlingen zouden bijvoorbeeld moeten leren dat licht zich rechtlijnig voortplant en dat sommige materialen goede warmtegeleiders zijn terwijl andere materialen juist goed isoleren.

Bij het doen van onderzoek komen echter ook denkvaardigheden en procedurele vaardigheden kijken. Deze onderzoeksvaardigheden zijn om een aantal redenen belangrijk. Ze kunnen je bijvoorbeeld helpen om de wereld om je heen te begrijpen (Chinn & Golan Duncan, 2018; Kind & Osborne, 2017) en om in het dagelijks leven beslissingen te maken over maatschappelijk-wetenschappelijke thema's (Sadler, 2004), iets waar iedereen mee te maken heeft. Door onderzoeksvaardigheden op school aan te bieden, worden dus kinderen voorbereid op het leven in een moderne kennismaatschappij (Trilling & Fadel, 2009).

Maar onderzoeksvaardigheden zijn ook belangrijk op school. Een belangrijke reden daarvoor is de toenemende aandacht voor onderzoekend leren, waarbij kinderen onderzoeksactiviteiten uitvoeren met als doel het opdoen van vakinhoudelijke kennis. Voor onderzoekend leren heb je onderzoeksvaardigheden nodig. Je zou het kunnen vergelijken met begrijpend lezen, waarbij het doel is om kennis op te nemen uit een tekst. Het aanleren van bijvoorbeeld leesstrategieën kan hierbij helpen, maar alleen als een kind de tekst ook daadwerkelijk kan ontcijferen. Technisch lezen is dus een voorwaardelijke vaardigheid voor begrijpend lezen. Met onderzoeksvaardigheden is het net zo: wil je kunnen leren van je onderzoek, moet je ook leren onderzoeken (Edelsbrunner et al., 2018; Lazonder & Harmsen, 2016; Zimmerman, 2007).

Dat leren onderzoeken gaat niet voor elke leerling even gemakkelijk. Dit proefschrift gaat over het aanleren van onderzoeksvaardigheden. Er wordt daarbij gekeken naar

---

<sup>1</sup> <https://www.slo.nl/thema/meer/tule/orientatie-jezelf-wereld/kerndoel-42/>

verschillen tussen leerlingen, verschillen tussen vaardigheden, en mogelijke manieren om die verschillen te overbruggen. Daarbij staan twee onderzoeksvragen centraal:

1. Wat kenmerkt verschillen in onderzoeksvaardigheden van bovenbouwleerlingen, en hoe kunnen die verschillen voorspeld worden?
2. Hoe kan in de onderwijsleersituatie ingesprongen worden op deze verschillen?

Voor deze vragen beantwoord worden, zal eerst een overzicht gegeven worden van de onderzoeksvaardigheden die in dit proefschrift uitgelicht zijn. Dan worden de aanpak, resultaten en conclusies van de verschillende hoofdstukken besproken. Aan de hand van deze resultaten worden ook bovenstaande onderzoeksvragen beantwoordt. Als laatste worden nog handreikingen voor de praktijk gedaan.

### **Wat zijn onderzoeksvaardigheden en hoe verhouden ze zich tot elkaar?**

Om goed te kunnen omschrijven hoe onderzoeksvaardigheden samenhangen en van elkaar verschillen, is het belangrijk om eerst te weten wat onderzoeksvaardigheden eigenlijk zijn. In de literatuur worden grofweg drie hoofdcategorieën onderscheiden: hypothesen opstellen, experimenteren en uitkomsten beoordelen. Deze deelvaardigheden hebben veel met elkaar te maken – een hypothese kan je helpen bij het opzetten van je experiment, en met het experiment verzamel je uitkomsten die beoordeeld moeten worden – maar zijn toch ook verschillend.

Hypothesen opstellen vindt plaats aan de beginfase van een onderzoek. Bij het opstellen van een hypothese moet je nadenken over de mogelijke uitkomsten van het onderzoek. Dat is een vaardigheid die zich pas tegen het einde van de basisschool echt begint te ontwikkelen. Experimenteren is de volgende stap. Hoewel het uitvoeren van een experiment natuurlijk heel leuk is, is ook een goede planning belangrijk: een experiment moet systematisch zijn opgezet, zodat de uitkomsten van verschillende varianten van het experiment goed vergeleken kunnen worden. Kleuters kunnen vaak al herkennen of een experiment zo is opgezet dat er een eerlijke vergelijking kan worden gemaakt, en oudere basisschoolkinderen kunnen ook leren om zelf een goed experiment op te zetten.

Om de uitkomsten van het experiment komt te beoordelen, moet een aantal stappen gezet worden. Eerst moeten de numerieke uitkomsten (ook bekend als data) worden geïnterpreteerd. Hoe makkelijk dat gaat is sterk afhankelijk van de data. Als er een duidelijk verband te zien is, kunnen zelfs kleuters dat al herkennen. Maar het herkennen van minder duidelijke verbanden of zelfs het ontbreken van een verband is lastiger. Het

beoordelen van de kwaliteit van de data is, hoewel er niet vaak expliciet aandacht aan besteed wordt, ook een belangrijke stap: is er bijvoorbeeld genoeg informatie verzameld, en zijn er uitkomsten die heel erg afwijken van de rest? Dit is voor kinderen een lastige stap, net als het uiteindelijke trekken van conclusies, waarbij op basis van de uitkomsten een uitspraak moet worden gedaan over het waarheidsgehalte van de hypothese.

De onderzoeksvaardigheden lijken, op basis van eerder onderzoek, dus behoorlijk van elkaar te verschillen. Maar net als bij elk ander schoolvak zijn er ook verschillen tussen kinderen: sommige kinderen kunnen beter onderzoeken dan andere kinderen. Omdat die verschillen in de loop van de basisschool lijken te groeien (Piekný & Maehler, 2013), is het in de bovenbouw extra belangrijk dat leerkrachten inspelen op verschillen tussen leerlingen.

Het doen van onderzoek vereist dus een groot scala aan vaardigheden die met elkaar samenhangen, maar ook verschillend zijn. Omdat er ook nog rekening moet worden gehouden met verschillen tussen leerlingen, kan het lesgeven in onderzoeksvaardigheden uitdagend zijn voor leerkrachten. Het is dan ook belangrijk om goed te begrijpen welke verschillen er bestaan, tussen kinderen en tussen vaardigheden, zodat er op die verschillen ingespeeld kan worden. In eerder onderzoek is weliswaar gekeken naar zulke verschillen, maar daarbij zijn vaak losse onderzoeksvaardigheden onderzocht, of zijn onderzoeksvaardigheden juist als één geheel behandeld. In dit proefschrift is geprobeerd om onderzoeksvaardigheden wel van elkaar te onderscheiden, maar niet te isoleren. Want hoewel er belangrijke verschillen zijn, horen hypothesen opstellen, experimenteren, uitkomsten interpreteren en conclusies trekken ook inherent bij elkaar.

### **Verschillen tussen leerlingen en tussen vaardigheden**

In de studie in hoofdstuk 2 is gekeken naar de onderzoeksvaardigheden van bovenbouwleerlingen van de basisschool. Hoewel deze kinderen tijdens de reguliere wetenschap- en technologielessen kennis hadden gemaakt met het doen van onderzoek, is er in deze lessen niet uitgebreid en ook niet expliciet stilgestaan bij het aanleren van onderzoeksvaardigheden. Om te onderzoeken of en hoe deze leerlingen van elkaar verschillen, en of de onderzoeksvaardigheden onderling verschillen in moeilijkheid, is een praktische taak gebruikt. Tijdens deze taak voerden kinderen zelf vier korte onderzoekjes uit en werden ze beoordeeld op in totaal vijf deelvaardigheden: hypothesen opstellen, experimenteren, interpreteren van data, beoordelen van data en conclusies trekken. Het doel van de studie was het beschrijven van verschillen, zowel tussen kinderen als tussen de deelvaardigheden.



Eerder onderzoek heeft al laten zien dat kinderen de deelvaardigheden niet op dezelfde leeftijd opdoen. Dat impliceert dat de deelvaardigheden verschillen in moeilijkheid – maar meestal worden slechts één of twee deelvaardigheden onderzocht, of worden de scores voor alle deelvaardigheden bij elkaar opgeteld. Daardoor was het nog onduidelijk of de deelvaardigheden voor kinderen van dezelfde leeftijd inderdaad verschillen in moeilijkheid. De studie in hoofdstuk 2 bevestigt dat onderzoeksvaardigheden verschillen in moeilijkheid: over het algemeen waren de scores het hoogst voor experimenteren, het laagst voor het beoordelen van data en het trekken van conclusies, en gemiddeld voor het interpreteren van data.

Eerder onderzoek laat ook zien dat er verschillen zijn tussen kinderen van dezelfde leeftijd. Zulke verschillen worden vaak omschreven aan de hand van zogenaamde voorspellers, eigenschappen van kinderen aan de hand waarvan kan worden verklaard welke kinderen beter of juist minder goed zijn in onderzoeken. In hoofdstuk 2 zijn drie van die voorspellers onderzocht om verschillen tussen kinderen te verklaren.

De eerste onderzochte voorspeller is begrijpend lezen, een vaardigheid waarvan al bekend was dat die samenhangt met onderzoeksvaardigheden, maar die in eerder onderzoek alleen gebruikt was in combinatie met papieren toetsen voor onderzoeksvaardigheden. In de studie in hoofdstuk 2 is onderzocht of de voorspellende waarde van begrijpend lezen in stand bleef als kinderen een onderzoekstaak deden waarvoor ze (bijna) niet hoefden te lezen. De tweede voorspeller die in hoofdstuk 2 onderzocht is, is rekenvaardigheid. Hoewel cijfers, getallen en berekeningen een prominente rol hebben in het doen van onderzoek, is rekenvaardigheid in het verleden weinig onderzocht als voorspeller voor onderzoeksvaardigheden. In dit onderzoek is rekenvaardigheid (gemeten met de schoolvaardigheidstoets hoofdrekenen), maar niet wiskundig redeneren, meegenomen als voorspeller. De laatste onderzochte voorspeller is probleemoplossend vermogen, een brede vaardigheid die raakt aan het doen van onderzoek: ontbrekende kennis (het probleem) wordt gevonden door middel van onderzoek (de oplossing). Er zijn echter ook abstractere vormen van probleemoplossend vermogen, die met kortere taken gemeten worden. In hoofdstuk 2 is zo'n kortere taak gebruikt om het probleemoplossend vermogen van kinderen in te schatten: de toren van Hanoi.

Begrijpend lezen was de enige voorspeller voor verschillen tussen kinderen. Omdat onderzoeksvaardigheden in hoofdstuk 2 zijn gemeten met een praktische test, kan dat niet komen doordat kinderen voor de test zelf veel moesten lezen – zoals in eerder onderzoek het geval was. Dat versterkt het beeld dat begrijpend lezen een goede voorspeller is van onderzoeksvaardigheden. Begrijpend lezen verklaart zowel onderzoeksvaardigheden in het

algemeen (waarbij scores voor alle vaardigheden bij elkaar opgeteld zijn), als vier van de vijf deelvaardigheden individueel. Alleen het opstellen van hypothesen werd niet door begrijpend lezen voorspeld.

### Leren onderzoeken op school

In hoofdstuk 3 en 4 wordt dieper ingegaan op de verschillen tussen kinderen en tussen onderzoeksvaardigheden tijdens het leren onderzoeken. Want hoewel eerder onderzoek laat zien dat niet alle onderzoeksvaardigheden zich tegelijkertijd ontwikkelen en dat niet alle kinderen even goed zijn in onderzoeken, is er minder bekend over hoe kinderen leren onderzoeken in een lessituatie. Zoals bij de meeste schoolvakken worden kinderen over het algemeen beter in onderzoeken als ze daar les in krijgen, maar leren niet alle kinderen even goed of even snel onderzoeken. Bij leren onderzoeken is er bovendien sprake van een set vaardigheden die inherent bij elkaar horen, maar ook verschillend zijn in aard en moeilijkheid. Dat betekent dat aan sommige vaardigheden meer aandacht en instructietijd besteed moet worden, en ook dat sommige kinderen meer instructie en ondersteuning nodig hebben dan andere kinderen.

Daarom is het belangrijk om te begrijpen hoe verschillen in het leren onderzoeken zich manifesteren in de onderwijsleersituatie. In hoofdstukken 3 en 4 is dat vraagstuk op verschillende manieren benaderd. Beide hoofdstukken beschrijven een studie waarin kinderen vijf weken lang elke week een uur les kregen in onderzoeksvaardigheden. Tijdens de eerste les is klassikaal besproken wat onderzoek doen inhoudt, en zijn vier onderzoeksvaardigheden uitgelicht: hypothesen opstellen, experimenteren, data interpreteren en conclusies trekken. Na deze uitleg gingen de kinderen in kleine groepjes aan de slag met hun eigen onderzoek, waarin deze vier onderzoeksvaardigheden aan bod kwamen. Elke les was volgens dezelfde structuur opgezet: uitleg of herhaling van de onderzoeksvaardigheden, 20 minuten zelf onderzoeken, het bespreken van de uitkomsten, nog eens 20 minuten zelf onderzoeken en een afsluiting waarin de uitkomsten en onderliggende principes besproken werden. Een uitzondering hierop vormde de laatste les: in deze les konden de kinderen maar één eigen onderzoek uitvoeren, omdat direct na de les de nameting afgenomen moest worden.

In hoofdstuk 3 is vooral gekeken naar verschillen tussen leerlingen als het gaat om leeruitkomsten, en of die verschillen verklaard konden worden aan de hand van gegevens uit het leerlingvolgsysteem. Kinderen maakten voorafgaand aan de lessenserie een voormeting, en na afloop van de laatste les een nameting. Na afloop van de lessenserie presteerden

kinderen beter op deze test dan vóór de lessenserie. Hoeveel kinderen vooruit gingen, hing samen met hun cito-scores op begrijpend lezen en rekenen. Begrijpend lezen bleek hierbij een sterkere voorspeller dan rekenen, en voor zowel begrijpend lezen als rekenen geldt dat ze niet in dezelfde mate samenhangen met elke deelvaardigheid.

Uit het onderzoek beschreven in hoofdstuk 3 bleek ook dat een deel van de kinderen vooruitgang boekten op de werkbladen die ze tijdens de lessen maakten, maar dat deze vooruitgang na ongeveer drie lessen stakte. Hoofdstuk 4 gaat verder in op verschillen in prestaties van leerlingen tijdens de lessen aan de hand van een latente transitieanalyse. In tegenstelling tot de analyses die in de voorgaande hoofdstukken gebruikt zijn, maakt latente profielanalyse geen gebruik van vooraf bekende factoren zoals rekenvaardigheid om verschillen te beschrijven. In plaats daarvan zijn kinderen in hoofdstuk 4 enkel gegroepeerd op basis van hun onderzoeksvaardigheden.

Op deze manier ontstonden vier vaardigheidsprofielen. In al deze profielen haalden kinderen vergelijkbare scores voor hypothesen opstellen en data interpreteren. Er waren kinderen die daarnaast op experimenteren en conclusies trekken relatief hoge scores haalden, de hoogpresteerders. Er waren ook kinderen die op alle onderzoeksvaardigheden relatief lage scores haalden, de laagpresteerders. De laatste twee profielen lijken een soort tussenfase tussen laag en hoog presteren te beschrijven. Niet alle kinderen lijken hierin op elkaar: sommige kinderen, de experimenteerders, scoren hoog op experimenteren en laag op conclusies trekken, terwijl anderen laag scoorden op experimenteren en hoog op conclusies trekken, de theoretici.

Na het vaststellen van deze profielen is ook gekeken naar de transities die kinderen maakten tussen de profielen. Daarbij vielen een aantal dingen op. Ten eerste wisselden kinderen heel vaak tussen profielen, en was er geen enkel pad dat door meer dan 10% van de kinderen gevolgd werd. Ten tweede bleek dat kinderen die eenmaal hadden geleerd te experimenteren daarna meestal goed bleven experimenteren – ze vielen dus niet terug naar de profielen laagpresteerders of theoretici. Het lijkt dus niet waarschijnlijk dat kinderen in de loop van hun leerproces beide ‘gemiddelde’ profielen doorlopen. In plaats van één duidelijke leerroute te zijn die alle leerlingen volgen, is het waarschijnlijker dat kinderen die van laag naar hoog presteren gaan, dat ofwel via het experimenteerders-profiel, ofwel via het theoretici-profiel doen.

## Inspelen op verschillen tussen leerlingen

In hoofdstuk 2 tot en met 4 zijn de verschillen tussen leerlingen en onderzoeksvaardigheden uitgebreid uitgewerkt. In hoofdstuk 5 wordt een studie beschreven waarin geprobeerd is om op de verschillen tussen leerlingen in te spelen. Kinderen volgden hiervoor vier lessen, die qua structuur en inhoud leken op de lessen die in hoofdstuk 3 en 4 gebruikt zijn. De zeven klassen die meededen aan dit onderzoek werden opgedeeld in drie groepen: een controlegroep en twee experimentele groepen. In alle groepen kregen alle kinderen dezelfde instructie. In de controlegroep kregen alle kinderen ook dezelfde ondersteuning via de werkbladen. In de twee experimentele groepen kregen sommige kinderen meer, en andere kinderen minder ondersteuning via de werkbladen. Om kinderen zo eerlijk mogelijk met elkaar te kunnen vergelijken, werd buiten de werkbladen om geen inhoudelijke ondersteuning geboden. Wel konden kinderen hulp vragen bij de praktische uitvoering van hun experiment. Als kinderen inhoudelijke vragen hadden, werd de ondersteuning van het werkblad voorgelezen.

Uit hoofdstuk 2 en 3 bleek dat de onderzoeksvaardigheden van kinderen voorspeld kunnen worden aan de hand van hun vaardigheid in begrijpend lezen en rekenen. Daarom werden die voorspellers in de eerste experimentele groep, de macro-adaptieve conditie, gebruikt om te bepalen welke kinderen meer en welke kinderen minder ondersteuning zouden krijgen via de werkbladen. In deze groep werden de kinderen dus vooraf ingedeeld op een ondersteuningsniveau, en werd dat ondersteuningsniveau niet veranderd gedurende de lessen.

Uit hoofdstuk 4 bleek dat het leerproces van kinderen heel dynamisch is. Daarom werd bij de tweede experimentele groep, de micro-adaptieve conditie, na elke les gekeken hoe leerlingen het deden. Als ze goed scoorden, kregen ze de volgende les minder ondersteuning via de werkbladen, en als ze niet zo goed scoorden kregen ze de volgende les meer ondersteuning via de werkbladen.

Uit de analyse van de voor- en nameting bleken dat kinderen in de drie groepen ongeveer evenveel geleerd hadden. Omdat veel kinderen in de twee experimentele groepen meestal ondersteuning op het middenniveau kregen, net zoals alle kinderen in de controlegroep, is ook nog gekeken naar de leerwinst van kinderen in de macro-adaptieve conditie die veel of juist weinig ondersteuning kregen. Deze kinderen zijn vergeleken met kinderen uit de controleconditie die vergelijkbare cito-scores hadden voor begrijpend lezen en rekenen. Ook hier was de leerwinst tussen groepen vergelijkbaar.

Als laatste is gekeken naar de vooruitgang op de werkbladen. Kinderen scoorden in de loop van de lessenserie steeds hoger op de werkbladen, en die verbetering lijkt samen te hangen met of zij in de controleconditie of één van de experimentele condities ingedeeld waren. Omdat het effect vrij klein was, kon niet worden bepaald welke conditie tot de grootste leerwinst leidde.

## Conclusie en implicaties voor de onderwijspraktijk

Het onderzoek in dit proefschrift laat een grote diversiteit zien in onderzoeksvaardigheden en het leren daarvan. Sommige vaardigheden bleken moeilijker dan andere, en daarnaast zijn er verschillen tussen kinderen. Deze variatie bleek niet uniform: sommige kinderen waren goed in één vaardigheid maar niet in de andere, terwijl dat voor anderen andersom was. Er is ook gekeken naar een manier om in te spelen op die verschillen, door via werkbladen extra ondersteuning te bieden aan kinderen die dat mogelijk nodig hadden. Hoewel de interventie in geen van de condities tot grote vooruitgang geleid heeft, leverde het onderzoek wel praktische ervaring op met adaptiviteit in een lessituatie zonder computers.

Alles bij elkaar genomen blijkt uit dit onderzoek opnieuw dat onderzoeksvaardigheden, zowel in de klas als in onderzoek, niet als één uniform construct gezien moeten worden. Tegelijkertijd laat het onderzoek ook zien dat onderzoeksvaardigheden niet in isolatie onderwezen of bestudeerd zouden moeten worden. De onderzoeksvaardigheden zijn dus verschillend in aard en moeilijkheid, maar horen ook bij elkaar. Dit is extra belangrijk met het oog op de lopende curriculumontwikkelingen in Nederland. Daarom wordt hieronder per onderzoeksvaardigheid besproken hoe in de les en in het lesmateriaal aandacht kan worden besteed aan verschillen, en hoe aanvullende instructie eruit kan zien.

Over het algemeen was de leerwinst in de studies uit dit proefschrift beperkt: zelfs bij kinderen die door herhaald oefenen hun onderzoeksvaardigheden verbeterden, stagneerde die verbetering rond de derde les. Omdat herhaald oefenen zo'n bescheiden leerwinst oplevert, lijkt het voor alle leerlingen belangrijk om gedurende het schooljaar aandacht te besteden aan leren onderzoeken. Een algemene, klassikale instructie en aanvullende ondersteuning via werkbladen lijkt hier niet genoeg voor te zijn.

Sommige onderzoeksvaardigheden, zoals hypotheses opstellen, blijkt voor de meeste kinderen vrij lastig. Het ligt dan ook voor de hand om daar met alle kinderen uitgebreid aandacht aan te besteden. Bij bestudering van de werkbladen in hoofdstuk 3 bleek dat veel

kinderen in de eerste les vrij hoog scoorden op hypothesen opstellen, en dat hun hypothesen sterk leken op de vraag die op het werkblad gesteld werd: ‘Wat denk je dat er met de uitkomstvariabele<sup>2</sup> gebeurt als je invoervariabele<sup>3</sup> verandert?’. In deze vraag zitten al veel onderdelen van een goede hypothese: de relevante variabelen worden genoemd en het is ook duidelijk dat er iets kan veranderen aan de uitkomstvariabele als de invoervariabele veranderd wordt. De vraag diende dus als een voorbeeld van een goede hypothese, dat door kinderen maar een klein beetje omgebogen hoefde te worden.

Na de eerste les daalden de scores op hypothesen opstellen, waarschijnlijk omdat kinderen dachten dat ze de vraag al kenden en daarom minder goed lasen. Om kinderen steeds opnieuw te helpen herinneren welke elementen in een goede hypothese thuishoren kunnen een aantal strategieën toegepast worden. De voorbeeldhypothese, die in dit proefschrift alleen als vraag in de werkbladen opgenomen was, zou op verschillende manieren kunnen worden aangeboden. Ook kunnen kinderen gestimuleerd worden om preciezer te zijn in hun hypothesen, door middel van extra opdrachten op de werkbladen maar ook in gesprek met de leerkracht. Een goede hypothese lijkt overigens erg op een goede conclusie, in die zin dat in beiden aandacht zou moeten zijn voor de onderzochte variabelen, het verwachte of geobserveerde effect en de richting van dat effect. De bovengenoemde suggesties gelden dus ook voor het formuleren van conclusies.

Experimenteren is veruit de meest onderzochte deelvaardigheid, maar niet voor alle leerlingen erg lastig om te leren. Bovendien scoren kinderen die eenmaal kunnen experimenteren consistent hoog op deze vaardigheid. Toch is het misschien niet optimaal om meteen vanaf het begin te differentiëren. In twintig minuten kunnen de meeste kinderen leren experimenteren, en uit eerder onderzoek (Lorch et al., 2017) blijkt dat kinderen die op deze manier leren experimenteren dat jaren later nog steeds kunnen. Het lijkt dus zinvol om deze korte instructie klassikaal aan te bieden, en daarna in de gaten te houden of er kinderen zijn die extra ondersteuning nodig hebben.

Het interpreteren van data blijkt voor alle kinderen moeilijk te zijn. Hoewel er geen direct beschikbare interventies zijn, blijkt uit eerder onderzoek dat sommige data makkelijker te interpreteren zijn dan andere data (Masnick & Morris, 2008; Piekny & Maehler, 2013). Omdat de kwaliteit van het experiment grote invloed heeft op de kwaliteit van de data die verzameld wordt, is het dus belangrijk om na te gaan of kinderen goed kunnen experimenteren. Daarnaast is het zo dat sommige experimenten, ook als ze goed

---

<sup>2</sup> De variabele die je meet, bijvoorbeeld hoe vaak een bal stuitert.

<sup>3</sup> De variabele die je verandert, bijvoorbeeld van welke hoogte je de bal laat vallen.

worden uitgevoerd, relatief lastig interpreteerbare uitkomsten hebben. Het experiment over de slingertijd is hier een goed voorbeeld van: enkel de lengte van een slinger maakt uit voor de slingertijd. Als je zou onderzoeken of het gewicht dat aan de slinger hangt ertoe doet, bijvoorbeeld door twee even lange slingers met verschillende gewichten naast elkaar te hangen en te kijken hoe lang elk doet over één slingerbeweging, zullen beiden dezelfde slingertijd hebben. Een dergelijk 'gebrek aan effect' is lastiger te interpreteren dan wanneer er juist wel een duidelijke relatie is tussen de invoervariabele en de uitkomstvariabele. Het kan dus nuttig zijn om eerst experimenten met relatief gemakkelijk interpreteerbare uitkomsten aan te bieden, en pas later experimenten met lastiger interpreteerbare uitkomsten.

## Referenties

- Chinn, C. A., & Golan Duncan, R. (2018). What is the value of general knowledge of scientific reasoning? In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific Reasoning and Argumentation* (pp. 77-101). Routledge.
- Edelsbrunner, P. A., Schalk, L., Schumacher, R., & Stern, E. (2018). Variable control and conceptual change: A large-scale quantitative study in elementary school. *Learning and Individual Differences, 66*, 38-53. <https://doi.org/10.1016/j.lindif.2018.02.003>
- Greven, J., & Letschert, J. (2006). *Kerndoelen primair onderwijs* [Curricular goals for primary education]. Ministerie van Onderwijs, Cultuur en Wetenschap.
- Kind, P. M., & Osborne, J. (2017). Styles of scientific reasoning: A cultural rationale for science education? *Science Education, 101*(1), 8-31. <https://doi.org/10.1002/sce.21251>
- Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning. *Review of Educational Research, 86*(3), 681-718. <https://doi.org/10.3102/0034654315627366>
- Lorch, R. F., Lorch, E. P., Freer, B., Calderhead, W. J., Dunlap, E., Reeder, E. C., Van Neste, J., & Chen, H.-T. (2017). Very long-term retention of the control of variables strategy following a brief intervention. *Contemporary Educational Psychology, 51*, 391-403. <https://doi.org/10.1016/j.cedpsych.2017.09.005>
- Masnack, A., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development, 79*(4), 1032-1048. <https://doi.org/10.1111/j.1467-8624.2008.01174.x>
- Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology, 31*(2), 153-179. <https://doi.org/10.1111/j.2044-835X.2012.02082.x>
- Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching, 41*(5), 513-536. <https://doi.org/10.1002/tea.20009>
- Trilling, B., & Fadel, C. (2009). *21st century skills - learning for life in our times*. Jossey-Bass.
- 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>





# Dankwoord

Hoewel mijn naam op de kaft van dit proefschrift staat, zijn er ook een heleboel mensen die ik wil bedanken voor hun bijdrage. En dat begint natuurlijk bij mijn promotoren. Ard en Inge, al tijdens mijn sollicitatie hadden we een goed gesprek over hoe onderzoek naar adaptiviteit in wetenschaps- en technologieonderwijs eruit zou kunnen zien en ik ben blij dat we dat gesprek in de afgelopen jaren zo prettig hebben kunnen voortzetten. De beroemde PhD-dip heb ik nooit gehad, en dat is grotendeels aan jullie te danken. Er waren natuurlijk wel tijdelijke dipjes, bijvoorbeeld toen ik door corona een dataverzameling moest afbreken, maar daar hielpen jullie me dan snel weer uit door alle mogelijkheden te schetsen die er óók nog waren. Ard, jij wil net als ik de details goed hebben. Daardoor voelde ik de ruimte om nauwkeurig te werken. Ik hoop dat ik door al je zorgvuldige tekstcorrecties een betere schrijver ben geworden. Inge, jij ziet altijd het grote plaatje en kijkt net vanuit een andere hoek. Ook dat heeft veel invloed gehad op hoe ik schrijf. Jouw Adaptive Learning Lab was bovendien, zeker in coronatijd, mijn thuisbasis op de RU.

Ook de manuscriptcommissie wil ik bedanken, voor de tijd die jullie hebben gestoken in het zorgvuldig lezen en beoordelen van mijn proefschrift en positieve reacties daarop.

Onderwijsonderzoek kan natuurlijk niet zonder de hulp van scholen, leerkrachten en leerlingen. In totaal deden er ruim 30 klassen vol enthousiasme mee aan mijn onderzoek! Daarbij waren meer leerkrachten betrokken dan ik op kan noemen, die mij hun lessen lieten overnemen voor pilots of onderzoek, maar ook het onderzoek gepromoot hebben bij hun collega's, zich in hebben gehouden als kinderen om hulp vroeg, en in coronatijd zelfs metingen hebben afgenomen zodat ik minder vaak naar school hoefde te komen. Zonder jullie had dit proefschrift er niet gelegen.

Van 2017 tot 2020 ben ik met veel plezier bijna elke werkdag naar Nijmegen gekomen, en dat is voor een groot deel te danken aan mijn fijne collega's. Het was altijd gezellig om samen te lunchen of even bij te praten als jullie tegenover mijn bureau op de printer stonden te wachten. Een aantal collega's wil ik in het bijzonder bedanken.

Hilde, wat was het fijn om de eerste jaren met jou een kamer te delen. We hadden samen precies een goede balans tussen focus en kletsen, en ook met serieuze onderwerpen kon ik

altijd bij je terecht. Marije, Lisa en Sascha, dankzij jullie was de terugreis naar de Randstad altijd gezellig. Carolien, toen we niet meer naar kantoor konden heb ik veel gehad aan onze belletjes over alle soorten promotie-struggles die je maar kan bedenken. Anne, Jolique, Rianne, Rebecca en Rick, de focus in onze corner office was soms ver te zoeken maar dat was misschien vooral omdat het daar zijn een welkome afwisseling van het thuiswerken was.

Noortje, dankzij het onderzoek waar jij met Ard al langer aan werkte had mijn promotietraject een vliegende start. Joep, jij had altijd de juiste kritische vragen als ik tussentijds mijn werk presenteerde. Bovendien gaf je mij het zetje om te beginnen aan een latente transitie-analyse, die uiteindelijk misschien wel de leukste resultaten uit dit proefschrift heeft opgeleverd. Jo, wat fijn dat ik jou tegen het einde van mijn promotietraject heb leren kennen. We worden enthousiast van dezelfde dingen, en ik hoop dat we elkaar nog vaak tegen gaan komen!

Jolique, wat fijn dat je één van mijn paranimfen bent. Voor wat extra gezelligheid kan ik altijd op je rekenen, maar je bent ook een onwijs harde werker en een doorzetter. Onze tweepersoons corona-schrijfweek daardoor gezellig en ook nog eens super-effectief. Anne, ook met jou als paranimf ben ik ongelooflijk blij. Ik vind het altijd fijn om met je samen te werken, vooral aan de creatieve projecten rondom ons onderzoek. We zitten vaak op hetzelfde spoor, maar wel op zo'n manier dat we echt verder komen dan wanneer ik in m'n eentje was doorgekacheld. Het is echt fijn om zo'n collega te hebben!

Als er op de universiteit alléén maar wetenschappers zouden werken, zou de helft van het onderzoek niet eens afkomen en de andere helft zou van beduidend lagere kwaliteit zijn. Katja, Lanneke, Christel en Lonneke, ik weet nog steeds niet zeker bij wie ik waarvoor moet zijn maar jullie gelukkig wel! Bedankt voor al jullie praktische en emotionele ondersteuning op de afdeling. Ook de Technical Support Group heeft een belangrijke bijdrage geleverd aan mijn onderzoek. Wilbert, bedankt voor het programmeren van de superflexibele Toren van Hanoi-taak die ik in mijn eerste onderzoek gebruikt heb. Sibrecht, nadat ik in mijn tweede onderzoek een provisorisch duikplankproefje gebruikt had, heb jij voor mijn derde onderzoek de duikplankjes uitgewerkt tot prachtig en volwaardig lesmateriaal.

Na twee pagina's aan collega's is het misschien tijd om ook even te vermelden dat er buiten het werk een hoop mensen zijn, die ik tot vervelens toe over mijn onderzoek op de hoogte heb gehouden en die toch nog steeds met me om willen gaan. Ik bof enorm met mijn vrienden en vriendinnen. Jullie zijn stuk voor stuk zorgzame, lieve mensen van wie ik altijd nieuwe dingen leer. Ik ga niet iedereen opnoemen omdat ik bang ben dat ik iemand vergeet, maar er zijn wel een paar mensen die ik specifiek wil bedanken.

Catheleyne en Aniek, wat was het gezellig om naast collega's ook wat vrienden in Nijmegen te hebben! Tamara, hoewel we elkaar niet vaak meer zien is het altijd leuk om je te spreken over wetenschaps- en technologieonderwijs. Bovendien heb je me enorm geholpen met het vinden van scholen voor mijn laatste onderzoek – bedankt! Reema, jij hebt een taak op je genomen die bijna niemand anders had kunnen doen: dankzij jou wist ik dat mijn informatiebrief óók voor mensen die nog niet zo lang Nederlands spreken goed te begrijpen is. Alexandra, op de middelbare school waren wij al onderwijskundigen in de dop en de commissies waar we ons toen mee bemoeid hebben bleken voorspellend voor de richting die we uiteindelijk gekozen hebben. Wat fijn dat je nu mijn bijna-buurvrouw bent! Evi, ik ben blij dat we het altijd weer gezellig hebben met elkaar – ook al is het iets te veel biertjes drinken inmiddels veranderd in wandelen, koken en boulderen. Suzan, jij leert me elke keer weer dat het oké is om emotioneel te zijn. GA-meisjes, bedankt dat ik, hoewel ik nog nooit in mijn leven een aflevering van Grey's Anatomy heb gezien, toch bij jullie mag horen.

Ulrika, jouw huis voelt als mijn tweede thuis en dat is me zoveel waard. Je bent slim, zorgzaam, zorgvuldig en creatief, en ik ben heel blij met jou als vriendin. Wat fijn dat we elkaars paranimfen konden zijn!

Papa en mama, jullie hebben me altijd gesteund in mijn keuzes (en dat waren er nogal wat), me laten zien dat het niet erg is om het oneens te zijn, en bovenal mijn liefde voor het leren van nieuwe dingen ondersteund. Het is dan ook dankzij jullie dat ik me zo thuis voel in de academische wereld. Jan en Henk, jullie zijn fantastische, superslimme en zorgzame broers. Hoewel jullie allebei met heel andere dingen bezig zijn dan ik, zijn jullie ook altijd geïnteresseerd in wat mij bezighoudt. Een eervolle vermelding voor Henk kan niet

achterwege blijven: toen jij de Toren van Hanoi-taak voor me testte was jij de enige die 'm binnen de tijd oploste, inclusief het laatste, extra moeilijke probleem<sup>1</sup>.

In de loop van mijn leven heb ik ook wat extra familie opgedaan. Letje, het alfa-vrouwje van de familie, wat ben ik ongelooflijk blij dat we je niet hebben weggejaagd door bij het avondeten te praten over de stelling van Pythagoras. Je bood een welkom tegenwicht voor al het bètageweld bij ons in huis en zorgde altijd voor vrolijke opschudding. Tot het laatste moment was je onvoorwaardelijk geïnteresseerd en betrokken. Ik ben blij en dankbaar dat ik je gekend heb.

Toen Maarten me ruim tien jaar geleden vroeg of hij me op zijn verjaardag voor mocht stellen als zijn vriendin, was ik het meest zenuwachtig voor de ontmoeting met zijn zussen. Dat bleek volkomen onterecht: jullie hebben me meteen welkom geheten in de familie. En Hubert en Wilma, wat is het leuk om jullie in de zomer tegen te komen op het IJsselmeer, maar ook om met jullie over onderwijs te praten. Fijnere schoonouders kan ik me niet wensen.

De belangrijkste persoon komt natuurlijk als laatste. Maarten, bij jou voel ik me thuis en veilig. Je bent zorgzaam en gezellig, maar ook principieel en dapper. Je zorgt voor me en leert me voor mezelf te zorgen. Ik heb nu al zin in alle avonturen die we nog samen gaan beleven, want als tweepersoons gezin kunnen we alles aan.

---

<sup>1</sup> [In het digitale proefschrift kan je hier klikken om 't ook te proberen!](#)

# Author biography

After graduating from secondary education, Erika Schlatter (Amsterdam, 1988) attended the University of Amsterdam. Here, she explored General Social Sciences as well as Artificial Intelligence before landing on Educational Sciences. After obtaining her bachelor's degree with honours in 2014, Erika was admitted to the Research Master of Child Development and Education. During this program, she participated in a research project on children's collaborative learning during mathematics lessons and designed and conducted research on qualitative modelling in secondary education. During her studies, Erika worked as a homework tutor for secondary school students, as a sports instructor in snowboarding and sailing, and as an educationalist at a company providing programming lessons to primary school children. After graduating in 2017, Erika started her PhD project at Radboud University. In this project, she studied differences in scientific reasoning between children, as well as across various scientific reasoning skills. She presented her work at various international conferences. Currently, she works as a post-doctoral researcher at Leiden University, where she studies children's adaptive expertise in mathematics.

# Publications

Schlatter, E., Bredeweg, B., Drie, J. V., & Jong, P. D. (2017, September). Can learning by qualitative modelling be deployed as an effective method for learning Subject-Specific Content? In *European Conference on Technology Enhanced Learning* (pp. 479-485). Springer, Cham.

Schlatter, E., Molenaar, I., & Lazonder, A. W. (2020). Individual differences in children's development of scientific reasoning through inquiry-based instruction: Who needs additional guidance? *Frontiers in Psychology, 11*, Article 904.

<https://doi.org/10.3389/fpsyg.2020.00904>

Schlatter, E., Lazonder, A. W., Molenaar, I., & Janssen, N. (2021). Individual differences in children's scientific reasoning. *Education Sciences, 11*(9), Article 471.

<https://doi.org/10.3390/educsci11090471>

Schlatter, E., Molenaar, I., & Lazonder, A. W. (2021). Learning scientific reasoning: A latent transition analysis. *Learning and Individual Differences, 92*, Article 102043.

<https://doi.org/10.1016/j.lindif.2021.102043>

Schlatter, E., Molenaar, I., & Lazonder, A. W. (resubmitted). Adapting scientific reasoning instruction to children's needs: Effects on learning processes and learning outcomes.