European Journal of Science and Mathematics Education
Vol. 8, No. 3, 2020 , 137-144

How to measure scientific reasoning in primary school: A comparison of different test modalities

Kristin Nyberg[1], Susanne Koerber[1], and Christopher Osterhaus[2]
[1] Freiburg University of Education, Department of Psychology, Germany
[2] Ludwig-Maximilians-University Munich, Department of Psychology, Germany
For correspondence: kristin.nyberg@ph-freiburg.de

Abstract:
Investigating scientific reasoning comprehensively with large sample sizes is a challenge, especially in young children where there is little evidence showing that paper-and-pencil instruments can be used to reliably measure competencies. The present study with 122 third graders compares three test modalities: an established interview (Science-K Interview), a novel group test (Science-K Group Test), and an established paper-and-pencil test (Science-P Reasoning Inventory), asking how scientific reasoning can be reliably measured in primary school. The children were presented with 15 items from the Science-K Interview, 15 parallel items from the Science-K Group Test, and 5 items from the Science-P Reasoning Inventory. The interview was presented first, followed by the group test and Science-P items. The Rasch model revealed a good fit of 23 items to the unidimensional model indicating that there is a single underlying ability involved in primary-school scientific reasoning that can be measured with both interview and group test. Although students performed significantly better in the group test than in the interview, the difference was not substantial and might partially be explained by the sequence of the tests. As expected, children solved more Science-K items correctly compared to the Science-P items, which assesses more advanced aspects of scientific reasoning. The important finding of the study is that group tests can reliably capture scientific reasoning in third graders. The use of group tests in primary school can facilitate conducting studies with large sample sizes and can be used for diagnostic purposes, especially in the context of science education in schools.

Keywords: scientific reasoning, primary school, group test, interview, test modalities, science education

Introduction

For several decades, researcher investigate the increasingly important ability of scientific reasoning in the cognitive development of children (Bullock & Ziegler, 1999; Carey, Evans, Honda, Jay, & Unger, 1989; Inhelder & Piaget, 1958; Klahr & Dunbar, 1988; Klopfer, 1969; Kuhn, 2010; Sodian, Zaitchik, & Carey, 1991). Scientific reasoning is considered important not only in the context of cognitive development, but also in its relation to science learning (e.g. Chen & Klahr, 1999; Lederman, Lederman, & Antink, 2013). Basic scientific-reasoning skills are present in early years (Piekny & Maehler, 2013; Samarapungavan, Mantzicopoulos, Patrick, & French, 2009), but development of the competence remains incomplete even by the end of secondary school (Bullock, Sodian, & Koerber, 2009; Kuhn & Pease, 2008). To understand and explain children's emerging abilities to reason scientifically and the relevance of such abilities to future course and career choices, comprehensive instruments and studies with large sample sizes are needed. Scientific reasoning is a complex construct and includes several components such as experimentation, data interpretation, and understanding the nature of science (for a review see Zimmermann, 2007). For many years researchers in numerous disciplines such as developmental and cognitive psychology and science education examined single subcomponents of scientific reasoning (e.g. Abd-El-Khalick, Lederman, Bell, Schwartz, 2002; Aikenhead, 1973; Koerber, & Sodian, 2009; Lawson, 1978).

Bullock und Ziegler (1999) for instance showed that primary school children have developed basic skills in experimentation, demonstrating that primary school students can select a controlled over a confounded experiment as the better test of their hypothesis. Kindergarteners seem to have basic

skills of experimentation: e.g., they can arrange unconfounded experiments with a limit of four variables (van der Graaf, Segers, & Verhoeven, 2015). Regarding data interpretation, another relevant component of scientific reasoning, primary school students show skills in interpreting data (Koerber, Mayer, Osterhaus, Schwippert, & Sodian, 2015) and kindergarteners can interpret simple patterns of covariation data (Koerber, Sodian, Thoermer, & Nett, 2005) and conclude from data presented in graphs (Koerber & Sodian, 2009). Additionally, primary-school children show basic conceptual understanding of hypothesis testing and evidence evaluation (Akerson & Donnelly, 2010). Koerber, Osterhaus et al. (2015) supported the findings and showed that basic skills of understanding the nature of science exist in 9-year-old children. Competencies in young children concerning their views of science are also found in the young children's views of science (CVS), which assess the views of the nature of science and nature of scientific inquiry orally in interview sessions due to limited reading and writing skills (Schwartz, Lederman, Lederman, 2008).

There are few scales to measure scientific reasoning in young children and most of the existing instruments do not measure scientific reasoning in more than one component. To measure skills comprehensively in primary school children, Koerber, Mayer et al. (2015) developed illustrated paper-and-pencil items with simple language to simultaneously assess three different aspects (experimentation, data interpretation and nature of science) of scientific reasoning in primary school children. The study investigated 1,581 second, third, and fourth graders. A Rasch model showed good fit of the data, which can be interpreted as evidence for a common conceptual core. The idea is that there is a common core, which is characterized by the understanding of the hypothesis-evidence relation (Kuhn, 2010). Furthermore, the results suggest a continuous development of scientific reasoning in primary school; the fourth graders performed significantly better than the second graders. To assess scientific reasoning skills of even younger children and enable longitudinal measurement from kindergarten through primary school, Koerber and Osterhaus (2019) developed an instrument (Science-K Inventory) with the purpose of measuring existing reasoning skills comprehensively in preschoolers. The Science-K Inventory comprises the three different components: Experimentation, data interpretation and nature of science. A one-dimensional Rasch model revealed good item-fit indices of all 30 items to the model, which strengthens the assumption of a common conceptual core comprised in scientific reasoning. The scale was validated on 227 six-year-old preschoolers and showed good reliability, thus establishing Science-K Inventory as a reliable and comprehensive measurement of scientific reasoning skills in preschool children. The 30-item scale was assessed by one-on-one interviews in five sessions. The students solved on average 42,5% of the scientific reasoning items correctly, indicating no ceiling effect and the opportunity to use the instrument further on in primary school (Koerber & Osterhaus, 2019).

Up to primary school, interviews seem to be the means of choice for measuring scientific reasoning skills. This is also based on a long tradition of childhood research and developmental psychology. Piaget (1978) for example emphasized the importance of interviews to investigate children's knowledge and ideas about the world (clinical interviews). Other reasons for the frequent use of interviews in young children could be insufficient reading abilities (which are usually necessary for paper-and-pencil tests), and the distraction by classmates during classroom testing. However, the question that arises is whether a group test is possible once cognitive skills develop. Especially for the often-required large-scale studies, paper-pencil tests (group tests) are preferred for time and cost reasons. Because of this, Town (1922) argued early for putting effort and attempts into the construction of valid group tests for children in kindergarten and primary school. However, there are important constraints to consider in the construction of group tests for young children. It is important to choose clear language and an appealing design of the questionnaire, depending on the target group (Mey, 2003). In addition, the objectivity of the test has to be ensured; this includes simple and standardized testing and test analysis (Lüke, Ritterfeld, & Tröster, 2016).

The main aims of the present study are therefore to compare the Science-K Interview with the Science-K Group Test and to test the suitability of the Science-K Inventory for measuring scientific reasoning in third graders. The result might provide a further test (modality) for the comprehensive and reliable measurement of scientific reasoning skills in primary school and could promote studies with large sample sizes. In a further step, the performance in the Science-K Inventory is compared to

that of the Science-P Reasoning Inventory, which measures a higher level of competence. If both group tests measure scientific reasoning reliably at the end of primary school it would give rise to the possibility to use group test for large-scale studies and for diagnostic purposes in science education. Third graders were selected as the sample; the results of the test difficulties among six-year-olds suggest that the test in primary school can still sufficiently differentiate (Koerber & Osterhaus, 2019) and because other studies indicate that group tests can be used at this age to reliably measure scientific reasoning (Koerber, Osterhaus et al., 2015; Osterhaus et al., 2017).

Methods

Participants

The sample consisted of N= 122 third graders (55 boys and 67 girls; M= 9 years 7 months; SD= 3,6 months) of 16 primary schools in Germany. The testing took place at the end of third year of primary school. Written consent was obtained from the parents of all children.

Materials

Science-K Inventory (interview and group test). The Science-K Inventory consists of 30 items with ten items each in experimentation (EXP), data interpretation (DAT) and nature of science (NOS), and assesses beginning competences of scientific reasoning. In each item children are provided with a short story and three possible answer options (one correct and two distractors), so that a maximum sum score of 30 points can be attained (1 point per item). NOS items assessed children's ideas of what scientists do and which kind of questions they ask. These items tested children's basic understanding of scientists as looking for explanations about the world. DAT items referred to children's interpretation of covariation patterns and their understanding of confounded data. Experimentation items assessed children's performance in the control of variables strategy and investigated their understanding of a conclusive test. For instance, in an experimentation item, children are told that Tom wanted to find out if Mia was good at doing puzzles. The participants were asked: "what should Mia do to find this out? a) piece together her favorite puzzle, b) piece together a puzzle with few puzzle pieces, or c) piece together a puzzle with many puzzle pieces (correct)". For a more detailed description of the items, see Koerber and Osterhaus (2019). The Science-K Interview and the Science-K Group Test measure scientific reasoning with 15 items each (five items per subskill).

Science-P Reasoning Inventory (group test). From the Science-P Reasoning Inventory (Koerber, Mayer et al., 2015), which measures scientific reasoning at the primary school level, five representative tasks were selected. In each task, three answer options are presented to the children (on a naïve level = 0 points, an intermediate level = 1 point, a scientifically advanced level = 2 points) and they have to agree or disagree with each of the answer options. The lowest level answer selected is taken as the final score on the entire item, so that children can obtain a maximum of 2 points per task, and a maximum sum score of 10 (see Koerber, Osterhaus et al., 2015 for further coding details). Two items each assessed EXP and NOS and one item assessed DAT. Similar to the Science-K Inventory, the two NOS items used from the Science-P Reasoning Inventory assessed children's understanding of what scientists do. The two experimentation items assessed performance in the control of variables strategy and the understanding of a conclusive test. In this latter item, for instance, children are asked to evaluate options, testing whether an aardvark has a good sense of smell. The options were attracting the aardvark with a) bread in a box (advanced), b) cheese on a box (naïve), or c) bread on a box (intermediate). In the data interpretation item, children's understanding of confounded data was tested by having them compare two fictitious data sets, which showed the effects of herbs on aliens' health, and evaluate the interpretations of the (confounded) data. For a more detailed description of the items see Koerber, Mayer et al. (2015) and Osterhaus, Koerber and Sodian (2020).

Text comprehension. As a control variable, text comprehension was assessed with a reading proficiency test for first to sixth graders (ELFE 1-6; Lenhard & Schneider, 2006). The test comprises of 20 multiple-choice items. The student received short story texts with a related question. The task was to choose the best of four given answers. Cronbach's alpha of the 20 items was excellent with .87.

Procedure

Half of the Science-K Inventory was tested in a one-on-one interview session. The other half of the Science-K Inventory and all of the Science-P items were tested in whole-class testing procedure (group test). In each modality the items were presented in an illustrated booklet. The items of the group test were read out by an experimenter and presented in a power point presentation. All students received the Science-K Interview first, followed by the Science-K Group Test. Each testing session lasted approximately 30 minutes.

Results

Scale Analysis: Science-K Inventory in third grade

To test whether the Science-K Inventory results in a reliable measurement of scientific reasoning in primary school children, we fitted a one-dimensional Rasch model to the data, using ACER ConQuest 2.0. (Wu, Adams, Wilson, & Haldane, 2007). Seven of the 30 items revealed poor infit and outfit mean squares (< 0.8 or > 1.2). The remaining 23 items indicated a good fit to the model (see Table 1). This finding supported the idea of unidimensionality. The reliability of the 23-item scale was in an acceptable range with a Cronbachs alpha of = .59, a maximum likelihood estimate (MLE) person-separation reliability = .56, a weighted likelihood estimate (WLE) person-separation reliability = .53 and an expected a posteriori estimate based on plausible values (EAP/PV) reliability = .57. A composite score was calculated for the 23-item scale. This composite score is used in subsequent analyses.

Core performance

Science-K Inventory. The distribution of the performance was skewed to the left. On average, the third graders answered 74.20% of the 23 Science-K Items correctly (SD= 12.77%). Separated according to the Science-K Group Test and the Science-K Interview, the children answered 78.84% (SD= 15.10%) of the items correctly in the group test and 71.44% (SD=15.18%) correctly in the interview. The mean score of the Science-K Group Test items, which was conducted in all children after the interview, differed significantly, but not substantially ($t(121) = 4.56$, $p = 0.000$) from the interview items. According to Cohen (1992) is it a very weak effect of $r = .03$. The correlation between the Science-K Group Test and the Science-K Interview was $r = .33$ ($p = 0.00$).

Science-P Reasoning Inventory. Regarding the Science-P items, students scored 28.61% (SD= 12.49%) of the items on the naïve level, 41.96% (SD= 9.34%) on the intermediate level and 29.18% (SD= 18.02%) on the advanced level. Across the five items, the children obtained a mean score of 5.06 (SD= 1.95; min= 0; max=9), which means that the students scored 50.58% (SD= 19.58%) of the possible points on average. To compare it to the performance in the Science-K Inventory, the performance in the Science-P Reasoning Inventory could be analysed dichotomously. Either agree with the advanced answer and disagreeing with the naïve and intermediate answer is counted as correct or both the advanced and intermediate answer is taken as correct (see Figure 1).

Table 1. Item fits for the Science-K Inventory

Tasks per component	Aspect	Percentage correct	Difficulty	Discrimination	Infit MNSQ	OutfitMNSQ
1 EX1 dog_group test[1]	Conclusive test	87.60%	-1.009	0.63	0.85	0.61
2 EX2 puzzle_interview	Conclusive test	86.01%	-0.521	0.26	1.03	0.98
3 EX3 soccer_group test	Conclusive test	70.73%	0.242	0.45	0.94	0.93
4 EX4 songs_interview	Conclusive test	50.00%	1.525	0.38	0.99	1.00
5 EX5 plants_group test[1]	CVS	90.98 %	-0.750	0.62	0.85	0.65
6 EX6 density_interview	CVS	86.88%	-0.582	0.21	1.05	1.05
7 EX7 turtle_group test[1]	CVS	95.90%	-1.402	0.59	0.88	0.59
8 EX8 milk_interview	CVS	86.34%	-0.835	0.27	1.01	0.89
9 EX9 slug_group test[1]	CVS	94.26%	-1.368	0.59	0.87	0.51
10 EX10 eggs_interview	CVS	93.44%	-1.276	0.17	1.04	0.91
11 DAT1 chewing gum_interview	Covariation data (perfect)	91.80%	-1.275	0.19	1.03	0.93
12 DAT2 juice_interview	Covariation data (imperfect)	44.26%	1.734	0.17	1.12	1.18
13 DAT3 bad teeth_interview	Covariation data (non-covariation)	90.89%	-1.044	0.35	0.97	0.98
14 DAT4 bad teeth_interview	Confounded data	35.24%	2.203	0.29	1.03	1.06
15 DAT5 healthy_interview	Confounded data	41.80%	1.857	0.21	1.10	1.12
16 DAT6 new trousers_group test	Confounded data	62.57%	0.788	0.48	0.94	0.92
17 DAT7 happy_group test	Confounded data	63.03%	0.705	0.28	1.04	1.10
18 DAT8 handkerchiefs_group test	Covariation data (perfect)	80.42%	-0.217	0.40	0.95	1.06
19 DAT9 nasal spray_group test	Covariation data (imperfect)	60.29%	0.930	0.31	1.07	1.09
20 DAT10 flowers_group test[1]	Covariation data (non-covariation)	91.80%	-1.00	0.64	0.85	0.58
21 NOS1 hedgehog_group test	NOS (do)	74.86 %	0.143	0.51	0.91	0.85
22 NOS2 colour_group test	NOS (do)	87.62%	-1.690	0.33	0.95	0.86
23 NOS3 leaves_group test	NOS (do)	59.93%	0.949	0.42	0.98	0.98
24 NOS4 ladybird_interview[1]	NOS (do)	100%	-	-	-	-
25 NOS5 swimming_interview	NOS (do)	77.86%	0.109	0.28	1.06	1.10
26 NOS6 weather_interview[1]	NOS (ask)	42.62%	1.845	0.09	1.15	1.23
27 NOS7 flowers_group Test	NOS (ask)	79.42%	-0.128	0.43	0.96	0.95
28 NOS8 stream_group Test	NOS (ask)	84.34%	-0.715	0.47	0.92	0.90
29 NOS9 universe_interview	NOS (ask)	66.39%	0.730	0.13	1.12	1.14
30 NOS10 sea_interview	NOS (ask)	78.68%	0.055	0.18	1.08	1.06

Note,[1]Items excluded due to poor infit and outfit mean squares.

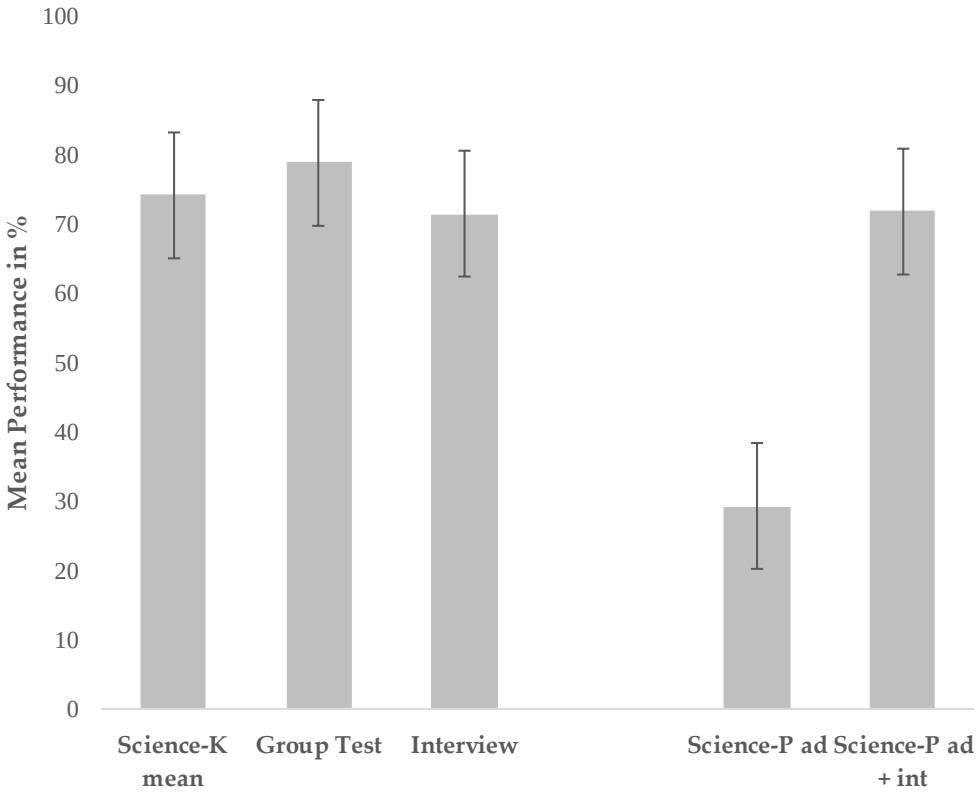

Figure 1. Mean performance in the Science-K Inventory (separately for interview and group test) and Science-P Reasoning Inventory. Science-P ad = only advanced answer is counted as correct and naïve and intermediate answers must be rejected; Science-P ad+int = advanced and intermediate answers are counted as correct and naïve answer must be rejected.

Control variable

Correlation between the performance in the Science-K and ELFE 1-6. To test whether reading abilities are particularly relevant for one or both test modalities (the Science-K Interview and Group Test) the correlations between the scientific-reasoning and reading test scores were calculated. The performances in the Science-K and the ELFE 1-6 correlated significantly with r= .28 (p= < .01). The correlation between the performance in the Science-K Interview items and ELFE 1-6 was *r*= .25 (p= < .01) and between the performance in the group test and ELFE 1-6 *r*= .21 (p= < .01). As expected, the correlations did not significantly differ (p= .37).

Science-K Inventory and Science-P Reasoning Inventory

Correlation. Performance on both tests correlated significantly *r*= .44 (p=0.000), revealing according to Cohen (1992) a medium to high effect. The correlation was *r*=. 39 when controlling for text comprehension, also a medium to high effect (Cohen, 1992).

Hierarchical Regression with Science-K and ELFE. The model showed an impact of the performance in the Science-K inventory on the performance in the Science-P items F(1,116) = 15.38 p= .000 and explained 19.80% of the variance. The model with the performance in Science-K Inventory and with ELFE showed a significant impact F(2,115) = 18.95 p= .000 and explained 24.77% variance of the performance in the Science-P items.

Discussion

The paper compares the Science-K Interview with the Science-K Group Test to assess scientific reasoning in primary school and further investigates the suitability of the Science-K Inventory for this purpose. One main result is that both interview and group tests can be used to comprehensively measure scientific reasoning skills in third graders. This is backed by the result of the Rasch model; items of both modalities (group test and interview) contribute to the good fit of the data to the

unidimensional model. Moreover, the fits of the remaining 23 items (from both Science-K Interview and Science-K Group Test) indicate the reliable measurement of scientific reasoning skills in third graders. The performance in the group test and interview differed significantly, but not substantially in favour of the Science-K Group Test. According to Cohen (1992) the difference represents a small effect. Guessing can be ruled out as an explanation for this difference in performance because the Science-K Interview presented children, just like the Science-K Group Test, with three answer options. The most likely explanation is a testing effect; the children always received the Science-K Group Test after the Interview, which—although no feedback was given—may have resulted in an increased familiarity with the content and a higher performance in the group test.

A further argument for the use of both modalities is the similarly strong relation of the interview and group test to the language measure. Like other studies in the area, this study showed a correlation between the text comprehension task and scientific reasoning (Koerber, Mayer et al., 2015; Koerber, Osterhaus et al., 2015; Mayer et al., 2014). Interestingly, both the group test and the interview correlate similarly strong with the text comprehension task, indicating that scientific reasoning is in general linked to language, independent of the modality. The result reinforces our assumption that reading ability plays no more important a role in the group test than in the interview, and suggests that the power point presentation has successfully eliminated any reading difficulties. A second outcome is that the Science-K Inventory allows a reliable and comprehensive measurement of scientific reasoning skills in third graders. Furthermore, the result of the scale analysis suggests a common underlying ability for 23 items of the Science-K Inventory in third graders. This is in line with prior research, which modelled scientific reasoning as a unidimensional ability (Koerber, Mayer et al., 2015; Koerber & Osterhaus, 2019). In comparison to the Views of Nature of Science (VNOS) (Abd-El-Khalick, Lederman, Bell, Schwartz, 2002) and CVS (Schwartz et al., 2008), both the Science-K Interview and the Science-K Group Test allow comprehensive measurement in a wider array of scientific reasoning skills in primary school. The insights into the subskills of the Science-K Inventory are not as deep as in the VNOS or CVS, but the Science-K Group Test could give an overview of existing skills in scientific reasoning and could be used for diagnostic purposes in primary school. The major advantage of a group test is simultenaous testing of a whole class, enabling larger sample sizes.

As expected, performance of the third graders was superior to that of kindergarteners (Koerber & Osterhaus, 2019), indicating that scientific reasoning is an ability that improves over time. Especially in this context, it is relevant whether the Science-K Inventory and the Science-P Reasoning Inventory measure scientific reasoning in a similar way, enabling the continuous measurement of scientific-reasoning skills in higher grades. Looking at the correlation (r= .44**) between the Science-K Inventory and the Science-P Reasoning Inventory, it can be assumed that both instruments measure similar concepts of scientific reasoning skills, but there are some differences which are important to consider when comparing the two measurements. The Science-K Inventory measures scientific reasoning skills in kindergarteners and hence capture basic skills of scientific reasoning. The Scince-P Reasoning Inventory is based on a conceptual model (Koerber, Mayer et al., 2015; Osterhaus et al., 2020; Sodian, 2018) and can assess a higher competence level of scientific reasoning using items that are more difficult and designed for older children. Due to the competence levels of the Science-P items the analysis was conducted in a different way. The answers of the Science-P items were divided into competence levels (naïve, intermediate, or advanced). To compare the percentage of correct answers of the Science-P items to those of the Science-K Inventory, they need to be analysed dichotomously. If the advanced answer is used as the only correct answer, the difference between the performance between the Science-K Inventory (74%) and the Science-P Reasoning Inventory (29%) is large. If the performance in the Science-P items is analysed with both the intermediate and advanced answer marked as correct and only the naïve answer as incorrect, then the percentage of the correct answers between Science-K Inventory and Science–P Reasoning Inventory (71%) is similar. This result indicates that at primary school age the Science-K Inventory might measure an intermediate level and could act as a precursor skill for advanced scientific reasoning skills. The results of the Science-P Reasoning Inventory suggest accurate measurement of scientific reasoning across developing skills, and that it could be used for higher grades, especially when the students become better in scientific reasoning.

Taken together, the results support the use of an additional test modality, the Science-K Group Test, to comprehensively measure scientific reasoning skills in primary school children. The use of group tests in primary schools allow larger sample sizes for the often-required large-scale studies, and the results offer validity for the use of groups tests for diagnostic purposes in primary school. The Science-K Group Test takes almost the same time as the interview, but it allows testing about 25 students at the same time compared to one student in an interview session. The remaining problem within the Science-K Group Test is still the cost of materials and the associated waste of resources. For future studies, the possibility of digital testing should be discussed.

In conclusion, our study demonstrates that paper-and-pencil tests, both the Science-K Group Test and the Science-P Reasoning Inventory, are reliable means to measure third graders' scientific reasoning competencies and that, more importantly, group test formats are suitable for research on complex reasoning abilities in primary school.

Acknowledgements

We are grateful to all research assistants for their help in the data collection and to all teachers, children, and parents for their friendly collaboration and support of this research.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) by grant KO 2276/5-1. The article processing charge was funded by the Baden-Württemberg Ministry of Science, Research and Culture and the University of Education, Freiburg in the funding program Open Access Publishing.

References

Abd-El-Khalick, F., Lederman, N.G., Bell, R.L., & Schwartz, R.S. (2002). Views of Nature of Science Questionnaire (VNOS): Toward Valid and Meaningful Assessment of Learners' Conceptions of Nature of Science. *Journal of Research in Science Teaching, 39,* 497-521. doi:10.1002/tea.10034

Aikenhead, G. (1973). The measurement of high school students' knowledge about science and scientists. *Science Education, 57,* 359-349.

Akerson, V., & Donnelly, L. A. (2010). Teaching nature of science to K-2 students: What understandings can they attain? *International Journal of Science Education, 32,* 97–124. doi:10.1080/09500690902717283

Bullock, M., Sodian, B., & Koerber, S. (2009). Doing experiments and understanding science: Development of scientific reasoning from childhood to adulthood. In W. Schneider & M. Bullock (Eds.), *Human Development from Early Childhood to Early Adulthood: Findings from a 20 year Longitudinal Study*. (pp. 183–208). New York: Psychology Press.

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, UK: Cambridge University Press.

Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C. (1989). An experiment is when you try it and see if it works': A study of grade 7 students' understanding of the construction of scientific knowledge. *International Journal of Science Education, 11*, 514–529. doi:10.1080/0950069890110504

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 122*, 155–159.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books. doi:10.1037/10034-000

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12,* 1–48. doi:10.1207/s15516709cog1201_1

Klopfer, L.E. (1969). The teaching of science and the history of science. *Journal of Research in Science Teaching, 6*, 87-95.

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development, 86*, 327–336. doi: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,* 510–533. doi: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,* 57–72. doi:10.1111/bjdp.12067

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

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,* 141–152. doi:10.1024/1421-0185.64.3.141

Kuhn, D. (2010). What is scientific thinking and how does it develop? In U. Goswani (Ed.), *Handbook of Childhood Cognitive Development* (pp. 497–523). Oxford, UK: Blackwell.

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C, White, S.H, Klahr, D., & Craver, S.M. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development, 60*, 1–157. doi:10.1207/s15327647jcd0703_1

Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cognition and Instruction, 26*, 512–559. doi:10.1080/07370000802391745

Lawson, A.E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching, 15*, 11-24. doi:10.1002/tea.3660150103

Lederman, N.G., Lederman, J.S., & Antink, A. (2013. Nature of science and scientific inquiry as contexts for the learning of science and achievement of scientific literacy. *International Journal of Education in Mathematics, Science and Technology, 1,* 138-147. doi:10.46328/ijemst.v1i3.19

Lenhard, W., & Schneider, W. (2006). ELFE 1-6: Ein Leseverständnistest für Erst- bis Sechstklässler. Göttingen: Hogrefe.

Lüke, T., Ritterfeld, U., & Tröster, H. (2016). Erprobung eines Gruppentests zur Überprüfung des Grammatikverständnisses auf der Basis des TROG-D. *Diagnostica, 62*, 242–254. doi:10.1026/0012-1924/a000157

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. doi:10.1016/j.learninstruc.2013.07.005

Mey, G. (2003). *Zugänge zur kindlichen Perspektive: Methoden der Kindheitsforschung*. Forschungsbericht. Berlin: Technische Universität, Institut für Sozialwissenschaften, Abteilung Psychologie.

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,* 450–462. doi:10.1037/dev0000260

Osterhaus, C., Koerber, S., & Sodian B. (2020). The Science-P Reasoning Inventory (SPR-I): Measuring emerging scientific-reasoning skills in primary school. *International Journal of Science Education*. Advanced online publication. doi:10.1080/09500693.2020.174825

Piaget, J. (1978). D*as Weltbild des Kindes.* Stuttgart: Klett-Cotta.

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,* 153–179. doi:10.1111/j.2044-835X.2012.02082.x

Samarapungavan, A., Mantzicopoulos, P., Patrick, H., & French, B. (2009). The development and validation of the science learning assessment (SLA): A measure of kindergarten science learning. *Journal of Advanced Academics, 20,* 502–535. doi:10.1177/1932202X0902000306

Schwartz, R.S., Lederman, N.G, & Lederman, J.S. (2008). *An Instrument To Assess Views of Scientific Inquiry: The VOSI Questionnaire*, Paper presented at the meeting of the National Association for Research in Science Teaching (NARST), Baltimore, MD, USA.

Sodian, B. (2018). The development of scientific thinking in preschool and elementary school age. A conceptual model. In F. Fischer, C.A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation. The roles of domain-specific and domain-general knowledge* (pp. 227-250). New York: Routledge.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development, 62*, 753–766. doi:10.1111/j.1467-8624.1991.tb01567

Town, C. H. (1922). A mass mental test for use with kindergarten and first grade children. *Journal of Applied Psychology, 6*, 89–102.

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,* 381–400. doi:10.1007/s11251-015-9344-y

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Camberwell, Victoria: ACER Press.

Zimmermann, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*, 172–223. doi:10.1016/j.dr2006.12.001