

Chapter 3

Science-P II: Modeling Scientific Reasoning in Primary School

Susanne Koerber, Beate Sodian, Christopher Osterhaus, Daniela Mayer, Nicola Kropf, and Knut Schwippert

Abstract Basic scientific reasoning abilities in primary-school children have been documented in numerous studies. However, an empirically tested competence-structure model has not been developed, most likely due to the difficulty of capturing conceptual understanding in paper-and-pencil tasks. The Science-P project contributes to this research area by constructing and testing a theoretical model of the development of scientific reasoning in primary school. Based on our own competence-structure model, derived from developmental research, we constructed a comprehensive inventory of paper-and-pencil tasks that can be used in whole-class testing. This chapter provides an overview of the development of our inventory, and reports three central findings: (1) the convergent validity of our inventory, (2) the significant development of scientific reasoning in primary school from Grades 2 to 4, and (3) empirical proof of our competence-structure model.

Keywords Scientific reasoning • Primary school • Competence modeling

3.1 Science-P

The Science-P project (*Science* competencies in *Primary school*) investigated the development of two central dimensions of science understanding: general scientific reasoning, and conceptual understanding in physics in primary school. This chapter focuses on the dimension “scientific reasoning” and reports central findings regarding the development of this form of reasoning from Grades 2 to 4. The

S. Koerber (✉) • C. Osterhaus
Freiburg University of Education, Freiburg, Germany
e-mail: susanne.koerber@ph-freiburg.de; osterhaus@ph-freiburg.de

B. Sodian • D. Mayer • N. Kropf
University of Munich (LMU), Munich, Germany
e-mail: sodian@psy.lmu.de; daniela.mayer@psy.lmu.de; nicola.kropf@psy.lmu.de

K. Schwippert
University of Hamburg, Hamburg, Germany
e-mail: knut.schwippert@uni-hamburg.de

development of conceptual understanding in physics is described in the chapter by Pollmeier et al. (2017) in this volume.

Whereas early studies of scientific reasoning focused primarily on secondary-school students, modern developmental research indicates the presence of basic scientific reasoning abilities in primary-school children (Zimmerman 2007) and even of beginning skills and understanding in preschool children (e.g., Koerber et al. 2005). The literature contains descriptions of two research approaches: (1) theory-oriented research focused on the developmental function and on qualitative change, mainly using interview-based studies (e.g., Carey et al. 1989; Kuhn 2010; Lederman 2007) and (2) research focusing on the psychometric modeling of science understanding (e.g., TIMSS, PISA), which usually involves large-scale assessments and complex models based on post-hoc-determined hierarchical levels of competence. Science-P aimed to bridge the gap between these two approaches by developing and empirically testing a theory-based model of scientific reasoning competence.

In line with the common conceptualization (e.g., Zimmerman 2007), we regard scientific reasoning as intentional knowledge seeking (Kuhn 2010) involving the generation, testing, and evaluation of hypotheses and theories, and reflecting on this process (e.g., Bullock et al. 2009). The resulting wide range of scientific reasoning tasks includes those related to experimentation strategies (e.g., control of variables [COV]), data interpretation and the evaluation of evidence (e.g., Kuhn et al. 1988), and the process of scientific knowledge construction (i.e., understanding the nature of science [NOS]). Despite the apparent variety of tasks, it is commonly assumed that understanding the hypothesis-evidence relation is fundamental to these diverse scientific reasoning tasks (Kuhn 2010; Zimmerman 2007); this assertion however has not been tested empirically.

3.2 Development of Our Inventory

Our inventory was constructed in three project phases (see Fig. 3.1). Based on an extensive literature review of interview-based and experimental studies, Phase I developed a series of paper-and-pencil tasks (see e.g., Koerber et al. 2011) that could be used in whole-class testing. In Phase 1a, we conducted several studies,

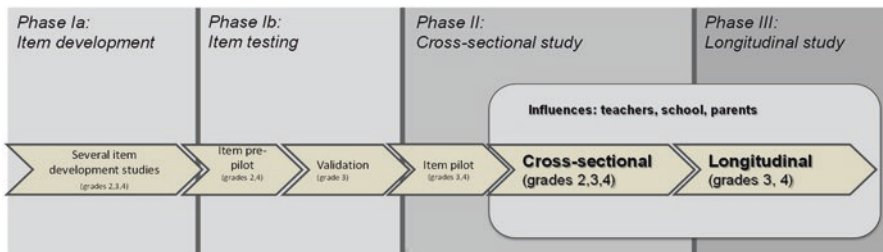


Fig. 3.1 Phases of the project Science-P

using multiple-choice (MC), forced-choice (FC), multiple-select (MS), and short open-answer tasks in one-on-one sessions. Each closed response format entailed answer options that corresponded to two or three hierarchical levels of competence, as postulated by the model (for an example of an MS task, see Fig. 3.2 from Koerber et al. 2015b). After designing and iteratively refining the tasks in several small studies, the first large-scale rotated-design study, involving 379 second and fourth





| | | |
|--|--|--------------------------------------|
| <p>Long ago, in the Middle Ages, people believed there were witches who could make people sick.</p> |  | |
| <p>A modern-day scientist traveled back to the Middle Ages with a time machine.</p> |  | |
| <p>Scientists in the Middle Ages thought that witches could make people sick. The modern-day scientist believed that bacteria could make people sick.</p> |  | |
| <p>The modern-day scientist showed the scientist from the Middle Ages the bacteria under the microscope and explained: "These bacteria are the reason why people get sick!"</p> |  | |
| <p>What would the scientist <u>from the Middle Ages</u> say to this?</p> | | |
| | <p>He would say this.</p> | <p>He would <u>not</u> say this.</p> |
| <p>1. "Of course you're right. Bacteria make people sick, not witches." naive</p> | <input type="checkbox"/> | <input type="checkbox"/> |
| <p>2. "Bacteria could be the witches' little helpers." advanced</p> | <input type="checkbox"/> | <input type="checkbox"/> |
| <p>3. "It may be true that there are bacteria here, but witches are still the ones who make people sick." intermediate</p> | <input type="checkbox"/> | <input type="checkbox"/> |
| <p>Which is the <u>best</u> answer?</p> | <p>No. _____</p> | |

Fig. 3.2 Example of an item assessing NOS (understanding theories) (Reprinted from Koerber et al. (2015b) with permission from Wiley & Sons. (C) The British Psychological Society)

graders, was conducted in order to test the fit of tasks ($N = 47$) and the applicability of the inventory in whole-class testing procedures (item pre-pilot). Phase 1b finished with a validation study of representative items of the inventory (see Kropf 2010, and below, Sect. 3.3). After constant improvement and extension of the item pool, Phase 2 comprised a large item pilot study involving 996 third and fourth graders. Based on item fits, biserial correlations, difficulty, and discrimination, 13 out of 83 tasks were excluded. The resulting item pool formed the basis for the further optimization and selection of tasks for the cross-sectional study (see below, Sect. 3.4), which also took place in Phase 2 and which tested more than 1500 second, third, and fourth graders. Phase 3 of the project was a longitudinal study (with two measurement series to date) that began with testing more than 1500 third graders (see below, Sect. 3.5).

Taking into account the diverse aspects of scientific reasoning, we aimed to provide a comprehensive inventory of scientific reasoning competence comprising five components: (1) a knowledge of experimentation strategies (e.g., the COV strategy), (2) an understanding of conclusive experimental designs for hypothesis testing, and (3) the ability to test hypotheses by interpreting data and evidence, and—on a more general level—to assess the understanding of NOS concerning (4) the goals of science and (5) how sociocultural frameworks influence theory development.

3.3 Convergent Validity of Paper-and-Pencil Inventory and Interviews

Whether the designed tasks adequately captured children's scientific reasoning competence was tested in a validation study comparing performance in a set of tasks with performance in an established interview (cf. Carey et al. 1989). The evidence for convergent validity is crucial, since a potential criticism of the use of paper-and-pencil tests is that they increase the probability of responding correctly by guessing (Lederman 2007). Indeed, paper-and-pencil tests might lead to arbitrary responses, and significant relations between children's answers in interviews and parallel MC tests are not always found. Whereas a slightly better performance might be expected in paper-and-pencil tests rather than interviews, due to the lower cognitive and language demands in the former, interindividual differences should be stable across the two methods when testing convergent validity. Because standardized interviews do not exist for all aspects of scientific reasoning, we exemplarily chose understanding NOS to establish the instrument's validity (see also Kropf 2010 for a related analysis of the instruments' validity, incorporating the component experimentation strategies).

3.3.1 Method

3.3.1.1 Participants

The participants comprised 23 third graders ($M = 8.10$ years, $SD = 5$ months) recruited from two primary schools in a rural part of Germany.

3.3.1.2 Material

Interview The Nature of Science Interview (NOSI; Carey et al. 1989; Sodian et al. 2002) focuses on the hypothesis-evidence relation: that is, the metaconceptual understanding of ideas (i.e., hypotheses, theories) underlying scientific activities and their differentiation from evidence. NOSI consists of several questions investigating children's understanding of science in general (e.g., "What do you think science is all about?") and of its central elements (e.g., ideas, hypotheses, experiments) as well as their relations (e.g., "What happens when scientists are testing their ideas, and obtain a different result from the one they expected?"). Based on prior research (Kropf 2010), the present study used a reduced version of NOSI, (nine of the 18 questions).

A three-level coding scheme was adapted from Carey et al. (1989, see also Bullock et al. 2009; Sodian et al. 2006) and further differentiated into the lowest level due to the youth of our participants and our focus on beginning abilities. The answers at Level 0 (the lowest naïve level, Level 1a, according to Sodian et al.) reflect a naïve understanding in which science is understood in terms of activities and without reference to ideas as formative instances of knowledge (e.g., "the goal of science is to make things work"). At Level 0.3, again a naïve level (Level 1b according to Sodian et al.), children regard science as information-seeking, but do not yet display an understanding of the hypothesis-evidence relation. Answers at Level 1 (the intermediate level) reflect a basic but not yet elaborated understanding of the differentiation between ideas and activities (e.g., "scientists consider things and think about why things are as they are; then they do research, perhaps they read what others have done, and then they probably ask a question why something is as it is, and they just do science"). Answers at Level 2 (the scientifically advanced level) indicate a beginning understanding of the relations between theories, hypotheses, and experiments, sometimes including an implicit notion of the role of a theoretical framework (e.g., "scientists have a certain belief or hypothesis, and then they try to confirm it by doing experiments or tests").

Paper-and-Pencil Tasks This study used five paper-and-pencil tasks presented in the format of FC, MC, or MS questions.

Control Variables Textual understanding was assessed using the ELFE 1–6 German reading proficiency test (Lenhard and Schneider 2006). Intelligence was assessed using the working-memory, logical-reasoning, and processing-speed subtests of HAWIK-IV, which is the German version of the WISC intelligence test (Petermann and Petermann 2008).

3.3.1.3 Procedure

Each child was tested twice. Half of the participants received the paper-and-pencil tasks first (whole-class testing) followed by the individual interview, with the order reversed for the other half. In the whole-group session, each child completed an individual test booklet under step-by-step guidance from an administrator using a PowerPoint presentation. Furthermore, a test assistant helped in the answering of comprehension questions.

3.3.2 Results

3.3.2.1 Pre-analyses

Pre-analyses revealed no significant effect either of order of presentation (interviews before paper-and-pencil tasks or vice versa), $F(1, 21) = 0.22$, *ns*, for the paper-and-pencil test, $F(1, 21) = 0.07$, *ns*, for the interview, or of gender, $F(1, 21) = 0.60$, *ns* and $F(1, 21) = 0.15$, *ns*.

3.3.2.2 Convergent Validity

We found a significant correlation between the scores for the paper-and-pencil test and NOSI ($r = .78$, $p < .01$). Whereas NOSI especially differentiated lower competencies (i.e., naïve conceptions at Level 0 or 0.3), the spread in the paper-and-pencil test was much larger (see Fig. 3.3). When partialing out intelligence and reading ability, the correlation between scores for NOSI and the paper-and-pencil test remained strong ($p_r = .70$, $p < .01$; partial correlation).

We also found that the level of difficulty differed significantly between NOSI and the paper-and-pencil test, in that children showed a significantly lower score in NOSI ($M = 0.22$, $SD = 0.14$) than in the paper-and-pencil test ($M = 0.95$, $SD = 0.43$; $t(22) = 10.20$, $p < .001$).

3.4 Scientific Reasoning: Development from Grades 2 to 4

After constructing a reliable scale and establishing its validity, we used the instrument (1) to systematically investigate the development of scientific reasoning from Grades 2 to 4, and (2) to investigate whether components of scientific reasoning are conceptually connected (for a more detailed description see Koerber et al. 2015a; Mayer 2012; Mayer et al. 2014).

In a rotated design, we presented more than 1500 children from Grades 2 to 4 with 66 paper-and-pencil tasks comprising several components of scientific reasoning. The children were also presented with an intelligence test (CFT) and a test of text comprehension (see Sect. 3.2). Furthermore, the parents completed a questionnaire about their socioeducational status (SES).

A unidimensional Rasch model revealed a good fit to our data, with only six items being excluded due to undesirable item fit statistics. The reliability was found to be good ($EAP/PV = .68$). Several multidimensional model comparisons supported the divergent validity of scientific reasoning, intelligence, problem-solving, and textual understanding, although these constructs are closely related to scientific reasoning, as indicated by strong correlations (between .63 and .74). Furthermore, different components of scientific reasoning (see Sect. 3.2) could be scaled together, indicating that they constituted a unitary construct. The results for the entire sample were the same as those for each grade separately. Identifying scientific reasoning as a unitary construct is especially impressive, given that the children were only second graders and that we used a comprehensive test with tasks involving different scientific-reasoning components in a single test.

Significant development was observed from Grades 2 to 3 and from Grades 3 to 4: this was independent of intelligence, textual understanding, and parental educational level. Previous studies of scientific reasoning in primary schools have employed single tasks from only one or two scientific-reasoning components, and the present study is the first to trace the development of scientific reasoning across different components using multiple tasks. The use of this inventory revealed development from Grades 2 to 4, despite scientific-reasoning competence not being explicitly and continuously addressed in the curricula.

Similarly to intelligence and textual understanding, the parental educational level and the time of schooling significantly impacted the children's scientific reasoning competence. However, since the obtained data are purely correlational, the direction and possible causation of these variables should be addressed in a future longitudinal study.

3.5 Competence-Structure Model of Scientific Reasoning: Hierarchical Levels of Competence

A central aim of the Science-P project was to develop and empirically test a competence-structure model. More specifically, our model is based on accounts of scientific reasoning that posit distinct hierarchical levels of naïve and intermediate understanding that children pass through before developing more advanced conceptions of science and the scientific method (Carey et al. 1989). Up to this point, testing such models has posed methodological difficulties, since these levels had not been implemented a priori in the tasks used in any previous large-scale study.

This was the first longitudinal study to test the competence structure described herein. In our refined inventory, the answer options for each item included all three hierarchical levels (naïve, intermediate, advanced), and the children were asked to consider each answer option individually (MS format) and also to choose the best answer option (MC format). From the resulting eight possible patterns of rejection and acceptance of each of the three answer options, the lowest level answer was identified as the final level. That is, an answer was coded as being naïve whenever the child endorsed a naïve level (regardless of the other answer options) and performance was coded as advanced only when the child accepted the advanced answer option and simultaneously refuted the naïve and intermediate options. An intermediate score was given in the case of acceptance of the intermediate and rejection of the naïve option, regardless of the acceptance of the advanced option. This form of MS assessment reduces the probability of correctly answering the items by guessing, which is a known problem of MC assessment.

The first measurement series of the longitudinal study (see Osterhaus et al. 2013) included a sample of more than 1300 third graders—a different sample from that in the cross-sectional study (reported in Sect. 3.3)—who answered 23 MS tasks on scientific reasoning (see Fig. 3.2). Again, intelligence and textual understanding were assessed.

A partial-credit model revealed a good fit to the data, supporting the hypothesized competence-structure model, which postulated that the three distinct levels represent the theorized hierarchical difficulties. For all but eight tasks, this assumption was supported by three indicators: (1) higher point-biserial correlations for higher categories (e.g., intermediate and advanced conceptions), (2) increasing ability level per category (naïve < intermediate < advanced conception), and (3) ordered delta parameters. This instrument, which includes hierarchical levels, exhibited acceptable reliability, and its divergent validity with respect to intelligence and textual understanding confirmed the results of the cross-sectional study presented in Sect. 3.3. The items differentiated sufficiently between children, although changing the item format to an MS format, and the stricter coding, made the items generally more difficult than in the cross-sectional study.

Together, these results confirm the validity of our competence-structure model, which posits three hierarchical levels. Therefore, we have succeeded in combining the methodological scrutiny of competence modeling with developmental accounts of the conceptual development of scientific reasoning (Carey et al. 1989).

3.6 Outlook

Future studies will include the results of the second measurement series, and will use multilevel analyses and structural models to determine competence gains and conceptual development, taking into account multiple factors of different levels of influence (e.g., intelligence, socio-economic status, teacher competence). Analyses of the developmental paths with respect to the hierarchical levels are currently underway.

An important future next step is to investigate the assumed mutual influence of content-specific science understanding (Pollmeier et al. 2017, in this volume) and scientific reasoning in development. The cross-sectional study of Pollmeier et al. found a close relation between both dimensions, and the results obtained in the present longitudinal study will facilitate identifying the direction of the influences.

In summary, the Science-P project contributes to our understanding of the relation between scientific reasoning and content-specific science understanding and its development in primary school. In addition, it has produced a competence-structure model of scientific reasoning in primary school and shed light on many of the important factors influencing the development of scientific reasoning, including intelligence, parental educational level, and school.

Acknowledgments The preparation of this paper was supported by grants to Susanne Koerber (KO 2276/4-3), Beate Sodian (SO 213/29-1/2); and Knut Schwippert (SCHW890/3-1/3) from the German Research Foundation (DFG) in the Priority Program “Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes” (SPP 1293).

References

- 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 the Munich Longitudinal Study* (pp. 173–197). Mahwah: Erlbaum.
- 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 junior high school students’ understanding of the construction of scientific knowledge. *International Journal of Science Education*, *11*, 514–529.
- 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.

- Koerber, S., Sodian, B., Kropf, N., Mayer, D., & Schwippert, K. (2011). Die Entwicklung des wissenschaftlichen Denkens im Grundschulalter: Theorieverständnis, Experimentierstrategien, Dateninterpretation [The development of scientific reasoning in elementary school: Understanding theories, experimentation strategies, and data interpretation]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *43*, 16–21. doi:10.1026/0049-8637/a000027.
- Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015a). 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., & Sodian, B. (2015b). Testing primary-school children's understanding of the nature of science. *British Journal of Developmental Psychology*, *33*, 57–72. doi:10.1111/bjdp.12067.
- Kropf, N. (2010). *Entwicklung und Analyse von Messinstrumenten zur Erfassung des wissenschaftlichen Denkens im Grundschulalter* [Development and analysis of instruments for the measurement of scientific reasoning in elementary school] (Unpublished doctoral dissertation). LMU München, München.
- Kuhn, D. (2010). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *Handbook of childhood cognitive development* (2nd ed., pp. 472–523). Oxford: Blackwell.
- Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. San Diego: Academic Press.
- Lederman, N. G. (2007). Nature of Science: Past, present, and future. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 831–880). Mahwah: Erlbaum.
- Lenhard, W., & Schneider, W. (2006). *ELFE 1–6. Ein Leseverständnistest für Erst- bis Sechstklässler* [ELFE 1–6. A reading proficiency test for children in Grades 1–6]. Göttingen: Hogrefe.
- Mayer, D. (2012). *Die Modellierung des wissenschaftlichen Denkens im Grundschulalter: Zusammenhänge zu kognitiven Fähigkeiten und motivationalen Orientierungen* [Modeling scientific reasoning in elementary school: Relations with cognitive abilities and motivational orientations]. Doctoral dissertation. Retrieved from <http://edoc.ub.uni-muenchen.de/14497/>
- 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.
- Osterhaus, C., Koerber, S., Mayer, D., Schwippert, K., Sodian, B. (2013, August). *Scientific reasoning: Modelling hierarchical levels of understanding*. Poster presented at the 15th biennial EARLI conference, München.
- Petermann, F., & Petermann, U. (Eds.). (2008). *Hamburg-Wechsler-Intelligenztest für Kinder IV (HAWIK-IV)* [Hamburg-Wechsler intelligence test for children IV (HAWIK IV)]. Bern: Huber.
- Pollmeier, J., Möller, K., Hardy, I., & Koerber, S. (2011). Naturwissenschaftliche Lernstände im Grundschulalter mit schriftlichen Aufgaben valide erfassen [Do paper-and-pencil tasks validly assess elementary-school children's knowledge of natural sciences]? *Zeitschrift für Pädagogik*, *6*, 834–853. doi:10.3262/ZP1106834.
- Pollmeier, J., Troebst, S., Hardy, I., Moeller, K., Kleickmann, T., Jurecka, A., & Schwippert, K. (2017). Science-P I: Modeling conceptual understanding in primary school. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education: Research, models and instruments* (pp. 9–17). Berlin: Springer.
- Sodian, B., Thoermer, C., Kircher, E., Grygier, P., Günther, J. (2002). Vermittlung von Wissenschaftsverständnis in der Grundschule [Teaching the nature of science in elementary school]. *Zeitschrift für Pädagogik*, Beiheft *45*, 192–206.
- Sodian, B., Thoermer, C., Grygier, P., Wang, W., Vogt, N., Kropf, N. (2006). *Coding scheme to the nature of science interview (BIQUA NOS)*. Unpublished paper, LMU München, München.
- Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*, 172–223. doi:10.1016/j.dr.2006.12.001.