Domain-General and Domain-Specific Scientific Thinking in Childhood: Measurement and Educational Interplay

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
PETER ADRIAAN EDELSBRUNNER
Mag. rer. nat., University of Graz
born on 15.01.1987
citizen of Austria

accepted on the recommendation of
Prof. Elsbeth Stern
Prof. Michael Schneider
Prof. Oliver Lüdtke

2017
για Μπόρα.
# Contents

Abstract ................................................................. I
Zusammenfassung .................................................. III

1 General Introduction ............................................. 1
   1.1 History of Research on Scientific Thinking ................. 4
   1.2 The Present Research .................................... 11
   1.3 Methodological Considerations ........................... 23

2 Psychometric Issues in Research on Scientific Reasoning .... 35
   2.1 Introduction ............................................. 36
   2.2 Literature Review ...................................... 47
   2.3 Results ................................................ 50
   2.4 Discussion ............................................. 58

3 Elementary School Students’ Verbal and Non-Verbal Knowledge about Experimental Design ........ 87
   3.1 Introduction ............................................. 88
   3.2 Method ................................................ 99
   3.3 Results ................................................ 108
   3.4 Discussion ............................................. 118

4 Variable Control and Conceptual Change: A Large-Scale Quantitative Study in Elementary School .... 135
   4.1 Introduction ............................................. 136
4.2 Method .......................................................... 142
4.3 Results ......................................................... 147
4.4 Discussion .................................................... 153

5 Enhancing Elementary School Students’ Domain-General Experimentation Skills by Inquiry-Based Content-Focused Physics Education 167
5.1 Introduction .................................................. 168
5.2 Method ........................................................ 169
5.3 Results ........................................................ 175
5.4 Discussion .................................................... 180

6 General Discussion 185
6.1 Synopsis of the Four Studies ............................... 185
6.2 Outlook ....................................................... 191

7 Appendix 194
Abstract

Scientific thinking encompasses domain-general thinking about principles of scientific inquiry, and domain-specific thinking about the contents of scientific concepts. In the present thesis, in four studies the measurement and the interplay of these two facets of scientific thinking are examined, with a focus on their development in elementary school children. In the first study, a review of psychometric modeling practices in research on scientific reasoning is undertaken. The most frequently applied psychometric technique turns out to be an approach to Rasch modeling that is often used in large-scale assessment programs. Supported by data simulations, it is argued that researchers’ current approaches to fit testing and interpreting results within this approach tend to be inconsistent, and to go beyond the limits of the data. Potential reasons for these current practices are discussed, and suggestions are provided for complementary psychometric approaches in future research. In the second study, patterns and interrelations in elementary school students’ verbal and non-verbal knowledge about experimental design are examined. In a sample of over 3000 first- to sixth-graders, patterns and interrelations in the different school grades across these two types of knowledge are estimated in latent variable models. It is found that during the first years of schooling, verbal and non-verbal knowledge are statistically separable, and children show heterogeneous patterns across the two kinds of knowledge. Fourth grade marks a turning point, from which on strong cross-sectional development in both kinds of knowledge is observed up to sixth grade. The strong heterogeneity found in students’ knowledge provides an explanation for prior inconsistent findings. In the third study, the contribution of students’ understanding of variable control in experimental design to their knowledge development in inquiry-based instruction is investigated. $N = 1809$ first to sixth graders received classroom-based instruction on the topic floating and sinking of objects in water. In a mixture modeling approach, the understanding of variable control turns out to be an informative predictor of students’ knowledge development from before to after instruction: The better students’ understanding of this important factor of experimental design, the more likely they are to restructure and integrate their prior knowledge during instruction. Taking prior lab-based findings to classrooms, these results show that students’ understanding of variable control matters in teacher-guided instruction. In the fourth study, it is examined whether students’ understanding of variable control can be augmented by receiving a high dose of inquiry-based instruction in which
variable control is of crucial importance but never explicitly instructed. In a multilevel path model, a dummy variable representing whether or not students received this intervention can explain variance of the change in $N = 181$ third-graders’ understanding of variable control. This result indicates that a high dose of domain-specific science learning can transfer to a domain-general science skill, opening new perspectives on the dynamics of science learning. The results of the four studies are discussed in light of their contributions to knowledge about the measurement and interrelations of the two types of scientific thinking, and suggestions are provided for potential future research to advance the present insights.
Zusammenfassung

Chapter 1

General Introduction

Scientific thinking encompasses two broad facets that can be described as scientific thinking that is domain-general, because it is based on principles that can be applied similarly across domains, and as scientific thinking that is domain-specific, because it is based on knowledge that can only be applied to a specific domain. Domain-general scientific thinking, for example, is demanded in the following item:

Recently, someone from Paul’s school class started leaving letters on his table, without signing them. Paul wants to find out who it is! He leaves a cookie on his table. And indeed, the next day the cookie is gone apart from some crumbs, and another letter has been left behind!

Good for Paul that he knows his classmates so well; he is quite sure that it is either Barbara or Sophie who leaves the letters, and he knows the following about them:

- Barbara likes eating Chocolate and Nuts, but no Cookies
- Sophie likes eating Cookies and Nuts a lot, but no Chocolate

Can you help Paul and tell him Who is the mysterious letter-buddy?

Figure 1.1. Can you help Paul with this domain-general riddle?
In this item, understanding of the determinacy and indeterminacy of evidence is demanded: Only one classmate likes cookies, and thus the evidence seems quite clear - it is a determinate situation. If however both classmates liked cookies, the evidence would be indeterminate, and Paul’s cookie-test inconclusive. The situation described in this item is quite context-specific. The underlying domain-general principle of determinacy, however, is principally valid across any context and content domains. For example, in inquiry-based science education, the understanding of determinacy often plays a role, in that any scientific evidence can be determinate, or indeterminate, and has to be interpreted accordingly.

Domain-specific scientific thinking, on the other hand, can only be meaningfully applied in its original domain, such as in the following item:

A person is standing in a resting boat and tosses a big stone into the water behind the boat. Which of the following statements are true?

- The boat moves in the direction the stone was thrown.
- The stone displaces water and this is why the boat moves just slightly back and forth.
- If you let an inflated balloon whizz through the air, principally the same happens.
- The boat moves contrary to the direction the stone was thrown.

*Figure 1.2.* Item demanding domain-specific scientific thinking. Borrowed from the *basic Mechanics Concept Test* (bMCT) in S. I. Hofer (2015).

In this item, in order to recognize that the fourth answer option is correct, applying conceptual understanding of Newton’s third law of motion is demanded. Application of this conceptual knowledge is an instance of domain-specific scientific thinking, because Newton’s third law is bound to the domain of mechanics. In the present thesis, the measurement, development, and education of these two types of scientific thinking are in focus.
Scientific thinking has seen major research interest since the beginnings of educational and developmental research. Educational scientists emphasize that apart from content knowledge, students should also acquire a *scientific way of thinking*. Related concepts used to be named and described as *reflective thinking* (Dewey, 1910), an understanding of *when and how to combine investigation with authority* (Herring, 1918), the ability to *think clearly* (Blair, 1940), a *scientific attitude* towards the problems of hie (Blair, 1940), *logical thinking* (Buck & Ojemann, 1942), and *critical thinking* (Ennis, 1964). In addition to educational aims, developmental scientists emphasize the need of developing skilled scientific thinking as a prerequisite for further cognitive development and personal growth. The developmental strand of research on scientific thinking was initiated in the 1950s and initially described related concepts as *formal reasoning* (Piaget & Inhelder, 1958) and *scientific reasoning* (Siegler & Liebert, 1975).

After a century of educational and developmental research on scientific thinking, some major insights have been gained and new questions have been raised. We now know in detail the developmental pathways of some psychological traits regarded as scientific thinking, and how these are influenced by education (Klahr, Zimmerman, & Jirout, 2011; Sandoval, Sodian, Körber, & Wong, 2014; Sodian & Bullock, 2008; Wilkening & Sodian, 2005; Zimmerman, 2000, 2007). The present research sheds light from a new perspective on one of the major insights about scientific thinking that is often emphasized but has not yet been researched in regular classroom education: The developmental interplay of domain-general and domain-specific scientific thinking. This interplay has been examined in detail in laboratory settings: Children’s knowledge about domain-general principles of scientific inquiry influences their knowledge development about domain-specific scientific concepts, and vice versa. The dynamics of this interplay are here examined from different perspectives and taken to education in real classrooms, to see whether and how, under circumstances of well-designed inquiry-based science education, this interplay matters. Meaningful measurement is the precondition for quantitative research, and contributions to the measurement of scientific thinking are described in turn in Chapters 2 and also in Chapter 3, the latter of which marks the transition to substantial questions that are examined in Chapters 4 and 5, and a general discussion is provided in Chapter 6.
1.1 History of Research on Scientific Thinking

Scientific thinking can be broadly described as the thinking involved in the skills necessary for inquiry, experimentation, evidence evaluation, and inference that are done in the service of conceptual change or scientific understanding (Zimmerman, 2007). A literature search was undertaken in the PsycInfo database with the terms "Scientific thinking" OR "Scientific reasoning". The literature search yielded an insight into the numbers of yearly publications in this topic, and seminal and historical research. As apparent from Figure 1.3, the number of publications in this area has been rising throughout the last decades\(^1\). Based on the results from this literature search, research on scientific thinking can be summarized in four periods that emerged in 1910: The Pre-Cognitivist, Cognitivist, Seminal, and Large-Scale periods (Figure 1.3).

A seminal monograph that initiated systematic philosophy, theoretical and empirical research on scientific thinking was published in 1910, marking the beginning of the first period: John Dewey’s *How we Think* (Dewey, 1910). Dewey (1910) reflections on scientific literacy and the following curricula reforms aimed at matching his vision of a scientifically literate general population (Dewey, 1910; Shamos, 1995). Although Dewey himself was philosopher and not part of the behaviorist movement, during the period of Behaviorism cognitive research was not actively undertaken, and therefore this phase is termed the Pre-Cognitivist phase. This book marked the first of 11 historical and seminal publications that are indicated in Figure 1.3 and with

\(^1\)This is not a surprise, since the general body of research grows; putting into Scopus ad hoc the word "grandma" brings a similar graph, which, however, if it is not set in relation to the overall number of publications in relevant journals or fields, does not yet indicate "increasing research interest in grandma":

![Graph showing publication trends](image-url)
more bibliographic details in Table 1.1.

NEED. Civilized society needs more scientific ability.

TEST. Put a cross on the line before the word that is in the wrong group of words; but if there is no such word, put a cross here: Animals: ___cat ___cow ___horse

– John Herring (1918), first line, and likely first item ever for assessing scientific thinking, from *Measurement of Some Abilities in Scientific Thinking*.

The next historical publication marked the beginning of research on measurement of scientific thinking. Almost exactly a century ago, in 1918, likely the first instrument to assess scientific thinking was published. Herring (1918) developed a questionnaire to assess the ability to "study, to handle problems, to work methodically, to plan independently their own ways of solving problems... ...in short, to use the scientific method".

![Figure 1.3](image_url)

*Figure 1.3.* Annual publications in the history of research on scientific thinking and reasoning since 1910. Historical and seminal publications indicated with dashed lines at respective publication year. Four research periods indicated by shaded areas, details discussed in text.
In 1940, Glenn Blair commenced his description of a validation study of the *Noll Test of Scientific Thinking*, another early instrument to assess scientific thinking, with the statement:

"One of the most important functions of the school should be to develop in children the ability to think clearly and to face problems in an unprejudiced manner. Children should learn early to discriminate between fact and opinion, and should develop the habit of suspending judgment when data at hand do not warrant the formation of a conclusion. In a word, children should learn to exhibit a scientific attitude toward the problems of life. Since this is one of the commonly accepted goals of education, some means should be devised for adequately measuring improvement in this trait".

Hence, the measurement of scientific thinking has been on researchers’ agenda since the early decades of educational research.

From an educational perspective, measurement should optimally be followed by education, and another milestone in research on scientific thinking is the first study investigating the effects of an educational intervention targeted at developing scientific thinking. Blair and Goodson (1939) compared groups of ninth-graders who either received science education focused on scientific thinking, non-focused science education, or no science education. The group of students who received the targeted science education indeed made the biggest gains on the Noll’s "What Do you Think?"-test. Thus, already from this first ever intervention study it was inferred that science education with explicit teaching of scientific thinking is necessary in order to support its development. Nevertheless, this study was the first of a myriad to examine how to best teach scientific thinking, a topic which since this beginning has been further worked on in ongoing research (Sandoval et al., 2014).
Table 1.1

*Historic and Seminal Publications in Research on Scientific Thinking, chronologically.*

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dewey</td>
<td>1910</td>
<td>How We Think</td>
</tr>
<tr>
<td>Herring</td>
<td>1918</td>
<td>Measurement of some abilities in scientific thinking</td>
</tr>
<tr>
<td>Blair &amp; Goodson</td>
<td>1933</td>
<td>Development of Scientific Thinking through General Science</td>
</tr>
<tr>
<td>Bruner et al.</td>
<td>1956</td>
<td>A Study of Thinking</td>
</tr>
<tr>
<td>Inhelder &amp; Piaget</td>
<td>1958</td>
<td>The growth of logical thinking from childhood to adolescence</td>
</tr>
<tr>
<td>Klahr &amp; Dunbar</td>
<td>1988</td>
<td>Dual space search during scientific reasoning</td>
</tr>
<tr>
<td>Kuhn</td>
<td>1989</td>
<td>Children and Adults as Intuitive Scientists</td>
</tr>
<tr>
<td>Sodian et al.</td>
<td>1991</td>
<td>Young Children’s Differentiation of Hypothetical Beliefs from Evidence</td>
</tr>
<tr>
<td>Dunbar</td>
<td>1993</td>
<td>Concept discovery in a scientific domain</td>
</tr>
<tr>
<td>Klahr et al.</td>
<td>1993</td>
<td>Heuristics for Scientific Experimentation: A Developmental Study</td>
</tr>
<tr>
<td>Schauble et al.</td>
<td>1995</td>
<td>Students’ Understanding of the Objectives and Procedures of Experimentation in the Science Classroom</td>
</tr>
<tr>
<td>Schauble</td>
<td>1996</td>
<td>The development of scientific reasoning in knowledge-rich contexts</td>
</tr>
<tr>
<td>Chen &amp; Klahr</td>
<td>1999</td>
<td>All other things being equal: Acquisition and transfer of the control of variables strategy</td>
</tr>
<tr>
<td>Zimmerman</td>
<td>2000</td>
<td>The Development of Scientific Reasoning Skills</td>
</tr>
<tr>
<td>Zimmerman</td>
<td>2007</td>
<td>The development of scientific thinking skills in elementary and middle school</td>
</tr>
</tbody>
</table>
The Cognitivist period began with two seminal publications that initiated two new distinct strands of research on scientific thinking. Jean Piaget, the founder of modern developmental science, together with Bärbel Inhelder initiated research on scientific thinking from the view of the emergence of formal logical thinking Piaget and Inhelder (1958). This book marked for example the beginning of research on the development of the control of variables strategy (CVS), a method for creating experiments in which a single contrast is made between experimental conditions (Chen & Klahr, 1999). Throughout the 20th century, CVS stayed as one of the major topics in research on scientific thinking and reasoning. CVS encompasses varying the variable of interest in an experimental design while keeping all other factors constant, thus allowing unambiguous causal inferences (Strand-Cary & Klahr, 2008). Today, we know that many but not all children typically develop these skills or precursors at ages 3-14, depending on the context and complexity of tasks (Chen & Klahr, 1999; Sodian, Zaitchik, & Carey, 1991; Strand-Cary & Klahr, 2008). Piaget led the basis for a strand of research examining this and further experimentation skills. Importantly, the logical thinking skills that this strand of research is concerned with are domain-general. Generality in this case means that the underlying rules can be applied across contexts or domains.

On the other hand, Bruner, Goodnow, and Austin (1956) initiated a strand of research concerned with development and change in people’s conceptions about scientific and natural phenomena. This is a domain-general strand of research, because it is concerned with the concepts that people hold about something, for example about a physical phenomenon. This research provided the basis for later research on conceptual change which, influenced by Thomas Kuhn’s descriptions of paradigm changes (T. S. Kuhn, 1970), tries to elucidate how humans adapt and modify concepts about physical phenomena. For example, Carey (1985) provided a detailed description of the development of children’s concepts in various domains, and Vosniadou and Brewer (1992) provides an elegant analysis of children’s concepts about the earth.

Throughout the Cognitivist period, researchers were mainly concerned with these two threads of scientific thinking. In 1988, the Seminal Period initiated with the work of Klahr and Dunbar (1988) who managed to bring these two threads of research together within the Scientific Discovery as Dual Search (SDDS) model. In the SDDS model, scientific reasoning is conceptualized as a problem solving process taking place in a hypothesis space and an experiment
space with the aim to develop and revise hypotheses that can explain empirical evidence. To engage in this problem solving process, three major cognitive processes are assumed to be necessary in the SDDS model: Hypothesis generation, experimental design, and evidence evaluation (Figure 1.4).

![Scientific Discovery as Dual Search (SDDS)](image)

**Figure 1.4.** Main components of SDDS model of scientific reasoning.

According to this model, hypotheses are generated influenced by prior knowledge about the investigated domain. Hypotheses are then mapped onto experiments that are conducted in order to produce evidence to test the hypotheses. The choice of experiments is again influenced by prior knowledge; for example, if children assume a causal relation between two factors, they are unlikely to design experiments that might disprove such a relation (Schauble, 1996). An experiment is then conducted, and evidence is interpreted in order to evaluate whether it promotes a hypothesis, or whether the hypothesis has to be changed in order to account for the accumulated evidence. This model of scientific thinking explains why and how domain-specific knowledge and domain-general experimentation influence each other, unifying the two prior threads of research on scientific thinking and reasoning. The SDDS model has served as a fruitful framework for research on scientific thinking across the last three decades, and it has also used as a framework for synthesizing empirical evidence (Morris, Croker, M., & Zimmerman, 2012; Zimmerman, 2000, 2007). Many more contemporary models of scientific thinking build on the three components identified in the SDDS model.

Another key theory about scientific was developed in research described in a seminal publication in 1989. D. Kuhn (1989) developed a theory of the development of scientific thinking that centered on the differentiation and coordination of theory and evidence. The proposition of
the relation between theory and evidence as they key to skilled scientific thinking and reasoning has been carried further by various researchers, including Körber, Mayer, Osterhaus, Schwippert, and Sodian (2014), Sodian et al. (1991) and Deanna Kuhn and her colleagues themselves (D. Kuhn, Iordanou, Pease, & Wirkala, 2008; D. Kuhn, Ramsey, & Arvidsson, 2015).

In 1999, research on the training of domain-general scientific thinking found a highlight with an elaborate study in which transfer across contexts and over time was compared between more strongly and less strongly guided scenarios towards teaching CVS (Chen & Klahr, 1999). In combination with this seminal study and further research, we now know that CVS can be trained successfully with remote transfer over time, particularly in teacher-guided interventions (e.g., Chen & Klahr, 1999; Ross, 1988; Strand-Cary & Klahr, 2008). Over the last 45 years, elaborate educational interventions have been developed to support the development of CVS and other aspects of experimental design. When the first intervention studies were conducted, it was believed that teaching experimental design might not be possible before adolescence. Also students’ background was seen as an influential factor; it was assumed that low achievers would have a hard time learning the principles of experimental design. This impression can be summarized in the conclusion of one of the earliest intervention studies, in which training had some effect and authors seemed rather surprised by this: "The goal of teaching low SES 14-year-olds to design and evaluate experiments is not completely unrealistic." (Case & Fry, 1973). After more than four decades of research, insights have been extended and the quality of intervention curricula has improved greatly. Intervention studies now encompass the whole range of education from early elementary school (Case, 1974) to university (Lin & Lehman, 1999), and low- and high achievers from diverse socioeconomic background (Case & Fry, 1973; Lorch Jr et al., 2010; Lorch et al., 2014; Zohar & Peled, 2008). In all of these populations, interventions have been designed that can successfully augment knowledge about experimental design in comparison to control groups (Chen & Klahr, 1999; Ross, 1988; Schwichow, Croker, Zimmerman, Höffler, & Härtig, 2015).

Of the remaining seminal publications in Table 1.1, various revolve around the mutual dynamics of domain-general and domain-specific scientific thinking. Particularly in lab-based studies, it has been found that the understanding of CVS relates mutually to conceptual belief revision in experimentation (Schauble, 1990, 1996).
Further scientific thinking skills that have been in the focus of research in the twentieth
century are the interpretation of covariation evidence (see e.g., Piaget & Inhelder, 1958; Shultz,
Fisher, Pratt, & Rulf, 1986; Shultz & Mendelson, 1975) and belief revision based on
counterevidence (e.g., Chinn & Brewer, 1998). Another topic also in the focus since the 1990s is
epistemic beliefs, that is, beliefs about knowledge and its development (see e.g., B. K. Hofer,
2004; B. K. Hofer & Pintrich, 1997; D. Kuhn, Black, Keselman, & Kaplan, 2000; Schommer,
Calvert, Gariglleti, & Bajaj, 1997).

Some new developments in research on scientific thinking and reasoning can be observed
since the late 1990s and early 2000s. Lately, Deanna Kuhn started moving beyond CVS (D. Kuhn,
2007; D. Kuhn et al., 2008; D. Kuhn & Pease, 2008), theorizing about the crucial skills apart from
CVS that are needed for successful scientific thinking and experimentation. As major skills, she
describes (a) the ability to coordinate effects of multiple causal influences on an outcome, (b) a
mature understanding of the epistemological foundations of science, recognizing scientific
knowledge as constructed by humans rather than simply discovered in the world, and (c) the
ability to engage in skilled argumentation in the scientific domain, with an appreciation of
argumentation as entailing the coordination of theory and evidence (D. Kuhn et al., 2008).
Deanna Kuhn and her team deliver classroom-based interventions to students under the theme
multivariable causal reasoning, which encompasses learning and applying CVS in designs with
multiple variables, and also the schooling of epistemic beliefs and argumentation, in order to
support understanding of the overarching theory-evidence relation (D. Kuhn et al., 2015).

The last period in research on scientific thinking initiated in the early 2000s with the rise of
large-scale assessment programs in which scientific thinking was encompassed. The emergence
of this period is partially documented in the first study of the present research.

1.2 The Present Research

Psychologists’ and educators’ hope is that scientific thinking can be educated effectively, in
contrast to more fixed personality traits such as intelligence. This concerns both scientific
thinking as content knowledge about science phenomena, and scientific thinking as
domain-general skills that are needed to engage in scientific inquiry, which are relevant in
scientific environments but also in everyday-life. The current thesis tries to tackle theoretical and
methodological issues in bringing prior developmental insights about these two facets of scientific thinking to the classroom.

The present research is embedded in a research framework of dynamic interrelations between domain-general and domain-specific scientific thinking, which is depicted in Figure 1.5. This framework focuses on three different aspects of both facets of scientific thinking that are important particularly in quantitative research. First, the framework acknowledges that effective measurement is a prerequisite for researching scientific thinking. Different approaches to measurement are suitable for tackling different questions, and the choice of measurement instrument and statistical measurement model is supposed to play a major role in this research area. Second, in this research framework education is acknowledged as a key variable in examining the interrelations of the two facets of scientific thinking, particularly in school contexts. Educating one facet in training interventions has the potential to reveal dynamic interrelations between the two facets: When domain-specific scientific thinking is trained, for example in interventions teaching children scientific concepts about natural phenomena, the success of this training might be influenced by children’s domain-general scientific thinking, and vice versa. Thus, taking into account educational interventions might enable triggering causal relations that inform about the dynamics underlying the development of scientific thinking.

The general questions that the present research tries to tackle based on this framework are the following. First, different approaches to measuring domain-general scientific thinking are explored and developed. A review is undertaken to examine current approaches, particularly in the statistical modeling of domain-general scientific thinking. Second, the development of domain-general scientific thinking is examined. Particularly, the development of children’s knowledge about CVS is examined, throughout elementary school. Current findings about this development are inconclusive (Morris et al., 2012; Schauble, 2003) and the present research tries to provide a detailed overview of patterns and also heterogeneity of children’s development in this regard. Third and fourth, the interrelation between domain-general and domain-specific scientific thinking is examined, focusing in one study on one direction of this interrelation, and in the other study on the other direction. The studies encompassed in the present research to tackle these four issues are now described in more detail.
Figure 1.5. The research framework of dynamic interrelations between domain-general and domain-specific scientific thinking used in the present research.

The first study is concerned with a prerequisite for examining development in scientific thinking: Its measurement, particularly that of domain-general scientific reasoning. In the paper, a literature review is conducted to examine which of the numerous available psychometric approaches researchers apply, and what their reasons for choosing specific approaches are. It turns out that mostly, an educational measurement-approach based on Rasch-modeling is applied that is currently predominant in large-scale assessment programs. This approach is tied to very specific scaling and model testing procedures which are based on debated statistical and philosophical background. These debates are summarized and alternative approaches suggested. In a simulation study, it is found that some of researchers’ inferences based on this psychometric approach might go beyond their data. Potential reasons for the predominance of this approach to the measurement of scientific reasoning are discussed, and suggestions for alternative approaches are provided. This study shows methodological issues in the measurement of scientific reasoning, tries to elucidate the underlying conceptual issues, and provides suggestions for alternative approaches that might optimally support theory development in future research.

In the second study, children’s verbal and non-verbal knowledge about experimentation is in the focus. The understanding of variable control is often assessed with multiple-choice questions but this might not represent whether children can adequately disseminate and discuss their knowledge about experimental design. Based on data from \( N = 3026 \) elementary school children’s open answers on a variable control assessment, a categorization scheme is developed to assess children’s verbal knowledge, and the underlying rationales they use for judging the quality
of experimental designs. Frequency analysis of children’s coded answers across school grades replicate and extend findings from the developmental literature. In latent variables models, it is found that students’ open answers are correlated but distinct from their multiple choice answers on the same questionnaire. Furthermore, in a mixture model, systematic heterogeneity is found in children’s verbal and non-verbal knowledge patterns, providing an explanation for prior inconclusive findings. The study provides a cross-sectional overview of students’ rationales underlying their reasoning about experimental designs in primary school. It also makes a methodological contribution with the development and validation of an elaborate coding and rating scheme, an assessment tool that provides a detailed look into students’ reasoning.

The third study examines an assumption that has often been made referring to lab-based findings but has not yet been examined in real classrooms. The assumption is that domain-specific science content knowledge and domain-general experimentation skills develop intertwined, which has been referred to as "mutual bootstrapping" of these two components of scientific thinking (Schauble, 1990, 1996). In the study, one side of this assumption is examined in primary school students who receive inquiry-based Physics instruction. \( N = 1809 \) students receive instruction on the Physics principles of object density and buoyancy force in the context of objects’ floating and sinking in water. In a theoretical discussion, it is shown that it is not at all obvious whether in the context of such Physics instruction, the developmental relation between content knowledge and experimentation skills holds, particularly under teacher-guided instruction. Before the instruction, the students receive an assessment of their understanding of variable control, since this well-researched experimentation skill is crucial for engagement and understanding in any experimental design. A mixture-modeling approach is applied to examine whether and how students’ achievement on the variable control assessment predicts their knowledge development during the instruction. The results show that this is indeed the case; students’ understanding of variable control predicts their knowledge development from before to after instruction. A regression-based analysis of the same data shows only marginal effects. This study makes a theoretical contribution by testing whether an often-assumed interrelation holds in a real classroom context. It also makes a methodological contribution by demonstrating that mixture modeling can offer detailed information about students’ knowledge development and how it is influenced by their cognitive characteristics, compared to relatively little information obtained.
from a traditional auto-regressive model.

In the fourth study, the interrelation between domain-specific and domain-general scientific thinking is again examined, now from the complementary perspective to the third study. It is examined whether a high dose of experimentation in domain-specific inquiry-based Physics instruction leads to improved development of domain-general experimentation skills. \( n = 81 \) third graders receive instruction on four basic Physics topics for 15 months on average, while a further \( n = 109 \) third graders deal as a comparison group. All students work on a questionnaire twice which assesses their understanding of variable control in experimental design, before and after they receive an extensive inquiry-based curriculum. In a multi-level model, it is found that in the intervention group, students’ understanding of variable control develops more strongly than that of students in the comparison group. Thus, receiving a high dose of domain-specific inquiry-based instruction supports this development, even though the domain-general principle of variable control is not explicitly instructed. This indicates a transfer effect from content-based instruction to domain-general experimentation principles. The study makes a theoretical contribution by testing this interrelation and transfer effect, supporting the assumption of intertwined development of content knowledge and experimentation skills in classroom instruction.

In a synopsis of the four studies, theoretical and methodological implications are discussed. Research on scientific thinking has seen a recent increase in literature, probably due to its inclusion in large-scale assessment programs. The present thesis complements this new phase of increased research activity by bringing former insights into classrooms, and by contributing to the methodological basis for further investigations.

1.2.1 The Swiss MINT Study

Three out of four studies of the present thesis are embedded in the Swiss MINT Study. This is a longitudinal study that was initiated in 2010 at the Section for Research on Learning and Instruction and the MINT Learning Center of ETH Zurich. The study was initiated with the idea to provide a large sample of Swiss primary school students with cognitively activating science education, continuously throughout all stages of schooling. The study is aimed at evaluating the long-term impacts of receiving this special curriculum. The present research was started in 2012 when I started contributing to this study. Chapters 3, 4, and 5 of this thesis are based on data
assessed within the framework of the Swiss MINT Study.

The general structure of the Swiss MINT Study is depicted in Figure 1.6. The study aims at recruiting primary school classes, preferably in second to fourth grade. The students from participating classes receive a science curriculum that was especially designed for the present study. Optimally, participating students receive the basic study curriculum in primary school and are further monitored and receive further instruction throughout their schooling career.

Currently, about 300 primary and secondary school classes from German-speaking Cantons (kind of federal states) of Switzerland participate in the study. In the relevant Cantons, primary school encompasses first to sixth grade. Children in participating school classes, first of all, receive four basic curricula, which consist of instruction in four Physics topics (Figure 1.6): Air & Air Pressure (Luft & Luftdruck), Sound & Sound Waves (Schall & Schallwellen), Floating & Sinking (Schwimmen & Sinken), and Building Bridges (Brücken & Was sie stabil macht).

The curricula are based on principles that have proven effective in empirical research on science education. The general approach throughout the curriculum is inquiry-based learning. In inquiry-based learning, a common kind of educational intervention in science education, students actively engage in thinking processes and activities of scientists (American Association for the Advancement of Science, 1993; Furtak, Seidel, Iverson, & Briggs, 2012). In the Swiss MINT Study, the implementation of this kind of learning is combined with general principles from cognitive-constructivist learning theory in order to optimally support students’ knowledge development.

![Figure 1.6](image_url) The general structure of the Swiss MINT Study, encompassing science education from primary school throughout secondary school.

The first stage of the Swiss MINT Study encompasses the four basic curricula. The instruction in these curricula is focused around teacher-guided inquiry-based instruction. The four
basic curricula have been developed by a team of science education experts at the University of Munster, Germany (Spectra Materials - KiNT-Boxes 1-4). The curricula are designed to support development of domain-specific content knowledge about basic physics concepts. The *Floating & Sinking* curriculum introduces the concepts of water displacement, object density, and buoyancy force. The *Air & Atmospheric Pressure* curriculum introduces air as non-visible matter that has weight and needs space. Additionally, it is demonstrated how air pressure can be functionally and mechanically used. The *Sound & Spreading of Sound* curriculum introduces the concepts of sound wave pitch and frequency, and wave movement. The *Stability of Bridges* curriculum introduces basic types of forces and principles of stable construction design. Hardy, Jonen, Möller, and Stern (2006) provide an extensive exemplary description of one curriculum (Floating & Sinking).

The curricula are deployed by students’ regular classroom teachers. In this way, students are taught under their regular classroom circumstances. This setting makes it difficult to control the exact implementation of the curricula. At the same time, however, it enables examining students’ development when they receive well-designed science education within their customary schooling environment. In order to prepare the teachers for deploying the curricula, they receive training by the team of the Swiss MINT Study at their schools or at ETH Zurich. For each of the four curricula, the teachers receive half- or full-day training. The teachers are then provided the KiNT-boxes, which for each curriculum comprise all necessary experimentation materials, worksheets, and theoretical background information.

The four curricula employ the same core educational principles. Children frequently engage in experimentation to explore the different physics concepts on their own, however under teacher guidance. The inquiry-based lessons are designed with a strong emphasis on instructional guidance and scaffolding. For example, prior knowledge is activated in teacher-led classroom discussions and in paper-based exercises before experimentation. Children write down their assumptions concerning the outcomes of the experiments, and they provide justifications for their expectations in a research notebook. After having conducted an experiment, children note the observed outcome and compare it to the expectations they noted in their notebook. The research notebook further contains additional content-related information and exercises. In concluding the lesson, the teacher secures children’s understanding of the underlying physics concepts in a teacher-led classroom discussion.
The four curricula together represent the initial stage of the Swiss MINT Study (Figure 1.6). In the further course of the study, participating students receive further curricula. Some of the further curricula build on the topics of the basic curricula on a more advanced level, in the manner of a spiral curriculum. Other more advanced curricula deal with new topics, such as magnetism, and later on topics from chemistry.

In order to monitor students’ knowledge development, they work on content knowledge assessments before and after undergoing each curriculum (Figure 1.6). In addition, before the first and after they have received the last of the four basic curricula, children also fill out a test which assesses their understanding of basic principles of experimental design. These assessments allow monitoring students’ development of scientific thinking throughout the curricula. The content knowledge assessments measure students’ domain-specific scientific thinking in each of the topics covered by the curricula, and the experimental design test measures domain-general scientific thinking.

Finally, in the beginning of the Swiss MINT Study some school classes were assigned to a waiting list condition. Students from these school classes worked on the same content knowledge and experimental design tests as the students in the intervention group, within a comparable time range. Students in the waiting list condition however did not yet receive the curricula. These students at first only filled out the tests, and received the curricula afterwards. In order to evaluate the impact of the curricula in comparison to students’ regular development, development of students in this waiting list condition can be compared to that of students in the regular intervention condition.

1.2.2 The Swiss MINT Study Variable Control Test

A specific assessment instrument that was developed for the Swiss MINT Study is applied in all empirical studies of the present research (Chapters 3, 4, & 5): The Swiss MINT Study Variable Control Test. This test encompasses items to assess students’ knowledge about two crucial principles of experimental design, conclusive hypothesis testing and the control-of-variables strategy. Conclusive hypothesis testing describes a hypothesis test in which the focal variable is varied such that it allows for conclusive evidence about its causal status. Knowledge about this principle corresponds to broad understanding of determinacy and
indeterminacy, that is, whether available evidence is a sufficient warrant for a specific conclusion (Falmagne, Mawby, & Pea, 1989; Klahr & Chen, 2003; Somerville, Hadkinson, & Greenberg, 1979). A conclusive hypothesis test corresponds to a test producing determinate evidence, as described in Figure 1.1.

The test encompasses 14 items. Example items are provided in Figure 1.7 and the whole test is provided in Appendix A. The different types of questions cover a broad construct representing conclusive hypothesis testing and various facets of the control-of-variables strategy (Schwichow, Christoph, Boone, & Härtig, 2016). Some of the tasks describe classic cover stories, such as the ramp- (Chen & Klahr, 1999), airplane- (Bullock & Ziegler, 1999), bouncing balls- (Lawson & Wollman, 1976), and mouse-task (Sodian et al., 1991). Some of the items deal with everyday situations such as plants that are grown under varying circumstances (Sodian et al., 1991), and other items with situations more abstract for the students, such as building airplanes (Bullock & Ziegler, 1999). From a psychometric view, the items differ with regard to three characteristics. The general characteristics of all 14 items, and how they differ in these three characteristics, are summarized in Table 1.2. All variations of the three characteristics are visible in the three example items in Figure 1.7.

First, five of the items are focused on assessing conclusive hypothesis testing, and nine items on assessing the control-of-variables strategy. One of the five items assessing conclusive hypothesis testing is provided in Figure 1.7B. In this item, children have to understand that attaching the carrot to a tree that both giraffes can reach was not a conclusive test and thus did not yield determinate evidence. Two of the nine items assessing the control-of-variables strategy are provided in Figure 1.7A & C. In the airplane item in Figure 1.7C, children have to understand that in order to find out whether the shape of the nose influences how much fuel an airplane uses, Mr. Miller has to build two airplanes that only differ in the shape of the nose, but in none of the other described relevant characteristics. Conclusive hypothesis testing and the control-of-variables strategy are typically not assessed within the same test. Whether children understand one of these two experimentation skills does however not indicate whether they also understand the other. In a basic experimental design, both aspects have to be taken account, because variation of the focal variable that allows determinate evidence has to be combined with the control of all other variables in order for an experiment to yield conclusive insights. The present test therefore has the
advantage that it can show whether children understand both of these crucial aspects of a basic experimental design.

Table 1.2

*Main characteristics of all items in the Variable Control Test*

<table>
<thead>
<tr>
<th>Item</th>
<th>Content</th>
<th>Reference</th>
<th>Construct</th>
<th>Format</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>Mouse</td>
<td>Sodian et al. (1991)</td>
<td>CHT</td>
<td>MC</td>
<td>Interpretation</td>
</tr>
<tr>
<td>Item 2</td>
<td>Airplane</td>
<td>Bullock &amp; Ziegler (1999)</td>
<td>CVS</td>
<td>MC</td>
<td>Design</td>
</tr>
<tr>
<td>Item 3</td>
<td>Parachute</td>
<td>CVS</td>
<td>MC</td>
<td>Design</td>
<td></td>
</tr>
<tr>
<td>Item 4</td>
<td>Ramp</td>
<td>Chen &amp; Klahr (1999)</td>
<td>CVS</td>
<td>CR</td>
<td>Evaluation</td>
</tr>
<tr>
<td>Item 5</td>
<td>Kite</td>
<td>CVS</td>
<td>MC</td>
<td>Design</td>
<td></td>
</tr>
<tr>
<td>Item 6</td>
<td>Giraffe</td>
<td>CHT</td>
<td>MC</td>
<td>Interpretation</td>
<td></td>
</tr>
<tr>
<td>Item 7</td>
<td>Nutcracker</td>
<td>CVS</td>
<td>MC</td>
<td>Design</td>
<td></td>
</tr>
<tr>
<td>Item 8</td>
<td>Bear</td>
<td>CHT</td>
<td>MC</td>
<td>Design</td>
<td></td>
</tr>
<tr>
<td>Item 9</td>
<td>Plant</td>
<td>Chen &amp; Klahr (1999)</td>
<td>CVS</td>
<td>MC</td>
<td>Interpretation</td>
</tr>
<tr>
<td>Item 10</td>
<td>Hammer</td>
<td>CVS</td>
<td>MC</td>
<td>Design</td>
<td></td>
</tr>
<tr>
<td>Item 11</td>
<td>Car</td>
<td>CVS</td>
<td>CR</td>
<td>Evaluation</td>
<td></td>
</tr>
<tr>
<td>Item 12</td>
<td>Bird</td>
<td>CHT</td>
<td>MC</td>
<td>Design</td>
<td></td>
</tr>
<tr>
<td>Item 13</td>
<td>Ball</td>
<td>CVS</td>
<td>CR</td>
<td>Evaluation</td>
<td></td>
</tr>
<tr>
<td>Item 14</td>
<td>Piglet</td>
<td>CHT</td>
<td>MC</td>
<td>Interpretation</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* CHT = conclusive hypothesis testing; CVS = control-of-variables strategy; MC = multiple choice single response; CR = constructed response.

Reference indicates original source of item cover story.

Second, the items differ in answer format. Some of the items deploy a multiple choice single answer-format, and some a constructed response format. The first type of answer format is that of multiple-choice-single-answer items. The items for conclusive hypothesis testing all have this multiple choice answer format. The answer format differs however within the nine items assessing the control-of-variables strategy. On the one hand, there are six items with a multiple choice single answer-format. These items encompass between three and four answer options. Of these, the right answer option is always that one according with the principle that only one variable is varied between experimental conditions (e.g., Figure 1.7C). Another three items assessing the control-of-variables strategy have a constructed response answer format (Figure 1.7A). These items combine a yes/no multiple choice answer asking whether a presented
experimental design was good to find something out with a short, open, written rationale. In this rationale, children are asked to provide a justification for their multiple choice answer (Warum ist das ein/kein gutes Experiment? Begründe deine Antwort). The Variable Control Test encompasses three such items on which children have to provide written open answers in the form of short rationales (Figure 1.7A). In all of these tasks, the students read descriptions of experimental designs and then choose and justify whether the respective design is appropriate for ascertaining the causal status of the focal variable. The crucial principle in all constructed response items is the control-of-variables strategy, because conclusive variation of the focal variable is given in all described experiments. Of the three constructed response items, two encompass consecutive descriptions of two experiments, and one item encompasses only one experiment. The items can thus be scored as three items, taking into account that they are based on three cover stories, or as five items, taking into account that overall there are five described experiments and answers. The principle of variable control is correctly applied in two of the five described experiments, and not correctly applied in three described experiments. In the example item (Figure 1.7A), children have to choose and describe that the described experiment is not good for scrutinizing whether a ball rolls farther on a steep ramp than on a flat ramp, because it is a confounded design with various non-controlled variables between the two conditions. The constructed responses serve two different purposes. First, they allow examining aspects of validity. That is, children’s open answers can be used to examine whether their reasoning behind the yes/no answer adequately reflects the assessment purpose of the item. Second, these constructed responses can be harnessed to examine in more detail children’s reasoning about experimental design. For scoring the constructed response items, either children’s yes/no choice can be scored, or their written constructed responses can be interpreted and scored. For coding and rating children’s answers, an elaborate categorization scheme was developed as part of the research in the present thesis, which is described in detail in Chapter 3.

Finally, the items differ in the stage of an experimental design that they are concerned with. Within the literature on the control-of-variables strategy, it has been acknowledged that there are four different facets in this regard that differ in educational relevance (Bryant, Nunes, Hillier, Gilroy, & Barros, 2015; Schwichow et al., 2016). These facets describe (A) whether children can choose the correct one out of a number of experimental designs, (B) whether they can make up a
design on their own, (C) whether they can correctly interpret evidence from a conducted experiment, and (D) whether they understand the rationale behind the control-of-variables strategy, that is, why a confounded experiment does not allow conclusions about the causal status of variables (Schwichow et al., 2016). In the present test, three of these four facets are covered. However, the facets differ not just for the control-of-variables items as in prior literature, but also for the conclusive hypothesis testing items. The first type of items in this regard is named here *design*-items (Table 1.2). In these items, children have to choose the correct one from a number of pre-defined experimental designs, as for example in (Figure 1.7C). In the second type, *evaluation* items, children have to judge whether a pre-defined experiment is good to find something out (Figure 1.7A). Finally, there are *interpretation*-items in which children have to draw correct inferences, depending on whether the respective principle was correctly applied (e.g., Figure 1.7B).

The variation among the 14 items on these three described characteristics reflects that the Swiss MINT Study, including the research in the present thesis, is aimed at examining children’s broad understanding of the two aspects of experimental design. A sum score across all items in the test yields information about a broad construct of basic understanding of experimental design, while splitting scores in accordance with either of the three major item characteristics can inform about children’s differential development on the different types of scores, and how these are influenced by educational interventions.
Introduction

Figure 1.7. Example items from experimental design test. (A) item 4: constructed response item assessing evaluation-aspect control-of-variables strategy; (B) item 6: multiple choice item assessing interpretation-aspect of conclusive hypothesis testing; (C) item 2: multiple choice item assessing design-aspect of control-of-variables strategy. Items adapted from Chen and Klahr (1999) & Bullock and Ziegler (1999).

1.3 Methodological Considerations

Some notes on the methodology and scientific practice in the present research are provided. First, a statistical description of mixture modeling follows, which is used in two studies of the present research. Then, a comment is provided on the research practices adhered to in the present
1.3.1 Mixture modeling

Mixture modeling is used in two of the four studies in this thesis. The mixture approach to statistical modeling has been first proposed by Pearson (1894) and it was further established as a tool for the social and behavioral sciences within Lazarsfeld & Henry’s framework of latent structure analysis (Lazarsfeld & Henry, 1966; Stouffer et al., 1950). In mixture modeling, a univariate or multivariate density is modeled as a linear superposition of more than one probability density function.

The Gaussian Mixture distribution is therefore a linear superposition of Gaussians:

\[ p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \]  

The observed density \( p(x) \) is the sum of \( K \) (univariate or multivariate) normal distributions with expectations \( \mu_k \) and variance-covariance matrices \( \Sigma_k \), weighted by the mixture weights \( \pi_k \) which are subject to:

\[ \sum_{k=1}^{K} \pi_k = 1 \]  

Thus they sum to 1, which implies that in a frequentist estimation framework they are identified with \( k - 1 \) degrees of freedom. A Gaussian mixture model is also called latent profile model.

Mixtures can also be modeled for categorical data, and both types of data can also be combined, such that the resulting model is the description of mixtures of mixed densities (Vermunt, 2004).

In addition, mixture models can also take into account dependencies of longitudinal data in an elegant way. In order to model which mixture (class or profile) a student or person belongs to over time, latent transition analysis can be modeled. A latent Markov structure is then added to the model, such that the probabilities of transitions between the discrete states of profile or class membership over time are modeled. The resulting equation is presented for the case of \( T = 2 \) time points, which is the maximum case in the present research, with \( I = 1, ... i \) categorical indicator variables (items) yielding answer vectors \( U \) across time points for \( k_t = 1, 2, ..., K \) mixtures:
\[
Pr(U = u | X = x) = \sum_{t=1}^{2} \sum_{k=1}^{K} \pi_k \tau_k | k (\prod_{t=1}^{2} \prod_{j=1}^{I} Pr[U_{ij} = u_{tj} | c_t = k | t])
\]  

(1.3)

where \(X\) can be a respondent-varying vector of covariates on which participants differ and which predict class or profile membership at time 1 and can also predict transition probabilities.

Mixture models serve both statistical and also theoretical purposes. From a statistical view, mixtures represent a nonparametric modeling approach that can basically be applied to any types of models. Mixtures owe their high range of applicability to the possibility of flexibly representing non-Gaussian distributions. This characteristic can first and foremost be used to model non-Gaussian data distributions (McLachlan & Peel, 2004). In Figure 1.8, this flexibility is demonstrated in mixtures of distributions which bear resemblance to spooky guises (Figure 1.8, upper panel) and to the shapes of Austrian mountain chains\(^2\) (Figure 1.8, lower panel).

\(^{2}\) Spooky mixture stems from

\[
spooky = \frac{250}{453} N(0, 30) + \frac{25}{453} N(-60, 5) + \frac{25}{453} N(60, 5) + \frac{25}{453} N(-61, 1) + \frac{25}{453} N(-59, 1) + \frac{25}{453} N(59, 1) \\
+ \frac{25}{453} N(61, 1) + \frac{50}{453} N(-27.5, .5) + \frac{50}{453} N(27.5, .5) + \frac{5}{906} N(0, .1)
\]

and mountain chain mixture from

\[
alpine = \frac{15}{92} U(-2, 4) + \frac{75}{184} N(0, 1.2) + \frac{25}{184} N(3, 1) + \frac{25}{92} N(6, 1) + \sum_{i=1}^{4} \frac{10}{184} N(i - 4, 1)
\]
Introduction

Figure 1.8. Funky mixture densities demonstrating the flexibility to model non-Gaussian data: A spooky mixture (upper panel), an Alpish mixture sharing similarity with the Grossvenediger-massif (lower panel).

This great flexibility can be harnessed for a myriad of further purposes. For example, it has recently been argued that in hierarchical modeling, a Gaussian assumption for the distribution of random parameters is often both overly strict and difficult to interpret. As a flexible alternative, an alternative nonparametric depiction in the form of a mixture has been proposed (Vermunt, 2003). In addition, the mixture assumption can also be applied to a promising new alternative of hypothesis testing, by setting up a Bayesian mixture model and estimating which mixture (i.e., models representing predictions of different hypotheses) data stem from (Kamary, Mengersen, Robert, & Rousseau, 2014; Robert, 2016). Finally, from a theoretical view, mixtures offer an elegant means for representing the assumption that model parameters systematically differ between different populations, for example groups of learners who differ in their knowledge structure (Harring & Hodis, 2016; Hickendorff, Edelsbrunner, McMullen, Schneider, & Trezise, in press). In the present thesis, this characteristic of mixture models is used to model and test predictions of different theories about learning and development.

Parameter estimation is typically achieved via an expectation-maximization algorithm yielding maximum likelihood estimates (McLachlan & Peel, 2004). For applied research, this
algorithm has the advantage that full information estimation is possible. This estimation uses all available information in the presence of partial missing data on cases, which prevents increased estimator variance by dropping cases with missing data.

A conceptual introduction to mixture modeling closer to the present context is provided in the relevant Chapters 3 and 4. Further information on mixture modeling is provided in McLachlan and Peel (2004, a broad statistical and applied overview) and Hickendorff et al. (in press, a conceptual introduction for research on learning and instruction).

1.3.2 Open Science and Methodology

In the conduct of the present research, I put emphasis on an open and transparent approach to science. Research materials and data have been stored openly at the Open Science Framework, and they have been attached at any journal submissions. This includes raw data, analytic data, \LaTeX-syntax and files for compiling the present document, and syntaxes for conducted analysis. This should allow any reviewers and other researchers to reconstruct and reproduce the research conduct and analysis reported in this thesis as closely as possible. The aim of these measures is to be transparent in my research endeavors because this helps understanding, critically reflecting, replicating and building on each other’s research and insights - and, from a practical perspective, helping each other and learning together. By sharing all scripts I can also share my expertise in data analysis and data visualization.

Open science takes a lot of time and effort. For my whole thesis research I publicly share data and scripts on my Open Science Framework-repository (https://osf.io/2agrk/) and for the simulations in Chapter 2 on GitHub (https://github.com/peter1328/simrasch). Not all data, materials and scripts are yet public as this was not always possible, but everything admissible will be uploaded. I recently finished my first preregistration and it helped me immensely develop my study and analysis planning skills. Open science is like the R statistics software package: Learning takes a lot of time but it benefits oneself and the community.

In my thesis I also tried to circumvent some shortcomings of common statistical methods. I tried to apply statistical models that are consistent with my theories and methods instead of relying on established practices that may be easier to apply but do not reflect my theories sufficiently well (such as Rasch modeling, which will be in focus in Chapter 2). Conceptual
introductions to these specific statistical approaches are provided in the relevant chapters, to aid readers’ following who are not familiar with these methods. This includes the application of mixture modeling (Chapter 3 & Chapter 4), and of Bayesian statistics in some analysis (Chapter 3 & Chapter 4). I did not always apply a Bayesian framework because for some types of analysis it is not as well understood and readily available as the classical frequentist framework. But I believe that the future of statistics is Bayesian and that it will influence our practices and inferences in future psychological and educational research.

References


Schwichow, M., Christoph, S., Boone, W. J., & Härtig, H. (2016). The impact of sub-skills and item content on students’ skills with regard to the control-of-variables strategy.


Chapter 2

Psychometric Issues in Research on Scientific Reasoning

Peter Edelsbrunner
ETH Zürich

Abstract

In the past decade, psychometric modeling has been frequently applied in research on scientific reasoning. A review is conducted to examine psychometric modeling practices in this field, including model choice, fit investigation practices, and the relation between psychometric models and researchers’ inferences. The review, examining 11 empirical research studies, reveals that the currently predominant psychometric approach in research on scientific reasoning is Rasch modeling, applied in a way strongly similar to practices in national and international large-scale educational measurement programs. This approach is discussed to provide strong measurement characteristics. However, corroborated through statistical data simulations, it is argued that these current practices are limited in informing about psychological structure, and various inferences researchers have drawn about the structure and cognitive basis of scientific reasoning might be unwarranted. In addition, inconsistent practices in setting up structural models and interpreting fit statistics undermine judging of the adequacy of researchers’ inferences. It is suggested that in future work researchers fully report important psychometric details and rationales for their approaches, take into consideration a broader range of psychometric modeling and fit testing.
practices, and pre-register confirmatory psychometric analysis in order to distinguish hypothesis testing from exploratory modeling. In a concluding section, the significance of these issues for research on scientific reasoning is discussed, and for further fields related to Psychology, Education, and beyond.

**Keywords**: scientific reasoning; item response modeling; Rasch model; literature review; simulation study

### 2.1 Introduction

Psychometric modeling has recently entered research on scientific reasoning. Some researchers in this field now apply psychometric modeling to examine the quality of assessment instruments, to scale obtained data, and to test hypotheses regarding the structure and correlates of scientific reasoning (Hartmann, Upmeier zu Belzen, Kroeger, & Pant, 2015; Körber, Osterhaus, & Sodian, 2015; Mayer, Sodian, Körber, & Schwippert, 2014). This might be seen as a favorable development, because psychometric modeling offers the opportunity to scale scores in accordance with specific models, and to explicitly model, estimate, and test assumptions that otherwise often stay implicit and untested (Raykov & Marcoulides, 2011). For example, psychometric modeling can be used to examine the reliability and some aspects of the validity of assessment instruments, and to examine whether the usage of a single overall score, or of various scores representing the variegated facets of scientific reasoning might be useful (Boone, Staver, & Yale, 2014; Reise, 2012; Reise, Bonifay, & Haviland, 2013).

Psychometric modeling is, however, attached to various issues that demand attention, because it goes hand in hand with strong assumptions about the investigated constructs and the modeled data (Christensen, 2013a). There are various schools of psychometrics that differ in their views on how to handle these assumptions. Which psychometric models should be applied, how the model structure should be defined, and whether and to which degree model fit is important for the interpretation of model estimates, are questions that proponents of different schools handle differently (Andrich, 2004; Linacre, 2010). How researchers deal with these psychometric issues can substantially influence their estimates from psychometric modeling, and ensuing inferences. It is therefore time to review psychometric modeling practices in research on scientific reasoning, and how prevalent practices may be related to theory development in this field.
In the present study, a review of psychometric modeling practices in research on scientific reasoning is undertaken. The review examines the prevalent practices in selecting and specifying psychometric models, investigating model fit, and interpreting results based on these practices. Examining these practices is aimed at elucidating the current role of psychometric modeling in this field, and at developing suggestions for how psychometric modeling might be handled in this field in order to optimally support theoretical progress.

Following, in order to provide the necessary background, an overview of the cornerstones of major theories about scientific reasoning is provided. This is followed by an overview of key issues in modern psychometric modeling, focusing upon different schools of psychometric modeling that favor different approaches to psychometrics. Then, the main review is reported, in which current practices in psychometric modeling in research on scientific reasoning are examined and summarized. The review focuses on researchers’ rationales for applying psychometric modeling, choosing specific models and model fit testing procedures, and on the inferences researchers draw based on results obtained from psychometric modeling. It is then discussed how the current practices might relate to theoretical developments in the field, and suggestions are developed for future modeling and reporting practices.

Key results from the literature review are 1) the prevalence of a specific Rasch modeling measurement perspective in research on scientific reasoning, 2) inconsistencies in researchers’ psychometric modeling and fit examination practices, which undermine judging the appropriateness of researchers’ inferences, 3) interpretations of psychometric dimensions as representations of traits, processes, and dynamics, which, as will be discussed, might go beyond the general potential of cross-sectional psychometric data. Backed up by statistical data simulations, these three insights will be discussed and then it will be reflected whether and how current practices might be expanded in order to optimally support theoretical progress in this field.

2.1.1 Scientific Reasoning

In accordance with the aim of the present review, an overview is provided of major theories of scientific reasoning, focusing on assumptions about the number and interrelations of components assumed to compose scientific reasoning. For more comprehensive overviews of theoretical specifics and further theories, interested readers are referred to Morris, Croker, M., and

Scientific reasoning encompasses the reasoning and problem-solving skills involved in generating, testing and revising hypotheses or theories, and in the case of fully developed skills, reflecting on the process of knowledge acquisition and knowledge change that results from such inquiry activities (Morris et al., 2012). Research on scientific reasoning emerged in the 1950s with the work of Piaget and Inhelder (1958) and Bruner, Goodnow, and Austin (1956) who started examining how children’s understanding of general principles of inquiry develops, and how they acquire concepts. Research on scientific reasoning continued within two related broad strands that focused on different facets of scientific reasoning: On the one hand, examinations of children’s understanding of general principles of scientific inquiry, such as the understanding of indeterminacy and variable control in experimental design (D. Kuhn et al., 1988; Somerville, Hadkinson, & Greenberg, 1979), and on the other hand examinations of children’s acquisition and development of concepts about scientific phenomena, such as the shape of the earth in geography, and density in physics (Carey, 1985; C. Smith, Carey, & Wiser, 1985).

These two strands of research were unified under a general model of scientific reasoning that has remained central throughout the last decades: The Scientific Discovery as Dual Search (SDDS) model (Klahr, 2002; Klahr & Dunbar, 1988; Zimmerman, 2000). In the SDDS model, scientific reasoning is conceptualized as a problem solving process taking place in a hypothesis space and an experiment space with the aim to develop and revise hypotheses that can explain empirical evidence. To engage in this problem solving process, three broad cognitive processes are assumed to be necessary, which constitute the main components of the SDDS model: Hypothesis generation, experimental design, and evidence evaluation (Figure 2.1).
Hypotheses are generated based on prior knowledge about the investigated domain, mapped onto experiments that are conducted in order to produce evidence to test the hypotheses. In the last step, the evidence is then interpreted in order to evaluate whether it promotes a hypothesis, or whether the hypothesis has to be changed in order to account for the accumulated evidence. In scientific inquiry, these three broad reasoning processes take place in an iterative manner. This model of scientific reasoning has served as a fruitful framework in many studies across the last three decades and also to synthesize evidence about scientific reasoning (Morris et al., 2012; Zimmerman, 2000, 2007). In this widely deployed model, scientific reasoning is thus assumed to be based on three broad components which represent interrelated facets of psychological processes.

Various later-developed models of scientific reasoning were built on the three components identified by Klahr and Dunbar (1988). In later models, the three components were further distinguished on a smaller grain size, or additional components were added, where components usually refer to psychological traits, processes, skills, or competencies. For example, early cognitive-developmental studies on scientific reasoning focused on examining the development of children’s understanding of the control-of-variables strategy. This strategy deals with ascertaining the causal status of a variable by keeping potential confounders constant (Siegler & Liebert, 1975). Later studies acknowledged the importance of additional, more complex components of experiment design and broader scientific reasoning, such as coordinating the effects of multiple causal variables (D. Kuhn & Pease, 2008). These later models also added further general components such as epistemic cognition, that is, beliefs and resulting cognition about knowledge
and its development (Cano, 2005; Kitchner, 1983), and recently the component of argumentation has been emphasized as a major component (Fischer et al., 2014; D. Kuhn, Iordanou, Pease, & Wirkala, 2008; D. Kuhn & Pease, 2008; D. Kuhn & Udell, 2003). As an example of a relatively broad model, in a recent synthesis eight interrelated components of scientific reasoning were conceptualized that included components similar to those of the SDDS model, and additional components such as problem identification and questioning (Fischer et al., 2014). Thus, there are different models of scientific reasoning that are somehow related to the SDDS model, yet differ in their breadth and level of detail.

A further major theory emphasizes scientific reasoning as a process of coordinating theory and evidence (D. Kuhn, 1989). Developmental studies have shown that children have great difficulties in distinguishing between their own theories and empirical evidence, and that mixing up these two ontologically related yet distinct entities is a major issue in the development of scientific reasoning (Körber, Mayer, Osterhaus, Schwippert, & Sodian, 2014; D. Kuhn, 1989; Sodian, Zaitchik, & Carey, 1991). It has therefore been assumed that the skilled coordination of theory and evidence plays a key role across the components of scientific reasoning (Körber et al., 2014). It is yet unclear to which degree this key role exists and constraints the general development of scientific reasoning, but an emphasis on this skill is put in recent efforts to support holistic development of scientific reasoning with elaborate educational interventions (D. Kuhn, Ramsey, & Arvidsson, 2015).

This short overview shows that scientific reasoning has been conceptualized in various ways. What most models have in common, however, is that they describe scientific reasoning as a complex set of interrelated components (Fischer et al., 2014; D. Kuhn & Pease, 2008; Zimmerman, 2007). In the present study, it will therefore be in focus how researchers conceptualize the components of scientific reasoning, and how this is reflected in their psychometric modeling practices.
2.1.2 Psychometric Modeling

In psychometric modeling, probabilistic models are applied that describe how characteristics of persons and items relate to the probability that persons solve items or not\(^1\). The most relevant model for the present study, the Rasch model, is described in its basic form, and then it is described how further models relate to the Rasch model.

**The Rasch Model.** The Rasch model is a psychometric model in which the probability that a person solves an item is predicted from an additive relation between the person’s ability level on a continuum describing a latent (not directly observed) trait, and the item’s difficulty level on the same continuum. The Rasch model is presented in Equation 2.1.

\[
p(x = 1) = \frac{\exp(\theta_p - \sigma_i)}{1 + \exp(\theta_p - \sigma_i)}
\] (2.1)

The model entails two types of parameters. Conceptually, the person parameter \(\theta_p\) describes the ability of a person working on the instrument, and the item parameter \(\sigma_i\), describes an item’s difficulty. The two parameters combine as follows: The probability of a person \(p\) to solve an item \(i\) depends on the difference between the person’s ability, \(\theta_p\), and the item’s difficulty, \(\sigma_i\). If both parameters are equal, the probability is .5 that the person will solve the item. If the ability parameter is smaller than the item parameter, the solution probability is lower than .5, and vice versa. The function linking the difference between person ability and item difficulty to the probability of solving an item is not a (linear) identity function but logarithmic. It yields a logistic link between parameters and observed data. The higher a person’s ability is, the higher the probability to solve an item, but the steepness of the function decreases asymptotically towards the probability boundaries of 0 and 1. In Figure 2.2A, the resulting item solution probabilities depending on person parameters \(\theta_p\) are plotted.

These item characteristic curves (ICCs) are plotted for two items (Figure 2.2A): One item with difficulty -2 (Figure 2.2A, left curve), meaning that persons with ability parameter \(\theta = 2\) have a 50% probability to solve this item. Another item with \(\beta = 0\) is plotted (Figure 2.2A, right curve), which is more difficult because a higher person parameter of \(\theta = 0\) is needed to solve this item.\(^1\)

---

\(^1\)the case described here counts for the frequent case of dichotomous (e.g., true-false) data, but as noted further below, models for multinomial data exist as well.
A Rasch person parameter ($\theta$) is visible that the two items’ ICCs are parallel and therefore do not overlap at any point, a crucial characteristic of items in the Rasch model, as will be discussed.

**Figure 2.2.** Item characteristic curves (ICCs) of two test items with all discriminations restricted to the same estimate (Rasch model, left), with discriminations freely differing between items (2PL model, right). Grey dashed lines indicate location of 50% solution probability, corresponding to items’ difficulty parameters.

This is the most basic form of the Rasch model. Models falling within the broader category of Rasch models can also be more complex. There can also be more than one person parameter or item parameter. The decisive characteristic of all models falling within the category of Rasch models is that answer probability is modeled simply as an additive function of the person and item parameters, which keeps the crucial characteristics of Rasch models. There are further and more complex Rasch models, for example for items with more than two ordinal answer categories, multidimensional for more than one person ability dimension, and mixture models for modeling subgroups with differing item difficulty parameters (e.g., for modeling development and change) (Jeon, Draney, & Wilson, 2015).

The simple additive relation of parameters in the Rasch model entails various favorable practical properties. When an instrument is in accordance with the Rasch model, its items can be interpreted as a sample from a universe of possible items that all assess the same trait to the same extent for any sample of test takers. Only the items’ difficulty differs. This allows sum scores from the instrument to be utilized as sufficient statistics; all information about a test taker
regarding the assessed trait is depicted in the sum score. *Specific objectivity* is given, and there is no need to examine further which specific items have been solved and which specific characteristics the test taker has; the sum score simply tells the same story about any test taker on any item. Due to measurement error, which is an important part of the Rasch model (Heene, 2013), the choice and number of items will only influence the reliability of person parameter estimates, but not their interpretation. This property is crucial for applied issues such as the possibility to link scores of instruments using partially different items. Test linking allows to draw valid inferences about ability development in longitudinal studies and in large scale studies in which by design not each participant works on all items (Davier, 2011).

These favorable properties bring statistical and theoretical burdens in the form of strong assumptions. The most common and general assumption of the Rasch model is unidimensionality. This assumption arises from the presence of only a single person parameter, $\theta$, that is modeled as the sole source of systematic variance in the item solution probabilities of different test takers. This dimension is assumed to be the same across all people, and across all items. Therefore, the trait should sufficiently well explain systematic differences in the probability that different persons solve different items of an instrument. From a statistical view, this entails that all correlations between the items can be explained by modeling this parameter - an assumption termed *local independence*. This assumption is closely linked to the stronger assumption of *local stochastic independence* which implies that when the trait is modeled, there are no further systematic influences on test takers’ item solution probabilities. Hence, when controlling for the trait, the test takers’ answers across the items are assumed to be stochastically independent of each other. Finally, the assumption most clearly demarcating the Rasch model from other models is that of equal item discriminations. It is assumed that all items assess the underlying latent trait to a similar extent, which relates to the demarcation of the Rasch model from further psychometric models.

The strong requirements of the Rasch model demand statistical fit examination. The most common suggested fit examination procedure among proponents of the Rasch school are item infit statistics (Christensen & Kreiner, 2013). Item infit statistics are based on the raw residual $R_{pi}$ (difference between predicted and observed score) of the score of person $p$ on item $i$, which is
Then, the item infit statistics are

\[ INFIT_i = \frac{\sum_{p=1}^{2} R_{pi}^2}{\sum_{p=1}^{2} \text{VAR}(X_{pi})} \quad (2.3) \]

Item infit statistics thus quantify the ratio of the variance predicted by the Rasch model and the empirically observed variance on an item. Infit statistics have an expected value of 1 under the Rasch model and their range is non-negative real numbers. Suggested cutoff values vary from broad 0.5-1.5 (Wright, 1979) to cutoffs for high stakes testing instruments with 0.8-1.2 (Wright, 1994). Generally, the most common suggestions for examining fit in the Rasch model encompass these and further fit statistics for items and persons, which follow the general principle that they provide estimates how closely items and persons behave as predicted by the Rasch model. The degree to which fit statistics indicate close fit between empirical data and the model, *invariant measurement* is given across any items and participants (Engelhard Jr, 2013), such that the same trait is measured across any persons and items.

**Item Response Models.** Another type of psychometric models is the family of item response models, which is associated with the broader framework of item response theory (IRT, see Sijtsma, 2011). Similar to Rasch models, in item response models the probability to solve items is modeled depending on characteristics of persons and items. The Rasch model in its various appearances can even be seen as part of the family of item response models. The decisive difference between the two families of models is that Rasch models only encompass those models in which relations between item and person parameters are modeled as additive. Item response models are not limited to models with additive functions. This makes item response models more general, but comes at the cost of characteristic bound to a purely additive relation between person and item parameters.

For example, a common item response model that does not fall within the category of Rasch models is the 2pl model (Thissen & Steinberg, 1986). In this model, not only a difficulty parameter for items is included, but also a second item parameter that describes differences in item discriminations. An item discrimination is conceptually similar to a factor loading in factor analysis: It describes how strongly the probability to solve an item is related to the underlying
person ability. Including item discrimination parameters describes that different items indicate person ability to different extent. This assumption might seem intuitively useful, since it might be difficult to achieve a set of items that all reflect the measured ability to a similar extent. Item discrimination parameters are however multiplied with the difference between person ability and item difficulty parameters in the 2pl model. This goes beyond a purely additive relationship of parameters. This multiplicative relation thus yields loss of the additivity characteristic, which by some researchers is deemed fundamental for measurement.

ICCs for two items under the 2pl model are plotted in Figure 2.2B. It can be seen that in comparison to items under the Rasch model (Figure 2.2A), ICCs under the 2pl model are not parallel and therefore do overlap. In the figure, two items are plotted that have the same difficulty parameter (0) but the item with the steeper curve has a higher discrimination parameter. The multiplicative relation of the discrimination and difficulty parameters induces the characteristic that the item curves are not parallel and therefore do overlap at some point. Overlapping ICCs imply that for persons with ability lower than the crossing point, items have a different difficulty order than for persons with ability above the crossing point.

Further parameters can be added to item response models, such as parameters that represent varying guessing probability (giving the right answer by chance, for example on multiple choice tasks) between items, and varying slipping probability (not giving the right answer by chance, for example because items have different distracting elements) (Revelle, 2004).

2.1.3 Two Schools of Psychometric Modeling

The often emphasized demarcation of Rasch models from further item response models has significant theoretical reasons. When person and item parameters are related in an additive manner, ICCs thus do not overlap. This keeps the characteristic that items have the same difficulty order for persons independent of their level on the ability continuum. Conceptually, this relates to the issue that the same trait is measured for different people, independent of their ability level. Two schools of psychometric modeling have developed since the 1960s that put different
emphasis on this characteristic. Which of the schools is followed therefore has implications for psychometric modeling practices and interpretations.

The first school is focused around Rasch modeling and named here the *Rasch school*. This school is based on a definition of measurement put forward for example by Thurstone (1928). This definition of measurement implies that in order for something to represent measurement, it must yield results that are independent of the measurement instrument used, and of the population it is applied on:

> A measuring instrument must not be seriously affected in its measuring function by the object of measurement (Thurstone, 1928)

This property, according to proponents of the Rasch school, is only given for Rasch models, due to the additivity characteristic, but not for any other psychometric models.

The second school of psychometric modeling is named here the *item response theory* (IRT) school. For proponents of the IRT school, psychometric modeling is statistical modeling in a broader sense than in the Rasch school. The aim of psychometric modeling in the IRT school is to construct models that fit data (Andrich, 2011). These models are built on the basis of theoretical assumptions about the relation of person and item characteristics with the probability of answers. The critical difference to the Rasch school is the status of the model. Under the Rasch school, the psychometric model is presumed fixed, and the data have to fit the model, because the Rasch model is conceptualized as representing measurement, and therefore it cannot be exchanged for any other model. Under the IRT school, the data are presumed fixed, and the model has to fit the data, and therefore model fit has higher priority than the data. If an initial model shows misfit to the data, it is usually adapted and modified until it is well in accordance with empirical data (Divgi, 1986; Hambleton, 2000).

These diverging viewpoints on model fit evoke different practices and interpretations. Proponents of the Rasch school do exclusively apply Rasch modeling, because they deem the Rasch model the only informative measurement model, while proponents of the IRT school choose whatever type of model fits well and is deemed informative. Regarding model fit, Rasch paradigms and their relation to development in science has been contested (Bird, 2013; Toulmin, 1974); thus we deem the concept of the two paradigms described by Andrich (2004) not yet sufficiently elaborate and critically reflected to be accepted.
proponents tend to keep things approximate (although different stringency habits exist within Rasch proponents; Linacre, 2010): Absolute fit is usually not in focus. Rather, approximate fit is examined - for example, whether item variances and person answer patterns are more or less what the model predicts (Christensen & Kreiner, 2013). It is evaluated whether data fit the Rasch model sufficiently well such that parameter estimates will be sufficiently exact for respective purposes (e.g., whether the aim is to estimate group level statistics, or high-stakes testing Linacre, 2006; Linacre & Wright, 2000; Wang & Chen, 2005). Also, particularly under the Rasch school it is acknowledged that absolute unidimensionality is in most cases not achievable. Instead, proponents of the Rasch school examine whether a psychometric dimension can be extracted across all items, and whether residual dimensions and dependencies are so small that they do not substantially bias parameter estimates on the extracted common Rasch dimension (Christensen, 2013b, 2013c). IRT proponents look more closely at fit of their model to the data, using a variety of model fit statistics and testing procedures (Maydeu-Olivares, 2013; Maydeu-Olivares & Montaro, 2013; Suarez-Falcon & Glas, 2003). There are various fit statistics to evaluate overall model fit (Maydeu-Olivares, 2013; Maydeu-Olivares & Montaro, 2013) and assumptions such as unidimensionality (Khalid, 2009), local independence (Koller, Maier, & Hatzinger, 2015), person invariance (Andersen, 1973), and item invariance (Khalid, 2009). IRT proponents do also often conduct multiple differential item functioning (DIF) analysis. In DIF analysis, it is examined whether item parameters are comparable between subgroups such as genders or across time (Hambleton, Swaminathan, & Rogers, 1991; Siersma & Eusebi, 2013).

The two described schools matter for the present review because they differ not only in researchers’ tendencies to choose specific models, but also to apply specific fit investigation practices, both of which may differently impact researchers’ theoretical inferences. Practicing either school may thus influence theoretical inferences in research on scientific reasoning. In the present review, it will therefore be examined in accordance with which school researchers apply psychometric modeling, and whether and how this might relate to researchers’ inferences.

2.2 Literature Review

A literature review was undertaken to examine psychometric modeling practices in research on scientific reasoning. Studies were reviewed in which psychometric modeling was applied on
Psychometric Issues

instruments meant to assess scientific reasoning comprehensively, with the aim either to psychometrically validate the instrument, or to relate scientific reasoning to other variables in order to test hypotheses or explore a research question. Inclusion criteria thus encompassed that research had to be based on empirical data assessed with an instrument meant to measure scientific reasoning, analyzed with psychometric models (which excludes classical test theory, in which sum scores are analyzed without an explicit psychometric model\(^3\)).

National and international large-scale assessment programs were added to the review for reasons of comparison. Large-scale assessment programs in which scientific reasoning is assessed are not reviewed in detail in the present study. They serve however as comparison studies, to be able to see commonalities and differences between practices in research on scientific reasoning, and practices in large-scale assessments on scientific reasoning\(^4\). Large-scale programs such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), and further national programs enjoy reputation of high quality methodology in researcher communities. It was thus assumed that practices in research on scientific reasoning could be influenced by large-scale assessment programs, and therefore these were included for comparison.

Literature search was undertaken in the PsycInfo and Scopus databases because PsycInfo is specialized in Psychology and related fields such as Education, and Scopus was chosen as a complementary broader database. Due to the different nature of the databases, a search including all search fields was undertaken in PsycInfo, while in Scopus search fields were limited to include title, abstract, and keywords. Search terms were based on a recent literature review that was undertaken on the topic of assessment instruments for scientific reasoning (Opitz, Fischer, & Heene, submitted). This review includes comprehensive a list of assessment instruments, and whether psychometric modeling was applied on them. Instruments mentioned in this review were

\(^3\)the distinction between classical test theory and modern psychometric modeling is however not clear-cut; factor analysis used to be regarded as an instrument of classical test theory, which, in its confirmatory versions, is similar to the psychometric models discussed here (Gebhardt, 2016).

\(^4\)It might be argued that large-scale assessment programs also represent kind of research, particularly because data gathered in these studies are often used for ancillary research. The aim of all of these programs is however applied policy-guiding assessment and thus the methodology is aligned with this aim, rather than with the aim to conduct research.
included in the present study, and to not miss further and newer instruments (the present review includes studies published until December 2016), the additional literature search was undertaken. Search terms were based on an extension of the search query used in Opitz et al. (submitted). The search terms and search process are depicted in Figure 2.3. Literature from the search in the two databases was first screened and to the obtained literature two further sources were added from Opitz et al. (submitted) that fulfilled our criteria.

The literature search resulted in 11 empirical research studies that fulfilled the inclusion criteria (Figure 2.3), and five large-scale assessment programs for comparison. In addition to nine references obtained from database search, seven references were added from Opitz et al. (submitted). These included two dissertations describing empirical research studies, and five national and international large-scale assessment programs in which comprehensive instruments for the assessment of scientific reasoning were applied using psychometric modeling (study abbreviations: IQB, NAEP, NASP, PISA, TIMSS).

Exclusion criteria were applied in a hierarchical manner (Figure 2.3): In a first screening of titles, abstracts and keywords, it was examined whether studies were concerned with scientific reasoning. If studies were likely to be concerned with scientific reasoning, full texts were obtained. After verification of topical relevance, it was checked whether studies were empirical and concerned with validating an assessment instruments, or a theoretical question about scientific reasoning. Then, it was screened whether comprehensive assessment of scientific reasoning took place (to exclude studies interested in individual facets), whether psychometric methods were applied, and then whether the psychometric methods fell under modern psychometric modeling as discussed here.
2.3 Results

The review results are presented in four parts. A summary is provided of the practices undertaken in large-scale assessment programs. Then, a summary of the review of the 11 empirical research studies is undertaken, focusing on three main topics: First, general strategies in setting up psychometric models are considered, specifically how the choice of psychometric model relates to researchers’ theoretical conceptualizations of scientific reasoning. Second, researchers’ practices in the investigation of statistical fit between models and data are summarized, to yield an overview what the most common and less common fit investigation practices are. Third, inferences that researchers draw based on results obtained from

---

5Research related to scientific reasoning is also conducted under the terms of scientific inquiry and scientific thinking. For terminological coherence, in the description of the reviewed studies these and related terms are replaced with the term scientific reasoning.
psychometric modeling are summarized. These three main topics are considered in turn, and afterwards insights from this review are summarized.

2.3.1 Large-scale assessment programs

A summary of the psychometric practices in the five large-scale assessment programs is provided in order to be able to evaluate how closely empirical research studies follow similar practices. The general approach is similar in all five programs: The psychometric model in all these programs is some form of the Rasch model (Ainley, Fraillon, & Freeman, 2007; Pant et al., 2013; M. Wu, 2004), applied unidimensionally so that it yields a single scaled score of scientific reasoning. Fit is mostly investigated in accordance with the Rasch school; fit of items and persons to the model is investigated in all of the five programs, and to different degree it is also examined whether items function similarly across groups of different persons or regions. These practices in large-scale assessment programs reflect that the Rasch school fits the focus of educational measurement: The aim is not to find a psychometric model that fits obtained data as closely as possible, which is generally more strongly in focus in the IRT school. The aim in educational measurement programs rather is to develop items and improve fit of obtained data to the Rasch model in an iterative process, reflecting the aim to achieve high quality measurement in accordance with the Rasch school.

2.3.2 Psychometric modeling practices

General characteristics, theoretical background and psychometric modeling practices of the 11 empirical research studies are summarized in Table 2.1.

The 11 reviewed empirical research studies generally showed mixed focus between instrument development and theoretical research: Some of the studies (four of 11) were concerned mainly with instrument development to assess scientific reasoning, another four studies had a focus on testing theoretical assumptions or exploring substantial research questions, and three studies combined both aims (Table 2.1). Samples sizes ranged from 155 to 2247 students with a median of 848 students, thus in the studies generally large sample sizes were obtained. Item numbers ranged from 5 to 166 items with a median of 45 items. Sample size and item numbers showed a Spearman correlation of .37, indicating that researchers aligned sample sizes
and item numbers to a moderate degree. As psychometric software, in most, specifically nine of the studies, the ConQuest software package was used (M. L. Wu, 2007). This software package is for example also used in the PISA study, to scale items and for some hypothesis tests. Some studies used the WinSteps (Linacre & Wright, 2000), FACETS (Linacre, 2012), and WinMIRA (von Davier, 2001) packages.

Concerning theoretical background, the studies varied greatly in the different components they described as main constituents of scientific reasoning. These components were described as sub-competencies (Grube, 2010), primary skills (Lou, Blanchard, & Kennedy, 2015), ability components (Körber et al., 2014; Kuo, Wu, Jen, & Hsu, 2015), and processes consisting of several skills (Nowak, Nehring, Tiemann, & Upmeier zu Belzen, 2013). In three studies, scientific reasoning was conceptualized to consist of one such general component, in four studies four components were conceptualized, in one study six components (Lou et al., 2015), and in three of the studies both one general component and alternatively three (Nowak et al., 2013) respectively five components (Körber et al., 2014), and seven components (Hartmann et al., 2015) were conceptualized (Table 2.1). There is thus no consensus how many components scientific reasoning has, but the average theoretical model assumes about four (interrelated) components.

---

6Notably, these packages are all commercial software, and in none of the studies free software such as the numerous packages available in the R software (for example the TAM package which is very similar to ConQuest but more versatile (Kiefer, Robitzsch, Wu, & Robitzsch, 2016)
Table 2.1

*Main characteristics and psychometric practices in reviewed studies.*

<table>
<thead>
<tr>
<th>Reference</th>
<th>Focus</th>
<th>Sample</th>
<th>Items</th>
<th>Software</th>
<th>Theory</th>
<th>Model</th>
<th>maxfit</th>
<th>relmod</th>
<th>ifitmod</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown et al. (2010)</td>
<td>I</td>
<td>343</td>
<td>29</td>
<td>CQ</td>
<td>4D</td>
<td>4D</td>
<td>na</td>
<td>1D</td>
<td>1D</td>
</tr>
<tr>
<td>Esswein</td>
<td>I, T</td>
<td>1613</td>
<td>27</td>
<td>WS</td>
<td>4D</td>
<td>1D</td>
<td>na</td>
<td>1D</td>
<td>1D</td>
</tr>
<tr>
<td>Grube (2010)</td>
<td>T</td>
<td>1562</td>
<td>60</td>
<td>CQ</td>
<td>4D</td>
<td>1D, 4D</td>
<td>4D</td>
<td>1D</td>
<td>1D</td>
</tr>
<tr>
<td>Hartmann et al. (2015)</td>
<td>T</td>
<td>578/2247</td>
<td>166/123</td>
<td>CQ</td>
<td>1D, 7D</td>
<td>1D</td>
<td>na</td>
<td>1D</td>
<td>1D</td>
</tr>
<tr>
<td>Heene (2007)</td>
<td>T</td>
<td>443</td>
<td>5</td>
<td>CQ, WS, FC, WM</td>
<td>1D</td>
<td>1D</td>
<td>na</td>
<td>1D</td>
<td>1D</td>
</tr>
<tr>
<td>Koerber et al. (2014)</td>
<td>I, T</td>
<td>1581</td>
<td>66</td>
<td>CQ</td>
<td>1D, 5D</td>
<td>1D</td>
<td>na</td>
<td>1D</td>
<td>1D</td>
</tr>
<tr>
<td>Kremer et al (2014)</td>
<td>I, T</td>
<td>1553</td>
<td>24</td>
<td>CQ</td>
<td>4D</td>
<td>4D</td>
<td>na</td>
<td>4D</td>
<td>4D</td>
</tr>
<tr>
<td>Kuo et al (2015)</td>
<td>I</td>
<td>1066</td>
<td>114</td>
<td>CQ</td>
<td>4D</td>
<td>1D, 4D</td>
<td>-</td>
<td>4D, 1D</td>
<td>1D</td>
</tr>
<tr>
<td>Lou et al (2015)</td>
<td>I</td>
<td>630</td>
<td>30</td>
<td>WS</td>
<td>6D</td>
<td>1D</td>
<td>na</td>
<td>1D</td>
<td>1D</td>
</tr>
<tr>
<td>Mayer et al. (2014)</td>
<td>T</td>
<td>155</td>
<td>20</td>
<td>CQ</td>
<td>4D</td>
<td>1D</td>
<td>na</td>
<td>1D</td>
<td>1D</td>
</tr>
<tr>
<td>Nowak et al. (2013)</td>
<td>I</td>
<td>537</td>
<td>90</td>
<td>CQ</td>
<td>1D, 3D</td>
<td>1D, 3D</td>
<td>3D</td>
<td>1D</td>
<td>1D</td>
</tr>
</tbody>
</table>

*Note.* Study focus: *I* = instrument development, *T* = theoretical research; *Sample* indicates sample sizes in order of multiple studies; *Items* indicates number of items Software: *CQ* = ConQuest, *WS* = WinSteps, *FS* = FACETS, *WM* = Winmira; *Theory* = dimensionality of theorized models; *Model* = dimensionality of estimated psycometric model(s); *maxfit* = dimensionality of model exhibiting best fit (na indicates not applicable because only one model was estimated); *relmod* = dimensionality of model for reliability estimation; *ifitmod* = dimensionality of model for itemfit estimations.
In psychometric modeling, theoretical components are modeled as dimensions. To varying degree, in most reviewed studies the number of psychometric dimensions did only partially or not at all overlap with the number of theoretical components (Table 2.1). Across all studies, the number of psychometric dimensions ranged from one to four dimensions. The number of psychometric dimensions was generally equal or lower than that of conceptualized components. In three cases (Esswein, 2010; Lou et al., 2015; Mayer et al., 2014) the number of modeled dimensions did not at all overlap with the number of theoretically conceptualized components of scientific reasoning. Investigations of reliability and fit were in all but two cases undertaken based on a unidimensional model, also in cases when initially additional models with other numbers of dimensions were estimated. In none of the studies a thorough discussion was undertaken regarding model choice for examining reliability and fit. Overall, researchers thus tended to fit a lower number of psychometric dimensions than they had conceptualized as components of scientific reasoning.

2.3.3 Fit investigation practices

Next, fit investigation practices are considered in more detail. In Table 2.2, practices in fit analysis are summarized. Notably, in all studies item infit statistics were investigated. In most of the reviewed studies, rather strict cutoff values of 0.8-1.2 were considered. In four studies, cutoff values were not explicitly mentioned, and in one study Esswein (2010) itemfit between 0.7 and 1.7 was considered appropriate without providing a reference or rationale for this judgment. To summarize, in most studies rather strict criteria were applied regarding item infit, but not in all studies references or rationales were provided, and in some studies no explicit discussion or mentioning of criteria was present.

Regarding the handling of misfit, in seven studies misfitting items were removed from the instrument based on inappropriate item infit statistics (Table 2.2). This is a common procedure in Rasch modeling. None of the studies followed the common demand that removed items be replaced with revised ones (Nielsen & Kreiner, 2013).
Table 2.2
Statistical fit investigation practices in reviewed studies.

<table>
<thead>
<tr>
<th>Reference</th>
<th>infit</th>
<th>crit (ref)</th>
<th>irem</th>
<th>iadd</th>
<th>outfit</th>
<th>DIF</th>
<th>lrt</th>
<th>misc fit</th>
<th>reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown et al. (2010)</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>PSR</td>
</tr>
<tr>
<td>Esswein</td>
<td>x</td>
<td>none (but 0.7-1.7 deemed good)</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td>RMSE</td>
<td>PSR, ISR</td>
</tr>
<tr>
<td>Grube (2010)</td>
<td>x</td>
<td>0.8-1.2 (Adams, 2000) + t</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>EAP/PV</td>
</tr>
<tr>
<td>Hartmann et al. (2015)</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>EAP/PV</td>
</tr>
<tr>
<td>Heene (2007)</td>
<td>x</td>
<td>0.8-1.2 (Wright, 2000) + t</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>RMSE</td>
<td>PSR, ISR</td>
</tr>
<tr>
<td>Koerber et al. (2014)</td>
<td>x</td>
<td>0.85-1.15 (-)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>EAP/PV</td>
</tr>
<tr>
<td>Kremer et al (2014)</td>
<td>x</td>
<td>0.8-1.2, t-2-+2 (Wu et al., 1997)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>EAP/PV</td>
</tr>
<tr>
<td>Kuo et al (2015)</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PSR</td>
</tr>
<tr>
<td>Lou et al (2015)</td>
<td>x</td>
<td>&gt;1.5 (-)</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>locdep, resPCA,RMSE</td>
<td>PSR</td>
<td></td>
</tr>
<tr>
<td>Mayer et al. (2014)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>EAP/PV</td>
</tr>
<tr>
<td>Nowak et al. (2013)</td>
<td>x</td>
<td>0.8-1.2 (Adams, 2002)</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>EAP/PV</td>
</tr>
</tbody>
</table>

Note. infit = item or person infit statistics; crit = applied itemfit cutoff criteria (+ reference); irem = items removed based on fit analysis; iadd = replacement items added after item removal; outfit = item or person outfit statistics; DIF = differential item functioning analysis; lrt = likelihood ratio test dimensionality comparison; MR = ; locdep = local dependence analysis; resPCA = residual principal component analysis; RMSE = root mean square error fit index; EAP/PV = expected a posteriori/plausible values, PSR = person separation reliability, ISR = item separation reliability.
As evident from Table 2.2, in some studies further fit statistics were used. These included outfit, which is similar to infit but more sensitive towards outlier patterns (Christensen & Kreiner, 2013), and in some studies differential item functioning (DIF) analysis were applied to examine whether item difficulty parameters are comparable between subgroups, which is an important characteristic of invariant measurement (Hambleton et al., 1991; Siersma & Eusebi, 2013). Furthermore, in one study (Lou et al., 2015), residual correlations between items (local dependence analysis) and a principal component analysis were applied to examine the strength of systematic residual variance in the data. To summarize, fit investigation encompassed infit and outfit statistics for items and persons, DIF analysis, residual and dimensionality analysis; the common thread, however, is to apply exclusively or mainly item infit statistics.

2.3.4 Interpretations of Statistical and Psychological Structure

In interpreting researchers’ inferences based on psychometric modeling, there are two kinds of inferences from psychometric modeling that should be distinguished (Bonifay, Lane, & Reise, 2016; Renkl, 2012). On the one hand, inferences can be made about the statistical structure of scientific reasoning. Statistical structure can inform about whether a common thread in statistical information is so strong that single scores should be model and reported, or whether there is additional reliable information such that reporting of multiple scores might be preferred (Reise, 2012; Reise et al., 2013). Information about statistical structure has applied significance for researchers and practitioners such as teachers, because it indicates which psychometric dimensions might be reliably modeled or deal for diagnostics in school classes and for individual students. On the other hand, inferences can be made about the psychological structure of scientific reasoning. This includes inferences about the nature, number, and interrelations of cognitive underpinnings for scientific reasoning, which might describe traits, processes, or cognitive demands. Information about psychological structure has different theoretical significance, because it is related to assumptions about the existence of psychological entities, and their causal relation to answers on items. Psychological structure thus refers to cognitive traits and processes.

In the reviewed literature, inferences were made that regard both the statistical, and the psychological structure of scientific reasoning to varying extent. For example, Nowak et al. (2013) concluded based on model comparison tests that three dimensions describing different
inquiry methods were present in their data. However, in this case it is not explicitly discussed whether the three dimensions were just of statistical nature, or indicated psychological traits; this inference could thus be interpreted as referring both to statistical, or to psychological structure. Lou et al. (2015) compared a theoretically based six-dimensional model to a unidimensional model and interpreted a lack of evidence for statistical multidimensionality as indicating one dominant underlying trait. Interpreting this single overarching trait, Lou et al. (2015) suggested that researchers and teachers use a single score from the instrument and interpret subscale scores with caution. Kuo et al. (2015) found that intercorrelations between four psychometric dimensions were higher than the individual dimensions’ reliabilities. They interpreted the high intercorrelations between the four dimensions as indicating imprecise measurement of the four dimensions and the results not being in accordance with a multidimensional construct. Grube (2010) based on fit comparison of a unidimensional and a four-dimensional Rasch model concluded that the four components can be empirically distinguished. However, in a subsequent modeling step, Grube (2010) added another single dimension on an additional level above the four dimensions and concluded that the four dimensions belong to an overarching construct. Based on moderate intercorrelations between the four dimensions, it was also concluded that the four components share a significant overlap in demands. Körber et al. (2014) fitted a unidimensional model to an instrument meant to assess five dimensions, and based on item infit statistics inferred that the construct itself represents a unitary, conceptually coherent psychological construct. Particularly the latter of these discussed inferences seem clearly referred to the psychological structure of scientific reasoning, and thus influencing psychological theory development in this field. These and likely inferences have implications for theory development about the psychological structure of scientific reasoning, because they regard cognitive processes, or the existence of a specific number of stable psychological traits and their interrelations.

2.3.5 Insights from the Literature Review

Three general insights are drawn from this review of psychometric modeling practices in research on scientific reasoning. First, Rasch modeling, applied in the sense of the Rasch school, is the prevalent practice in research on scientific reasoning. In most of the studies, absolute model fit was not in focus. Rather, the aim in most studies was to ensure items were approximately in
accordance with the Rasch model such that sufficient measurement precision in the sense of the Rasch model would be achieved. Second, there was a high amount of inconsistency in psychometric modeling practices both within individual studies, and between different studies: Cutoffs for fit statistics were not always based on references or explicit rationales, the dimensionality of psychometric models was in few cases based on rationales or discussed in sufficient detail such that inconsistency with theoretical models could be explained (Körber et al., 2014; Mayer et al., 2014), and the criteria for evidence being interpreted as supporting unidimensional or multidimensional models differed strongly across studies. Third, in various of the reviewed studies researchers drew statistical and also theoretical conclusions about the dimensionality and structure of scientific reasoning based on results from the Rasch model. Based on item infit statistics, model comparison tests, and intercorrelations between dimensions, researchers drew inferences about the structure (Nowak et al., 2013), coherence (Körber et al., 2014), and cognitive demands (Grube, 2010) of scientific reasoning. These three findings are now discussed in turn, focusing on how these practices might relate to theory development in research on scientific reasoning, and providing suggestions for altered and extended future practices.

2.4 Discussion

2.4.1 The Rasch School as Prevalent Approach to Psychometric Modeling

The first finding from the review is that Rasch modeling, in the sense of the Rasch school, is the prevalent approach to psychometric modeling. Does the Rasch model deserve this special status in research on scientific reasoning, and is its prevalence a surprise? Particularly for researchers from central Europe, but also from diverse OECD- and further countries, it might not come as a surprise that Rasch modeling is the prevalent approach to psychometric modeling in research on scientific reasoning. This trend is likely driven by application of the Rasch model in highly visible large-scale assessment programs, the five of which that were described here taking place in Germany (IQB Federal State Comparison; Pant et al., 2013), the USA (NAP, NAEP; Donovan, Hutton, Lennon, O’Connor, & Morrissey, 2008; Donovan, Lennon, O’connor, & Morrissey, 2008; National Assessment Governing Board, 2007), the OECD countries (PISA; OECD, 2006), and further countries around the world (TIMSS; I. V. S. Mullis, Martin, Ruddock, O’Sullivan, & Preuschoff, 2009; I. V. Mullis et al., 2003). The trend towards Rasch modeling also
Psychometric Issues

go hand in hand with the trend to conceptualize constructs in educational measurement as
competencies. A competence is conceptualized as a complex ability that is closely related to
performance in real life situations (Hartig, Klieme, & Leutner, 2008). Constructs under this label
are often conceptualized as very broad and as encompassing multiple skills, abilities, knowledge
and likely traits. Also in some of the reviewed studies, scientific reasoning was sometimes
conceptualized, or described, as a competence (Grube, 2010; Hartmann et al., 2015; Körber et al.,
2014; Mayer et al., 2014).

The foundation of this frequent use of Rasch modeling is its strong grounding in the
underlying philosophy of invariant measurement (Andrich, 2004; Boone et al., 2014;
Engelhard Jr, 2013). Researchers in scientific reasoning and beyond often stick to this approach
because Rasch modeling is displayed as the only means to achieve valid measurement in
educational and psychological measurement literature (Boone et al., 2014). Among theoretical
statisticians and philosophers of measurement, however, the theoretical basis and statistical
uniqueness of the Rasch model are strongly debated. These debates might go unnoticed by
applied researchers, who likely rely on textbooks and tutorial literature in which space for
theoretical discussions is limited. It is therefore indicated to provide a concise overview of some
issues concerning the Rasch model and its stance in the world of psychometric modeling.

Let us approach the question of the special status of the Rasch model from the perspective
of a teacher who would like to use a Rasch calibrated measure to track the level and development
of scientific reasoning in her school class, and to diagnose and meet the educational needs of
individual students. For this and similar situations, it is relevant that the Rasch model is first and
foremost a model of individual differences. For a Rasch calibrated measure, invariant
measurement does hold for the measurement of interindividual differences, but only to a limited
degree this does inform about measurement on the level of the individual. It has been shown in
simulations, formally proven, and empirically corroborated that a structural model of individual
differences does not have to hold on the level of individuals (P. C. M. Molenaar, 2004). For
example, in personality research the prevalent big five model is continuously found to describe
quite well the structure of interindividual differences, but when measures are taken repeatedly on
the same individuals, resulting personality structures are found to differ strongly between persons
(Borkenau & Ostendorf, 1998). It is yet to examine to which extent fit statistics in the Rasch
model ensure that a structure describing individual differences holds within individuals (P. C. Molenaar & Campbell, 2009) (for a related suggestion see Stenner, Fisher, Stone, & Burdick, 2013).

As a related statistical issue, Rasch models are often specified with a random effect for person ability parameters; under this specification, the model is conceptualized for populations of persons with different abilities, but not for individuals (Robitzsch, 2016). This implies that under the common specification Rasch models do not even demand and depict that the model holds for individuals, but this would be relevant for the described situation and from broader policy perspectives in education and beyond. Another related statistical question is whether the currently most common approach to estimating parameters in the Rasch model, marginal maximum likelihood estimation, is consistent with the property of specific objectivity. It has been debated that this property is only given under the alternative conditional maximum likelihood estimation, in which person raw scores instead of parametric assumptions are used to inform item parameter estimation (Mair & Hatzinger, 2007, see Robitzsch, 2016 for a diverging perspective). In the reviewed studies, conditional maximum likelihood was not applied although it is available in a well-developed free software package (Mair, Hatzinger, & Maier, 2010). Depending on which opposing view on this issue one follows, the property of specific objectivity of the Rasch model may thus not be given. It has furthermore been argued that specific objectivity (Robitzsch, 2016) as well as sufficient statistics (Humphry, 2011) may also be obtained with other psychometric models than the Rasch model. These and further debates are likely to continue, and for researchers in scientific reasoning this discussion should make visible that the uniqueness of the Rasch model is debated among statisticians.

Finally, the attractiveness of the Rasch model is partially grounded in its likening with definitions of measurement in the natural sciences demanding that a measurement instrument be independent of what it measures (Thurstone, 1928). Here, it should be noted that manifold theories and definitions of measurement exist that vary in breadth and emphasis on different aspects (Finkelstein, 2003; Mari, Maul, Irribarra, & Wilson, 2016). The Rasch model is discussed as a stochastic realization of conjoint measurement for quantitative structures within representational measurement theory (Perline, Wright, & Wainer, 1979). It is however widely debated whether and in which way Rasch modeling is related to such measurement (Kyngdon,
This debate is linked to the further question whether structures of psychological and educational constructs are at all quantitative, a property which many measurement theorists deem a pivotal prerequisite for measurement. Whether a structure is quantitative can be tested, but testing this property is difficult and has never been done for constructs in Education and Psychology, questioning whether actual measurement in these disciplines has ever taken place (Michell, 2000, but for a recent counterargument see Mari et al., 2017). Thus, what measurement really is, and whether and how it can be achieved in Education and Psychology are highly complex questions that cannot simply be answered with the Rasch model, as researchers might suspect when looking into Rasch school proponents’ textbooks. The attractiveness of the Rasch model for Education and Psychology might however be justified from a more practical perspective. For example, the independence of characteristics of persons and items is a property that has the intuitive appeal of fairness which is important for policy making and decision communication. This characteristic might still be best justified by sticking to the Rasch model. Thus, this overview is not aimed at discrediting Rasch modeling as a useful tool, nor its special status in educational and psychological measurement that is has maintained and cultivated over the last couple of decades. However, researchers should be aware that the Rasch model is neither a guaranteed recipe, nor with certainty the only recipe, for yielding informative insights. Whether the Rasch model or other types of models should be applied eventually should be based on detailed considerations of various factors and aims (Robitzsch, 2016). Following the Rasch model in accordance with large-scale assessment programs and practices in competence assessment might be the most easily justifiable approach. Sticking to this approach across a variety of educational and psychological questions might however limit its informativeness. Particularly, informativeness depends also on the question what exactly is scaled in a measurement model, that is, whether a psychometric dimension depicts a theoretically meaningful construct. Whatever is measured with the Rasch model does not yet have to be a useful construct, an issue that is taken up next.

Various alternatives exist for researchers who would like to go beyond the Rasch school and apply models from the broader IRT school. Typical models that were developed concurrently with the Rasch model and its extensions are the 2PL, 3PL, and 4PL models, in which differing item discriminations, guessing probabilities, or slipping probabilities across items are modeled.
(Revelle, 2004). Also nonparametric alternatives are available, most notably Mokken scaling (Mokken, 1971), and categorical IRT in which participants can be practically allocated to a finite number of ability levels (Bartolucci, Bacci, & Gnaldi, 2014). Another approach to psychometric modeling particularly for practical purposes is cognitive diagnosis modeling, in which for example based on students’ answer patterns it could be estimated which scientific reasoning skills students already possess, and which not (de la Torre, 2009). All of these approaches are available in the free R software, for example in the TAM (Kiefer et al., 2016) & irtProb (Raiche & Raiche, 2009), mokken (Van der Ark et al., 2007), multiLCIRT (Bartolucci et al., 2014), and CDM (Robitzsch, Kiefer, George, & Uenlue, 2014) packages. Conditional maximum likelihood estimation is available in the eRm package (Mair & Hatzinger, 2007). A commented syntax demonstrating the estimation of these models in the R software is provided in the online supplementary materials of this article. These models differ in statistical assumptions, demands towards assessments instruments and sample sizes, and scope. Applying a broader range of these models and comparing the gained information might help researchers yielding new ideas and insights, and varied perspectives on the information they seek about scientific reasoning.

2.4.2 Five steps Towards Self-Fulfilling Inferences

The second and third insights from the literature review are discussed concurrently. These concern inconsistencies in modeling and fit investigation practices, and how these are related to researchers’ statistical and psychological inferences. In order to discuss the interplay of these two issues, a prototypic scheme is presented in Figure 2.4. In this scheme, five practices are presented which have occurred in the reviewed literature and demonstrate the interrelated emergence of the two issues.

First, items are typically developed under a multidimensional theoretical model of scientific reasoning (Figure 2.4(1.) & (2.)). Then, fit of data to the psychometric model is frequently evaluated based on item infit statistics in a unidimensional model (Figure 2.4(3.)). If some items show misfit according to varying fit criteria, they are removed until the remaining items fit the unidimensional model (Figure 2.4(4.)). The result is that all (remaining) items are interpreted to fit the unidimensional model. It is concluded that the unidimensional model shows adequate fit (note the mixing up of data fit and model fit occurring in this step), and that scientific reasoning is
in accordance with this unidimensional structure (Figure 2.4).

![Figure 2.4. Five steps to circular inference.](image)

This scheme makes apparent that the predominant practices identified in the literature make it difficult for researchers to find evidence that refutes the unidimensional Rasch model. Although most researchers conceptualized multiple components of scientific reasoning, in the large part of reviewed articles a unidimensional model was fitted. By mostly applying itemfit statistics, such a statistical model cannot be rejected, because itemfit statistics do not test the Rasch model. In addition, cut-offs for itemfit statistics differed across studies, and rationales for a specific choice of cut-offs were seldom provided. A lenient choice of cut-offs for itemfit statistics further supports the practice that eventually adequate fit of the unidimensional model is inferred and interpreted.

A crucial issue in this line of reasoning is that itemfit statistics do not test the Rasch model, but whether items fit the model. This might appear to be a subtle difference, but it is important and fosters the described inferences. In the described scheme, ignoring this difference finally leads to a circular inference: A model is fitted but not tested, and subsequently it is concluded that the data revealed the correctness of the model, although its correctness has been assumed a priori.

In order to demonstrate the problematic nature of this line of reasoning, data simulations were conducted. The simulations were aimed at demonstrating how sensitive data-model fit
statistics are in contexts described here\textsuperscript{7}. For these simulations, a plausible multidimensional model of the structure of scientific reasoning was set up. The model that data were simulated from is depicted in Figure 2.5(A). Four interrelated dimensions with intercorrelations varying between $r = .3$ and $r = .7$ were assumed. This scenario was assumed to represent a realistic expectation of the structure of scientific reasoning. It is for example similar to the study by Körber et al. (2014), in which 20 items were developed based on a theoretical model assuming four interrelated components of scientific reasoning.

![Figure 2.5](image)

\textit{Figure 2.5.} Model for data simulations (A) and schematic depiction of common variance across items (B). Depictions for item size of $i = 20$.

\textsuperscript{7}It should be noted that other researchers have evaluated and discussed item infit statistics (Christensen & Kreiner, 2013; Heene, Bollmann, & Buhner, 2014; A. B. Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008; R. M. Smith, Schumacker, & Bush, 1998; R. M. Smith & Suh, 2003) and the present simulations have a demonstration purpose, rather than the purpose to exhibit generalizable statistical insights. The simulation conditions were therefore tailored towards reflecting the main characteristics of the reviewed studies.
The aim of the data simulations was to examine whether and to which extent item infit statistics would be sensitive towards this scenario of multidimensionality. To examine this question, data were sampled from the depicted model, varying sample sizes between \( n = 100, 300, 500 \) and 1000 to resemble the range of most sample sizes from the reviewed literature. The number of items was also varied between \( i = 12, 24, 60, \) and 120, in order to simulate a range in item numbers similar to that in the reviewed studies. 1000 datasets were sampled from this four-dimensional model of scientific reasoning, for each of the combinations of sample sizes and item numbers (with the items always spread evenly among dimensions), yielding an overall number of 16000 datasets. All of the datasets were random draws from a four-dimensional Rasch model structure simulated in the eRm software package (Mair & Hatzinger, 2007); the outlined four-dimensional Rasch model therefore was the real model in all datasets. For each of these simulated datasets, a unidimensional Rasch model was however fitted, and the percentage of items tagged by infit statistics was recorded for each simulated and fitted dataset. Specifically, the number of items exceeding the item infit cut-off range of 0.8 - 1.2 was recorded. This cut-off was the most frequently chosen in the reviewed literature. These simulations should indicate whether the item infit statistics would react to the modeling of one psychometric dimension on data that originate from multiple dimensions.

![Figure 2.6](image)

**Figure 2.6.** Sensitiveness of item infit statistics towards multidimensionality under varying sample and item size.
The result of the percentage of items that item infit statistics tagged to be misfitting is depicted in Figure 2.6. As can be seen, for each combination of sample size and item number, the item infit statistics did not react to multidimensionality. The maximum percentage of items that item infit statistics reacted to on average across 1000 datasets was about 1.7%, under the condition of $i = 12$ items and a sample of $n = 150$ participants. With increasing sample size and number of items, this percentage further decreased. These simulations show that item infit statistics are not sensitive towards detecting multidimensionality.

These quick simulations showed that itemfit statistics are not sensitive towards multidimensionality. Inferences based on the assumption that itemfit statistics in a unidimensional model point towards a coherent trait, fit of the unidimensional model, or a common set of underlying cognitive processes are thus likely to be too far-fetched.

How can it be explained that item infit statistics, the by far most common approach to fit testing in the Rasch school, do not react to this clear pattern of multidimensionality? The explanation is what proponents of the Rasch school do often explain but can be difficult to grasp. Item infit statistics do not test the assumptions of the Rasch model. They test whether a common Rasch dimension can be extracted from data, for which multidimensionality in many cases is not a problem. It is thus researchers’ focus that determines how many dimensions will be extracted from data. If researchers want to extract one dimension, item infit statistics are aimed to show whether this one dimension is estimated with sufficient precision such that it represents the common variance across items. If researchers want to extract four dimensions, item infit statistics do the same, for each of the four dimensions individually. These statistics are thus not aimed at evaluating whether a model shows absolute fit, but whether items fit a pre-determined model sufficiently well. This conceptual explanation holds for other types of item- and person-fit statistics. In accordance with the Rasch school, it is not the task or scope of person and item fit statistics to find, prove, or disprove the right or wrong model. The scope of these and related data-model fit statistics is to indicate whether from a set of items a common scale can be extracted with sufficient precision such that person parameters will adequately represent the common

---

The number of items was simulated to be equal across dimensions; under deviations from this scenario itemfit statistics likely tag more items. However, an equal or almost equal number of items across dimensions is the regular scenario in psychometric studies.
variance across all items.

If itemfit statistics support the modeling of a general dimension across four intercorrelated factors, what exactly is it that is represented in this common dimension, across items that are meant to assess the complexity of scientific reasoning? To illustrate this question, in Figure 2.5(B) the process of extracting common variance from items meant to assess four dimensions of scientific reasoning is depicted. In this hypothetical example, the four dimensions represent items concerned with epistemic cognition (a dimension often associated with scientific reasoning that encompasses for example beliefs about knowledge and its acquisition and development, Schommer, Calvert, Gariglietti, & Bajaj, 1997) and with the three facets of the SDDS model. The common variance across all items is depicted in dark gray, which catches part of the variance of all the items. There are various possibilities what such common variance, scaled in a single Rasch model dimension, might represent.

A first possibility what common variance might represent is method variance. Systematic variance is often caused by method, for example when tasks of a measure have similar wording, are presented in pen-and-paper form, or use similar answer scales such as multiple choice or constructed response (Pohl & Steyer, 2010). While it might be doubted that common method variance accounts for a full stably estimated Rasch dimension, it could at least be responsible for part of the common variance across items that measure different traits.

A possibility that is also related to common demands across items is that general reasoning plays a major role in all items. This possibility is corroborated by evidence from the reviewed literature. Mayer et al. (2014), for example, found that a model assuming a general reasoning dimension and scientific reasoning dimension to overlap perfectly showed better fit than a model assuming these two constructs to be separable. Still, based on a non-perfect correlation between the two constructs in a two-dimensional model, the authors concluded that the two constructs are distinct. It seems however, from this and further studies (Körber et al., 2014; Nowak et al., 2013), that scientific reasoning and general reasoning are highly correlated, raising the question how much of a scientific reasoning-dimension is indeed specifically scientific.

Another related possibility might be further common cognitive demands, which however do not have to be specific to scientific reasoning. For example, most measures in the reviewed literature are based on items that demand reading comprehension, and also often some spatial
ability. Some studies have examined this possibility and these constructs seem to explain substantial amounts but not all of the variance in scientific reasoning dimensions (Körber et al., 2014; Mayer et al., 2014; Nowak et al., 2013).

The perhaps most conspicuous source of common variance across scientific reasoning items is mutual developmental influence. Scientific reasoning does often take place in an iterative circle of inquiry (Manlove, Lazonder, & Jong, 2006), which implies that the activities implied in this circle do depend on, and are likely to influence each other. For example, learning to design an experiment will offer new opportunities to interpret evidence, and more advanced interpretations of evidence might contribute to better understanding of theory development. In this way, the components of the SDDS model and further facets of scientific reasoning are developmentally intertwined, because they constrain and support each other developmentally. This intuitive assumption implies that a common dimension of scientific reasoning reflects common variance stemming from developmental interrelations between components of scientific reasoning. Such mutual influence between cognitive abilities has been assumed to constitute general dimensions in other areas, such as general reasoning (Van Der Maas et al., 2006).

2.4.3 Suggestions for Future Practices

This discussion of inconsistencies in psychometric modeling, fit investigation practices, and the resulting inferences, indicates that the currently predominant psychometric practices in research on scientific reasoning might foster circular inferences. Psychometric models are set up that are tested with itemfit statistics, which do not test the Rasch model, but inferences are drawn regarding the psychological structure of scientific reasoning. Future avenues are now discussed how these current practices might be expanded in order to support further theory development about scientific reasoning.

First, it has been discussed that itemfit of a unidimensional Rasch model indicates common variance across items. It is however not known what the common variance depicted in a Rasch dimension stems from. In order to examine the source of the common Rasch dimension, for example longitudinal studies could be conducted, to examine mutual influence of different scientific reasoning abilities. In this way, it could be examined whether and how different components of scientific reasoning influence each other or relate to each other’s development over
time. This might reveal constraints and dependencies in developmental patterns, which can represent a source of a common Rasch dimension. Furthermore, in studies varying assessment method, it could be examined to which extent the common Rasch dimension can be explained by method effects, which could in turn reveal the influence of common cognitive demands. For example, by varying the amount of pictorial material in items, and of multiple choice, constructed response, and hands-on tasks, it could be examined to which extent common variance is triggered by common cognitive demands. Another suggestion is to conduct experimental studies: If one component of scientific reasoning is specifically trained, and this training over time propagates to another component, it may be assumed that mutual influence between the two components exists. These suggested longitudinal and experimental designs, and designs varying stimulus and answer material, might be most fruitfully deployed during or shortly before strong developmental phases in scientific reasoning. In the course of such developmental phases, dynamics between different components and cognitive demands of scientific reasoning could then be monitored.

A further question of educational interest which is linked to experimental designs is whether what is represented in the Rasch model dimension of scientific reasoning can be trained by educational interventions. Even if this dimension overlaps rather strongly with general reasoning and other cognitive abilities, it might be hoped that what is specifically scientific to this dimension might be better trainable than stable general cognitive abilities (Melby-Lervåg & Hulme, 2013). So far, there have been no studies aimed at training this general dimension of scientific reasoning.

In the reviewed literature, scientific reasoning has however been correlated with educational variables, such as school grade and year and field of studies (Esswein, 2010; Grube, 2010; Hartmann et al., 2015; Körber et al., 2014; Kremer, Specht, Urhahne, & Mayer, 2014; Mayer et al., 2014). The danger of looking into such relations based on a unidimensional Rasch model, however, is that different components of scientific reasoning can have differential relations to these and further external variables. In a unidimensional model, such differential relations can lead to an overall biased average estimate, and the more accurate, differential information that would be gained from a multidimensional model is lost. An obvious way to circumventing this issue is to fit a multidimensional model in addition to a unidimensional model. In this way, it can be examined whether modeling multiple components of scientific reasoning reveals differential relations that would be hidden in a multidimensional model.
Particularly in analysis of scientific reasoning across age groups or across educational levels, analysis of differential item functioning should as well be applied. This type of analysis allows examining structural changes in scientific reasoning. Often, such structural changes are expected during development; such information is however not modeled and thus not visible if the same Rasch model is estimated across age or educational level.

Two further measures are suggested to alleviate the current practices in fit examination and dimensionality assessment. One measure is the careful inspection of a Wright map in which it is tagged which theoretical component items belong to (Wilson, Allen, & Li, 2006). On a Wright map, persons and items are depicted on the same graphical dimension, and it might support detecting effects of interventions or other external variables which impact only one component of scientific reasoning. It can however be doubted whether in a unidimensional model the true effects can be made visible with a Wright map, because the common Rasch dimension across items does catch the limited variance that is common to multiple components. Variance that is specific to multiple components might not be part of a common Rasch dimension, which would undermine seeing differential impact in a Wright map. A second measure is to use tools for recognizing the presence, strength and impact of multidimensionality and further violations of model assumptions, which are indeed easily available. There are various free, powerful statistical tools for detecting violations of assumptions of the Rasch model. Approaches suggested in the Rasch school include conducting a principal component analysis on Rasch model residuals, and examining local dependencies between items (Linacre & Wright, 2000). Further tools are rarely mentioned in literature by proponents of the Rasch school, because they focus on fit of the model rather than data, which the Rasch school is generally not interested in. These tools include tests for person and item homogeneity (Andersen, 1973; Kreiner & Christensen, 2013), detecting heterogeneity arising through combinations of observed covariates (Strobl, Kopf, & Zeileis, 2015) or through unmeasured covariates (Rost, Carstensen, & Von Davier, 1997), the M2-statistic and alternatives for examining goodness of fit (Maydeu-Olivares, 2013), nonparametric fit statistics that preserve high power for detecting violations in small samples (Koller et al., 2015), and under Bayesian estimation it can be examined whether data generated by the model match empirical data (Fox, 2010). Particularly accommodating for researchers who are acquainted with structural equation modeling is that psychometric models can be fit as structural equation models where
various well-known fit indices such the Root mean square error of approximation (RMSEA), comparative fit index (CFI), and square root mean residual (SRMR) are available (Glockner-Rist & Hoijting, 2003). Recently, these typical fit indices also became available directly within the framework of psychometric modeling (Maydeu-Olivares & Joe, 2014).

All mentioned statistics and analysis are available in powerful, user-friendly and free software packages (Kiefer et al., 2016; Mair & Hatzinger, 2007; Rosseel et al., 2017). If multidimensionality should be acknowledged and modeled, these packages also offer the opportunity to estimate multidimensional (Rasch and further IRT) models, in which information about the various components of scientific reasoning can be made visible (Reckase, 2009).

As a general measure to improve the report of psychometric modeling, it is suggested that researchers describe, in sufficient detail such that readers can reconstruct decision processes, their rationales for (a) choice of model (e.g., the Rasch model) and model structure (e.g., dimensionality of fitted models), (b) fit investigation practices (e.g., fit indices, model comparison tests) and how and why these should be related to hypotheses and exploratory questions (e.g., hypothesis concerning psychometric structure of test, exploration of data-model fit), and (c) choice of cut-offs and decision rules for fit investigation and further psychometric tests (e.g., cut-offs for item & person infit & outfit indices, decision rules about group difference tests). Such descriptions might ensure that psychometric modeling practices become more transparent to readers and reviewers, and also that researchers ensure the validity of inferences. It is also important that researchers discuss the theoretical reasons for their model and fit testing choices because the models themselves are devoid of any information; the information arises from underlying theory and the combination with prior evidence (Sijtsma, 2012). In cases such as the reviewed literature, this is relevant for example when researchers conduct a comparison test and find that a unidimensional Rasch model shows better fit according to indices such as the AIC or BIC than a multidimensional model. From a statistical perspective, fit indices such as AIC and BIC identify models with an optimal balance of model fit and number of parameters (Bozdogan, 1987). Such a finding does however not allow strong psychological inferences unless it is based on strong theory and additional data. Looking in detail at this problem, it becomes obvious that psychometric data are usually cross-sectional. Theories about psychological constructs however involve causality and the existence of distinct processes. It is commonly acknowledged that
cross-sectional, non-experimental data generally do not allow conclusions about causality and processes. Limitations of cross-sectional data tend to be overlooked in disguise of psychometric models. Whether a parameter representing a common Rasch dimension is meaningful from a psychological perspective can therefore not be decided based on model comparisons tests applied to regular cross-sectional data, because such data do not adequately represent parameters that should stand for causality and cognitive processes. Focal parameters in model comparison tests, such as those representing a common Rasch dimension, therefore are devoid of substantial meaning and are arguably downgraded to data aggregation tools. Therefore, if researchers aim at testing psychological hypotheses about scientific reasoning, it is crucial that they report in detail the theory underlying their psychometric models. This will enable reconstructing why a certain set of data should be suitable for testing strong theoretical assumptions by means of psychometric modeling.

For the reassurance of reliable hypothesis testing under the application of psychometric modeling, a further measure is proposed. Specifically, it is suggested that researchers report their hypothesis testing procedures before they conduct the respective tests. The reason for this proposal is that without strict procedures and decision rules that are specified before hypothesis tests are conducted, hypothesis tests become unreliable and devoid of information. The circumstances under which statistical hypothesis tests function well are complex and very difficult to meet (Wagenmakers, 2007). Generally, in order for common statistical hypothesis tests to work properly, all tests that will be used for inferences, and the sample these tests will be applied on have to be fixed a priori. In the reviewed literature, these conditions for hypothesis tests to work have not been met. Psychometric models were often adapted after item exclusions and fit comparison tests, which however undermines the functioning of hypothesis tests; \( p \)-values and likely procedures that should lead to decisions about hypothesis do not work under these circumstances. When researchers adapt their models and tests to data, the probability that a hypothesis test leads to wrong inferences approaches one (Simmons, Nelson, & Simonsohn, 2011). It is therefore suggested that researchers determine in detail beforehand which analysis they will conduct to test hypothesis about scientific reasoning, and openly disclose their analysis plan online (for example on the Open Science Framework; http://osf.io) before they engage in psychometric modeling that should culminate in hypothesis testing (Chambers, Dienes, McIntosh,
Rotshtein, & Willmes, 2015). This measure does not prevent researchers from exploratory analysis. Psychometric modeling is an iterative process and particularly in Rasch modeling, it is emphasized that careful inspection of data in the sense of exploratory analysis is necessary to aid precise measurement (Boone et al., 2014). Pre-defining analysis does however ensure that confirmatory analysis that are used for drawing inferences about hypothesis are clearly separated from such exploratory analysis, such that these two different types of statistics do not get mixed up (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Exploratory analysis can still give rise to hypothesis which have to be tested on new data afterwards. It is hoped that this measure will ensure the reliability of psychometric modeling practices, because reliable hypothesis testing might be particularly difficult to ensure in applied psychometrics.

Finally, the present issues and insights are not limited to the specific field of research on scientific reasoning. Rasch modeling and more general unidimensional psychometric modeling is recently frequently used in many areas, encompassing health science, intelligence and personality research, and psychopathology (Caspi et al., 2014; Gignac, 2016; Musek, 2007; Wilson et al., 2006). The discussed limitations of the Rasch model and further psychometric models applied to cross-sectional data are also valid in those fields, and fruitful discussions might follow of researchers’ practices and inferences in these areas.

**Conclusion**

The present review of psychometric modeling practices in research on scientific reasoning brought forth various informative insights and points for discussion. Rasch modeling, the predominant approach to psychometric modeling in this field, bears strong measurement characteristics, but in the reviewed literature it is related to sometimes oblique and inconsistent modeling and interpretations.

Most importantly, the present study has shown that Rasch modeling is related to some strong inferences that seem unwarranted given the nature of this modeling approach: Fit of items to the Rasch model does yet not imply that a common dimension of scientific reasoning bears useful information about the existence of causal cognitive processes; and in the presence of multidimensionality, which is likely to go undetected by common fit indices, inferences about the development and relations of scientific reasoning that are based on a unidimensional model can be
misleading. It is therefore suggested that in future endeavors researchers acknowledge the multiple abilities, skills, and further facets of scientific reasoning in psychometric modeling.

It was further suggested that researchers reflect and decide beforehand on their research questions, accept the limitations of correlational data, and report the details of their modeling strategy that bears on hypothesis testing before they conduct the relevant research. In this way, the transparency of psychometric modeling strategies will be increased, limitations of data acknowledged, and confirmatory research more clearly separated from exploratory psychometric modeling. With these suggestions, information gain might be increased regarding educational and psychological questions in research on scientific reasoning, both when researchers rely on the Rasch model as psychometric method of choice, or decide to apply further methods beyond the Rasch school of psychometric modeling.

**Supplementary Materials**

Supplemental material including syntaxes and data is stored publicly available at the Open Science Framework under https://osf.io/zmwnv/, and an R package for the data simulations under https://github.com/peter1328/simrasch.

**Acknowledgements**

I am grateful to Fabian Dablander for exchange about the present study, who also wrote large parts of the R package for the presented and further data simulations.

**References**


doi:http://dx.doi.org/10.14786/flr.v2i2.96


Psychometric Issues


Schommer, M., Calvert, C., Gariglietti, G., & Bajaj, A. (1997). The development of
episemological beliefs among secondary students: a longitudinal study. *Journal of
Educational Psychology, 89*, 37–40. doi:10.1037/0022-0663.89.1.37

Siegler, R. S. & Liebert, R. M. (1975). Acquisition of formal scientific reasoning by 10-and
doi:10.1037/h0076579

Siersma, V. & Eusebi, P. (2013). Analysis with repeatedly measured binary item response data by
doi:10.1002/9781118574454.ch14

doi:10.1016/j.measurement.2011.03.019


flexibility in data collection and analysis allows presenting anything as significant.
*Psychological science, 22*, 1359–1366.

and sample size considerations for polytomous data. *BMC Medical Research Methodology,
8*, 33. doi:10.1186/1471-2288-8-33


Smith, R. M. & Suh, K. K. (2003). Rasch fit statistics as a test of the invariance of item parameter


Chapter 3

Elementary School Students’ Verbal and Non-Verbal Knowledge about Experimental Design

Peter A. Edelsbrunner, Anne Deiglmayr
ETH Zürich

Abstract

Verbal knowledge about experimental design is a prerequisite for collaborative engagement in experimentation, and non-verbal knowledge for individual engagement. Thus, these two kinds of knowledge differ in developmental and educational relevance. Students’ verbal explanation knowledge and non-verbal application knowledge about experimental design, and their interrelation were examined throughout elementary school. Over 3000 first- to sixth-graders worked on written tasks with open answers assessing their verbal knowledge about crucial principles of experimental design, and on multiple choice tasks assessing their non-verbal knowledge. Fourth grade turned out to be a significant developmental period for both kinds of knowledge. Codings of rationales underlying students’ open answers indicate that up to fourth grade, their verbal knowledge about experimentation is often based on tautological reasoning, factors irrelevant to the validity of experimental design, and biased by prior content knowledge about the task domains. Latent variable analysis indicate that as well up to fourth grade, students’
Knowledge about Experimental Design

verbal knowledge can be separated from their non-verbal knowledge. In fifth and sixth grade, both types of knowledge are on a high level in most students, and statistically inseparable. These findings replicate and substantiate prior findings, and provide new insights into students’ knowledge and reasoning about experimental design throughout elementary school. 

Keywords: scientific reasoning; item response modeling; Rasch model; literature review; simulation study

3.1 Introduction

Researchers, policy makers, and practitioners emphasize that children should develop knowledge about major principles of experimental design early on because experimentation is a powerful tool for examining and learning about causal relations in school settings, science, and everyday life (National Research Council, 1996, 2013; Sandoval, Sodian, Körber, & Wong, 2014). Accordingly, various investigations have been concerned with the development of children’s knowledge about experimental design (Sodian & Bullock, 2008; Wilkening & Sodian, 2005; Zimmerman, 2000, 2007), and the possibility to support this development with educational interventions (Klahr, Zimmerman, & Jirout, 2011; Sandoval et al., 2014; Schwichow, Croker, Zimmerman, Höffler, & Härtig, 2015).

An often overlooked virtue of knowledge about experimental design is that different kinds of knowledge exist that have different relevance for students’ schooling and everyday lives. Knowledge about experimental design can be verbal; this explanation knowledge is relevant for collaboration and argumentation with peers, teachers, or parents. Knowledge about experimental design can also be non-verbal; this application knowledge is more relevant for engaging in experimentation and interpreting insights on one’s own. The differential development, and interrelations of these two types of knowledge have implications for developmental theory and education: Does the knowledge children need to solve verbal knowledge tasks develop hand in hand with that they need to solve non-verbal knowledge tasks, or are these two types of knowledge separable and need different educational support? The present study examines these questions, focusing on the development and interrelation of students’ verbal and non-verbal knowledge about experimental design throughout elementary school. In addition, verbal knowledge is usually not in the focus of large-scale investigations. In the present study, in a large
sample of first- to sixth graders, students’ reasoning when they exhibit verbal knowledge is examined in detail, to examine the factors contributing to its development.

3.1.1 Knowledge about Experimental Design in Elementary School Students

Two major principles of experimental design are conclusive hypothesis testing (CHT; Sodian, Zaitchik, & Carey, 1991) and the control-of-variables-strategy (CVS; Bullock & Ziegler, 2009). These two principles yield the design underlying a controlled experiment: The focal variable has to be varied such that it tests the hypothesis (CHT), while all other potentially confounding variables have to be kept constant (CVS). There are also other, more advanced principles, but these two principles are basic building blocks for understanding experimentation.

CHT is related to broader understanding of determinacy and indeterminacy, that is, whether available evidence is a sufficient warrant for a specific conclusion (Falmagne, Mawby, & Pea, 1989; Klahr & Chen, 2003; Somerville, Hadkinson, & Greenberg, 1979). In observational evidence, an indeterminate situation is given when the evidence related to a question is not sufficient for drawing a conclusion. For example, when a toy consists of plugged round parts and the question is which of two boxes the parts were taken from, but both boxes contain round parts, the evidence is indeterminate (Piéraut-Le Bonniec, 1980). If however only one box contains round parts, the situation is determinate and thus the observational evidence is sufficient to infer that the toy was built with parts from that box. Thus, determinacy is concerned with ensuring that available evidence eliminates all uncertainty about an outcome (Fay & Klahr, 1996). In experimental design, the CHT principle pertains to the focal variable of which the causal status is in question. In order to produce determinate evidence, the focal variable has to be varied such that it can yield determinate evidence. An example for indeterminate experimental evidence is when an object that smells strongly to humans is hidden to find out whether a German short hair dog can smell better than humans: Even humans could have smelled the object, therefore, if the dog finds it, this is not determinate evidence, and therefore no conclusion can be drawn. In another

1The principles in focus in this study are pivotal, but there are other, more and less advanced steps in the development of knowledge about experimentation, for example generic understanding of causality, and the coordination of multiple variables; these and further knowledge about experimental design are discussed in Croker and Buchanan (2011), Kuhn, Iordanou, Pease, and Wirkala (2008), Osterhaus, Körber, and Sodian (2016), Piekny and Mähler (2013), Sodian et al. (1991)
classic task, Sodian et al. (1991) asked children to decide whether in order to find out if a small mouse or a big mouse is in the house, a bait of cheese should be put in a small box overnight in which only the small mouse fits, or in a big box in which both mice fit. If the next morning the bait is missing from the big box, this leaves an inconclusive state of evidence, because both mice could have taken the bait.

In addition to ensuring that the variation of the focal variable allows for determinate evidence, all other variables that could influence the outcome have to be kept constant between experimental conditions. This principle is referred to as the control-of-variables-strategy (CVS Chen & Klahr, 1999), sometimes also as vary-one-thing-at-a-time (VOTAT; Tschirgi, 1980). For example, in a classic task children are presented different options how an experiment could be designed in order to find out whether the shape of the nose of an airplane influences fuel use (Bullock & Ziegler, 1999). In order to solve the task, they have to understand that only the shape of the nose should be varied but all other characteristics should be kept similar between experimental conditions, such that the design represents a controlled comparison that allows conclusions. CVS, in combination with CHT, yields the basis of any experimental design, which then depending on the context is combined with further principles such as randomization and local control (Osterhaus et al., 2016).

In research, CHT and CVS are not always distinguished; that the focal variable is varied in accordance with the hypothesis is often implied in definitions of CVS (e.g. Chen & Klahr, 1999; Strand-Cary & Klahr, 2008). From a developmental view, it is however likely that CHT is an easier and less complex principle than the full CVS principle: Understanding that a variable has to be varied in accordance with a hypothesis such that it yields determinate evidence is the core of CVS, in which all other variables in addition have to be kept constant. There is evidence supporting the notion that CHT is less complex to comprehend than full CVS. Basic understanding of causality, which is strongly connected to CHT, develops in early infancy (Carey, 1995), and basic knowledge about CHT indeed in some children develops already in early childhood, earlier than knowledge about CVS usually does (Sodian et al., 1991). CHT can thus be seen as a first step in understanding experimentation, followed by more advanced understanding of the relation of indeterminacy to confounding variables which yields the CVS principle.

Developmental and educational researchers emphasize the importance of acquiring
knowledge about principles of experimental design early on. The developmental relevance rests on the importance of knowledge about experimentation for the development of more advanced scientific thinking, and that it is related to the general ability to coordinate theory and evidence, which is at the heart of scientific endeavors (Kuhn, 1989; Kuhn, Ramsey, & Arvidsson, 2015). The educational significance of knowledge about experimental design is additionally grounded in its importance for knowledge acquisition in inquiry contexts (Edelsbrunner, Schalk, Schumacher, & Stern, 2016, in press; Schauble, 1996), and also its importance for general science achievement, which goes beyond that of general cognitive abilities (Bryant, Nunes, Hillier, Gilroy, & Barros, 2015). The development of knowledge about experimental design thus has significant developmental and educational implications.

Regarding the development of CHT and CVS, a partially outdated assumption is that knowledge about principles of experimental design cannot develop before adolescence. This assumption was introduced by Jean Piaget within his definition of formal reasoning, and in observational studies he gathered evidence that knowledge about abstract principles of experimental design does not develop before adolescence (Piaget & Inhelder, 1958). An increasing body of evidence refutes Piaget’s theory and suggests a more complex picture.

Knowledge about principles about experimental design can develop in childhood, and this development can be supported by educational interventions. Precursors of knowledge about CHT and CVS develop in some children already at preschool age or in elementary school age, and a lot of evidence points towards a phase of increased development during age 9 to 12 (Bullock & Ziegler, 1999; Sodian & Bullock, 2008; van der Graaf, Segers, & Verhoeven, 2015; Zimmerman, 2007). Elaborate educational interventions to improve knowledge about experimental design have therefore been developed that encompass the whole range of education from early elementary school (Case, 1974) to university (Lin & Lehman, 1999), in low- and high achievers from diverse socioeconomic backgrounds (Case & Fry, 1973; Lorch Jr et al., 2010; Lorch et al., 2014; Zohar & Peled, 2008). In all of these populations, interventions led to stable knowledge gains in comparison to control groups (Chen & Klahr, 1999, 2008; Ross, 1988; Schwichow et al., 2015). This evidence suggests that Piaget’s assumption was incorrect, because knowledge of principles about experimental design can develop in childhood and can be trained in educational interventions.
Knowledge about Experimental Design

Piaget’s assumption can however not be fully refuted, because not all children develop thorough knowledge about CHT and CVS before adolescence, and even in some university students wrong conceptions about major principles of experimentation remain (Boudreaux, Shaffer, Heron, & McDermott, 2008). Such wrong conceptions that can remain stable throughout adolescence have been investigated through the lens of conceptual change research (Körber, Mayer, Osterhaus, Schwippert, & Sodian, 2014; Körber, Sodian, Kropf, Mayer, & Schwippert, 2011). In conceptual change theory, it is assumed that children have naïve conceptions about the world. Triggered by experiences which cannot be explained by these conceptions, they develop new, more adequate conceptions (Carey, 2009). In this process, children develop a number of intermediate conceptions until they reach a knowledge state of scientific concepts that are in accordance with current scientific views (Schneider & Hardy, 2013). In the context of experimental design, various wrong assumptions about principles of experimental design have been found in children that correspond to naïve conceptions (Siler & Klahr, 2012). A common misconception which undermines producing determinate evidence is children’s assumption that the focal variable should be kept constant at the level deemed best for achieving a subjectively positive outcome (Siler & Klahr, 2012; Zimmerman & Glaser, 2001). Keeping the focal variable constant however cannot produce determinate evidence that allows a conclusion. This misconception is related to the the assumption often found in children that the aim of experimentation is producing a positive outcome, rather than figuring out the causal status of variables (Schauble, Klopfer, & Raghavan, 1991; Zimmerman & Glaser, 2001). Producing a desirable outcome can however rather be accredited to an engineer, and thus the conception underlying this strategy is termed the engineering model of experimentation (Schauble et al., 1991). This naïve conception of experimentation is related to the phenomenon that children are often strongly biased by their own prior assumptions about the investigated domain (Penner & Klahr, 1996). For example, regarding the floating ability of objects in water, many children based on their everyday experiences assume that object weight is the decisive characteristic that determines floating ability (Edelsbrunner et al., 2016; Hardy, Jonen, Möller, & Stern, 2006). Crucially, such prior assumptions are not just often wrong from a scientific view, they can also influence how children set up and interpret experiments (Penner & Klahr, 1996). Another misconception among children is that all potentially relevant variables should be varied, in the
airplane example the shape of the nose, but also the shape of the rudder and number of wings, in order to find out about causal relations (Siler & Klahr, 2012; Tschirgi, 1980). This would however not lead to a decisive conclusion, because any of these factors could contribute to potential outcomes. There is evidence that at the age of about 12, an understanding develops that the aim of experimentation is the testing of effects rather than the prediction of results (Penner & Klahr, 1996). At about the same age, children systematically design experiments in which only the focal variable is varied (Penner & Klahr, 1996; Tschirgi, 1980). Children thus show a variety of naïve conceptions about experimental design, many of which according to prior findings remain throughout childhood.

3.1.2 The Role of Different Assessments Methods

Various assessment methods have been used to examine children’s knowledge about experimental design, and not all led to the same conclusions. These include discovery tasks (Schauble, 1990) and interviews (Penner & Klahr, 1996), analog and digital performance tests (Kuhn, 2007; van der Graaf et al., 2015), and questionnaires consisting of multiple choice items (Osterhaus et al., 2016; Schwichow, Christoph, Boone, & Härtig, 2016) and open answer items (Babai & Levit-Dori, 2009), under lab-based conditions (Klahr & Dunbar, 1988; Schauble, 1990) and under more realistic or school-like circumstances (Penner & Klahr, 1996). While some assessment methods indicated that children at a certain age do not possess knowledge about experimental design, other methods showed divergent findings. A prominent example is children’s ability to differentiate theory from evidence, an ability closely related to the development of knowledge about experimental design. Kuhn (1989) argued that children cannot differentiate between theory and evidence and that this is the key ability in children’s development of knowledge about experimental design and broader scientific thinking. In her seminal research, Kuhn (1989) gathered evidence for this assumption on 7- to 9-year old children. But applying different methods this assumption had to be partially revised: In simple choice-tasks that demand differentiation between theory and evidence, some children seem to draw this distinction and can correctly justify their choices (Piekny & Mähler, 2013; Sodian et al., 1991). Thus, findings and inferences can depend upon the choice from the broad repertoire of methods to assess children’s knowledge about experimental design.
This dependence on method can be partly explained by task characteristics that influence task difficulty. Impact of task characteristics on task difficulty has been examined particularly for questionnaire tasks. For example, the number of variables that children have to take into account in an experimental design task is positively related to its difficulty (Staver, 1984, 1986; van der Graaf et al., 2015). Another significant task characteristic is whether children are asked to design experiments on their own, or whether they are asked to choose from a variety of predefined experimental designs (Schwichow et al., 2016). In early studies children were often given the task to design an experiment, and it was then evaluated whether they correctly applied principles of experimental design. Various studies using this type of tasks have shown that designing an experiment is a very difficult task for children and that only few succeed (Kuhn, 2000; Kuhn et al., 1988; Schauble, 1996). In more recent studies, however, instead of designing experiments on their own, children were asked to evaluate or choose between various predefined experimental designs that are either controlled or not (Piekny, Grube, & Mähler, 2014; Piekny & Mähler, 2013). With this kind of task, much more children succeed by making the right choice, even in early school grades (Bryant et al., 2015; Bullock & Ziegler, 1999; Körber et al., 2011). There are thus various task characteristics that influence which tasks children can solve at which age.

Beyond an influence on task difficulty, task effects can also be related to substantially different knowledge and cognitive processes they demand and trigger (Osterhaus et al., 2016; Schwichow et al., 2015). Most notably, children’s achievement on open answer questionnaire tasks differs from that on multiple choice tasks (Schwichow et al., 2015). One reason for this discrepancy is that on multiple choice tasks, children can guess between usually few options, while on open answer tasks guessing is not possible because they have to develop answers on their own (Staver, 1984, 1986). Open answer tasks are therefore generally more difficult to answer correctly than multiple choice tasks (Schwichow et al., 2015). Another, more intriguing reason for the discrepancy between open answer and multiple choice tasks is that they might substantially differ in the knowledge they evoke and demand (Kuechler & Simkin, 2010; Martinez, 1999). Meta-analytic evidence has shown that open answer tasks show systematically bigger effects of educational intervention aimed at improving children’s knowledge about experimental design than multiple choice tasks do (Schwichow et al., 2015). It has therefore been assumed that open answer tasks demand broader or different knowledge than multiple choice
3.1.3 Different Kinds of Knowledge about Experimental Design

There are various frameworks for categorizing different kinds of knowledge. Knowledge can be implicit or explicit, depending on whether a person is aware of one’s knowledge (Ziori & Dienes, 2008). It can be conceptual and thus facilitate understanding of principles, or procedural and thus facilitate problem solving (Schneider & Stern, 2010). In a similar framework, declarative knowledge is experiential or factual, and procedural knowledge is goal-oriented and serves problem solving (Corbett & Anderson, 1994). What these frameworks have in common is that they are based on cognitive characteristics of knowledge.

In the present study, a framework is used that is based more strongly on behavior yet related to cognitive representations. It is from the Knowledge-Learning-Instruction (KLI) framework, in which different kinds of knowledge are characterized in terms of their functioning in performing on a set of related tasks (Koedinger, Corbett, & Perfetti, 2012). In the focus of the KLI framework is the alignment of different kinds of knowledge with relevant assessments and instruction in order to optimally promote learning. In the KLI framework, verbal knowledge is distinguished from non-verbal knowledge. Verbal and non-verbal knowledge are defined based on the kind of behavior they enable. Verbal knowledge enables explaining something but not doing it, while non-verbal knowledge enables doing something but not explaining it, and if both kinds of knowledge are given, they enable explaining something and doing it (Koedinger et al., 2012). This distinction can be related to the other frameworks that are based on cognitive characteristics of knowledge, but it is defined based on what one can do2.

The distinction between verbal and non-verbal knowledge does not imply that no general verbal ability is required to perform on non-verbal knowledge tasks. Also on non-verbal knowledge tasks, children have to read or follow verbal instructions, questions, and contextual information. The distinction between verbal and non-verbal knowledge tasks lies in the format of the answer that has to be provided. Koedinger et al. (2012) describe the difference in the way activated knowledge functions: Verbal knowledge functions for tasks that demand readily  

2the procedural/conceptual distinction also emphasizes behavior but is more strongly based on cognitive representations (Schneider & Stern, 2010)
verbalizable knowledge about the rationale for a principle, while non-verbal knowledge functions as non-verbalized application knowledge. In the present study, verbal knowledge is therefore defined as explanation knowledge that is activated when answering open answer tasks, and non-verbal knowledge as application knowledge that is activated when answering multiple choice tasks.

Crucially, only verbal knowledge about experimentation allows children to share their ideas and insights with peers, teachers, or parents. Developing verbal knowledge enables students to explain principles, and developing non-verbal knowledge enables students to apply principles. In experimental design, verbal knowledge is thus defined as the knowledge enabling students to engage in verbal tasks such as explaining principles of experimental design, and non-verbal knowledge as enabling students to engage in non-verbal tasks such as setting up an experimental design. Verbal knowledge is thus essential for engaging in argumentation. Argumentation, which can be trained in educational interventions (Kuhn et al., 2015; Kuhn & Udell, 2003) can take two forms: It can be a process of arguing with other people, but it can also be an interior, individual process (Kuhn, 2000). Individual reasoning is the basis for collaborative argumentation, in which reasons are exchanged to solve a problem or to convince (Mercier, 2011; Mercier, Boudry, Paglieri, & Trouche, 2017). Individual and collaborative argumentation are supposed to be closely related, and to trigger similar cognitive demands (Billig, 1996; Kuhn, 1991). Therefore, children’s individual verbal knowledge about experimental design can be seen as a prerequisite for reason and argumentation about experimentation in collaborative inquiry settings. Non-verbal knowledge on the other hand is a prerequisite for engaging in experimentation on one’s own; only if children have non-verbal knowledge about the underlying principles they are able to succeed in contexts that demand choosing between various options how to design an experiment, or setting up experiments on their own. Verbal and non-verbal knowledge therefore differ in their developmental and educational relevance. Notably, only verbal knowledge allows engaging in tasks that demand sharing or discussing insights in inquiry situations. For example, in inquiry-based learning, a common kind of educational intervention in science education, students often collaboratively engage in thinking processes and activities of scientists (American Association for the Advancement of Science, 1993; Furtak, Seidel, Iverson, & Briggs, 2012). In and beyond inquiry-based learning, collaboration and argumentation are crucial parts of students’
Knowledge about Experimental Design

education that support learning (Asterhan & Schwarz, 2007; Deiglmayr & Spada, 2011; Mercier et al., 2017). It is therefore of significant educational importance to examine at which age children develop verbal knowledge about experimental design, which factors influence this development, and if verbal knowledge and non-verbal knowledge go hand in hand.

3.1.4 The Present Study

The aim of the present study was to examine elementary school children’s verbal knowledge about experimental design in detail, including the reasoning children apply when they exhibit verbal knowledge, and whether its developmental level can be distinguished from that of children’s non-verbal knowledge. To this end, two different types of tasks were posed on a large sample of first- to sixth-graders. In the first type of tasks, assessing verbal knowledge, children had to judge the validity of experimental designs that they were presented in written descriptions with images. The children then had to provide written rationales for their answers, reflecting the verbal knowledge underlying their judgments. The crucial principle that the children had to understand and apply in these tasks was that all the presented experimental designs were conclusive, that is, the variable of interest was correctly varied between experimental conditions, but not in all cases the full principle of variable control (CVS) was correctly implemented. In some tasks, confounding variables were correctly controlled, and in some tasks they were not. Children’s written rationales on these tasks were coded in order to categorize which reasoning the children applied to judge the validity of the presented experimental designs. These categories were also rated as reflecting different levels of children’s verbal knowledge about experimental design. These open answer tasks thus demanded explanation knowledge and were used to quantitatively rate the level of children’s verbal knowledge about experimental design, and at the same time to qualitatively code the reasoning underlying their verbal knowledge.

In the second type of tasks, the children were again presented experimental designs, but on these tasks they had to provide multiple choice answers. In some of these multiple choice tasks, the children had to choose between different options of experimental designs, in some of which the principles of conclusive hypothesis testing or variable control were correctly applied, in others not. On some other multiple choice tasks, they had to interpret which conclusions could be drawn from experiments in which the two principles of CHT and CVS were either correctly applied or
Knowledge about Experimental Design

not. These multiple choice tasks thus demanded application knowledge and were posed on children to assess their non-verbal knowledge about crucial principles of experimental design.

The first aim with these two types of tasks was to disentangle two types of knowledge about experimental design: Children’s verbal knowledge, as shown in their written rationales on the open answer tasks, and their non-verbal knowledge, as shown in their answers to the multiple choice tasks. These two types of tasks do demand different types of answers. It is however not known whether and to which extent the verbal knowledge required to solve the open answer tasks and the non-verbal knowledge required to solve the multiple choice tasks can be separated from each other: Is children’s level of verbal knowledge perfectly correlated with their level of non-verbal knowledge, or are these two kinds of knowledge separable? If these two kinds of knowledge are separable, it would imply that some children lack one kind of knowledge or the other to different extent. Such knowledge discrepancy might influence children’s ability to engage in inquiry-based education and other collaborative and non-collaborative situations in which experimental design plays a role. The present study therefore allows distinguishing whether open answer tasks and multiple choice tasks differ in difficulty, and also whether they activate the same kind of knowledge. It might be expected that children’s overall achievement on the verbal knowledge tasks will be lower than that on the non-verbal knowledge tasks, because on the multiple choice tasks used to assess non-verbal knowledge guessing is possible (Schwichow et al., 2015; Staver, 1984, 1986). This expected difference in absolute levels of scores on the two tasks however does not yet indicate to which extent children’s knowledge on the two types of tasks will be correlated and can be separated. This is a correlational question. To examine the interrelatedness of the two kinds of knowledge, in the present study latent variable modeling is applied, to separate assessments of the two constructs from measurement error and examine their interrelatedness on the level of latent variables (Fox, 1983).

The second aim was to examine which knowledge patterns children show on the two kinds of knowledge in the different school grades. There might be systematic interindividual differences in children’s knowledge, for example such that there are children with a high level of non-verbal knowledge but a low level of verbal knowledge, or the other way round. Examining whether such patterns exist and how they differ between school grades can provide detailed information about interindividual differences in children’s knowledge about experimental design, and which patterns
teachers might expect in different school grades. To examine this question, latent class analysis is applied, a modeling technique in which homogeneous answers patterns on a number of indicator variables are searched for, and the frequency of different patterns can be related to external variables (Hickendorff, Edelsbrunner, McMullen, Schneider, & Trezise, in press). In the present study, the school grade children attended at the time of the assessment is taken into account as an important external variable. This analysis is applied to examine knowledge patterns across the different types of tasks assessing the two kinds of knowledge, and to see how the frequency of identified knowledge patterns differed across school grades.

The third aim was to examine in more detail the factors underlying children’s level of verbal knowledge in the different school grades. The discussed literature suggests a major role of biases and naïve conceptions about experimentation, such as the influence of children’s prior content knowledge on their experimental inquiry, and wrong conceptions about the aim and principles of experimentation (Schauble et al., 1991; Siler & Klahr, 2012). To get a detailed look into these and further factors in children’s development of verbal knowledge, the reasoning underlying their answers on the open answer tasks was examined using qualitative coding. The frequencies of different types of reasoning were examined across the different school grades, to examine in detail what factors contribute to children’s level of verbal knowledge about experimental design throughout elementary school.

3.2 Method

3.2.1 Sample

The sample encompassed \( N = 3026 \) first- to sixth-grade school children who worked on a pen- and paper- questionnaire assessing their verbal and non-verbal knowledge about experimental design. The children attended elementary schools from German-speaking cantons of Switzerland in which a large part of the population has the Swiss nationality. This included schools from various German-speaking cantons, such as Zurich, a densely populated area in which about 80% of elementary students’ parents have the Swiss nationality and welfare reception rate was at about 5% at the time of assessments, and Appenzell, a rural area in which unemployment rate during that time was almost below 1%. The sample comprised the whole range of SES but the rate of poverty and welfare recipients in Switzerland is generally low, almost
all residents are expected to finish higher secondary education, and about a third tertiary education (OECD, 2014).

The children were from elementary school classes that participated in the Swiss MINT Study, a longitudinal study aimed at examining the long-term effects of early-initiated science education. The subject most closely related to experimentation which exists in some form at all Swiss elementary schools is "Human beings and their environment", in which however Swiss elementary school teachers typically focus on (social) geography and biological sciences, rather than natural sciences such as physics or chemistry (Metzger & Schär, 2009). In Swiss elementary schools, during the years of the assessments for the present study there have been mostly school-specific educational standards. Only recently, educational standards across German-speaking Switzerland have been adopted, and these do not include targeted education of principles of experimentation before secondary school (D-EDK, 2016). It can thus be expected that children in the present sample had not yet received any formal education covering principles of experimental design.

At entry into the Swiss MINT Study, all students filled out the questionnaire for the present study before working on other assessments or receiving any intervention within this study. 3027 students filled out the questionnaire but the data of one student were not used because his answers were written down by a supporting teacher and this could bias some of the analysis. In the final sample there were thus 3026 students, from 146 first- to sixth-grade classes, $M_{age} = 8.98, SD = 1.40$, range = 6 - 14 (only one student was 14 years old), 48.3% ($n = 1404$) females. More than half of students were from third- or fourth-grade classes (54.76%, $n = 1657$) because in the Swiss MINT Study the focus is on recruiting from these grades. From each school grade there were over 100 students. The distribution of students by grades, including demographic characteristics, is provided in Table 3.1. The participants came from five subsequent cohorts, with assessments taking place between May 2011 and August 2015.
Table 3.1

Sample characteristics in different school grades.

<table>
<thead>
<tr>
<th>grade</th>
<th>students (%)</th>
<th>girls (%; n_{mis})</th>
<th>age (SD; n_{mis})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>240 (7.31%)</td>
<td>110 (45.8%; 0)</td>
<td>7 (0.55; 9)</td>
</tr>
<tr>
<td>2</td>
<td>753 (24.88%)</td>
<td>378 (50.2%; 22)</td>
<td>8 (0.54; 8)</td>
</tr>
<tr>
<td>3</td>
<td>984 (32.52%)</td>
<td>502 (51.0%; 36)</td>
<td>9 (0.63; 5)</td>
</tr>
<tr>
<td>4</td>
<td>673 (22.25%)</td>
<td>328 (48.7%; 59)</td>
<td>10 (0.62; 4)</td>
</tr>
<tr>
<td>5</td>
<td>262 (8.66%)</td>
<td>129 (49.2%; 0)</td>
<td>11 (0.61; 0)</td>
</tr>
<tr>
<td>6</td>
<td>114 (3.77%)</td>
<td>58 (50.9%; 0)</td>
<td>12 (0.57; 1)</td>
</tr>
</tbody>
</table>

Note. n_{mis} = number of missing data points on respective variable for respective group. Median age in years is reported.

3.2.2 Procedure

All children worked on the questionnaire within a regular school lesson chosen by their teachers. The teachers conducted the assessments based on instructions that they received from the Swiss MINT Study team. Children sitting next to each other received one of two questionnaire versions with different task orders to prevent copying among neighbors. The students worked on the tasks on their own. In first grade classes, the teachers read the tasks aloud and went through them together with the children, to ensure that also those students who had not yet received instruction in reading could follow the procedure and work on the tasks.

3.2.3 Assessment Instrument

The questionnaire to assess children’s verbal and non-verbal knowledge about experimental design encompassed fifteen tasks. Example tasks are provided in Figure 3.1 and the whole questionnaire is provided in Appendix A. The different types of questions cover a broad construct representing CHT and the varied facets of CVS (Schwichow et al., 2016). The tasks are analogous to classic tasks such as the ramp- (Chen & Klahr, 1999), airplane- (Bullock & Ziegler, 1999), bouncing balls- (Lawson & Wollman, 1976), and mouse-task (Sodian et al., 1991). Some of the items deal with everyday situations such as plants that are grown under varying circumstances (Sodian et al., 1991), and others with situations more abstract for the students, such
as building airplanes (Bullock & Ziegler, 1999).

**Figure 3.1. Example items from the experimental design questionnaire.** (A) a verbal knowledge task; (B) a non-verbal knowledge task about conclusive hypothesis testing; (C) a non-verbal knowledge task about the control-of-variables strategy. Tasks adapted from Chen and Klahr (1999) and Bullock and Ziegler (1999). Translated by study authors.

**Verbal knowledge.** To assess students’ verbal knowledge about experimental design, the questionnaire encompassed five tasks in which the students had to provide written open answers in the form of short rationales (Figure 3.1A). In these tasks, the students read descriptions of experimental designs and then chose and justified whether the designs were appropriate for finding out about the causal relevance of a variable of interest. The crucial principles that the
Knowledge about Experimental Design

children had to understand in these tasks were that all the presented experimental designs were conclusive, that is, the variable of interest was varied between experimental trials (CHT), but not in all cases the principle of variable control (CVS) was correctly applied. In the example item (Figure 3.1A), the students had to choose that the described experiment was not good to find whether a ball rolls farther on a steep ramp than on a flat ramp, because it was a confounded design with various non-controlled variables between two conditions. The children then had to provide a written rationale for their choice, reflecting their verbal knowledge about experimental design. These open answers were coded and rated based on a categorization scheme.

**Categorization scheme.** The categorization scheme was developed to code which kind of reasoning students applied in each of their five open answers, and to rate based on these codings on which level students’ verbal knowledge was. It is thus a combined coding scheme and rating scheme (Meier, Spada, & Rummel, 2007): The coding categories are qualitative, but the different coding categories fall at the same time within different quantitative rating categories. The general approach taken to the development of the scheme was a data-driven bottom-up approach, however mixed with a top-down approach (Chi, 1997). Top-down development of coding schemes is based on available prior findings and theories about the studied phenomena. There is rich literature available describing students’ experimentation strategies and misconceptions about scientific inquiry (e.g., Schauble, 1990; Schauble et al., 1991; Siler & Klahr, 2012). For example, there is various evidence indicating that children judge experimental design based on their prior content knowledge about the investigated domain (Schauble, 1990), and on the misconception that the aim of experimentation is to produce specific effects (Schauble et al., 1991). It was expected that related reasoning schemes would occur in students’ open questions. Mainly, however, a bottom-up approach was used in order to be able to also yield new categories from the data. In prior studies, exclusively theory-driven top-down approaches were used to code whether students’ open answers fell into a limited number of pre-defined broad categories; for example, Piekny et al. (2014) coded whether children gave a pre-defined justification for a choice or not. Similarly, Kuhn et al. (1988) and Croker and Buchanan (2011) coded based on pre-defined categories whether children gave belief based or evidence based justifications. The present categorization scheme is to the best of our knowledge the first developed to freely capture the various kinds of reasoning underlying the verbal knowledge students apply in their judgments of experimental design.
Knowledge about Experimental Design

The first version of the coding and rating scheme was developed based on data of \( n = 73 \) students from four mixed grade school classes encompassing students from first to sixth grade. Based on the open answers of these students, a draft of the categorization scheme was developed. The main lead in the development was taken by both authors of the present article. A student assistant was then trained to use this draft version of the categorization scheme, and re-coded the data. In collaboration with the student assistant, the scheme was then revised based on encountered coding difficulties. The revised and final version of the coding scheme was then used to train another student assistant. The first study author and this trained student assistant then both blindly from each other coded answers of another \( n = 189 \) students to put the final version of the categorization scheme to test and estimate interrater-reliability. The final version of the scheme worked well, and interrater reliability was estimated at Cohen’s \( \kappa = .71 - .89 \) across the five tasks for the codings, and Spearman’s \( \rho = .80 - .95 \) for the ratings. These reliability indices are both in a quite high range. Consequently, the first study author coded all remaining open answers for the present study, dealing as master coder (Syed & Nelson, 2015). In Appendix C and the online supplementary materials, examples and coding rules for typical unclear cases are provided.

The scheme has three categories declared as missing data (see descriptions below) and twelve substantial coding categories that fall within three rating categories (Table 3.2). The three rating categories reflect three developmental stages that were visible in children’s open answers. In the first rating category, no knowledge (naïve conceptions), there were four coding categories. These encompassed typical student biases and misconceptions. In this rating category fell codings of answers unrelated to the question students had to answer, answers pointing out factors irrelevant to experimental design, answers based on children’s prior content knowledge, and answers based on tautological reasoning. For codings in this category, children received a rating of 0 points. In the second rating category, partial knowledge (intermediate conception), there were three coding categories. These categories included correctly mentioning a single controlled or confounding factor in an experimental design, or providing a correct general hint that too many things were varied or that little was varied. For codings in this category, children received a rating of 1 point. In the third rating category, full knowledge (scientific conception), there were four coding categories. These categories included correctly mentioning at least two or all controlled or confounding factors in an experimental design, providing a general explanation of a good
experiment, and providing a full correct proposal how the present experiment should be conducted. For codings in this category, children received a rating of 2 points. In terms of conceptual change theory, the three developmental stages extracted from children’s open answers thus describe lack of adequate knowledge or the presence of typical misconceptions (first stage; naïve conceptions), being able to and describe individual but not all factors crucial for experimental design (partial knowledge; intermediate conceptions), and being able able to describe general principles or all factors relevant for experimental design (full knowledge; scientific conception).

Non-verbal knowledge. To assess students’ non-verbal knowledge about experimental design, the questionnaire encompassed ten tasks in which the students had to provide multiple choice answers\(^3\). Five tasks were concerned with CHT and another five tasks with CVS. In the CHT tasks (Figure 3.1B), students read a short background story and then concise descriptions of three or four possible designs. They then either had to choose which design should be implemented, or, based on the description of an already conducted design, they had to choose which inferences could be drawn. In the example item, students had to choose that either of a bigger and a smaller giraffe might have eaten a carrot, because both could reach the tree it was attached to, thus this was not a conclusive test, and evidence was indeterminate (Figure 3.1B). The other five non-verbal knowledge tasks assessed CVS (Figure 3.1C). In these tasks, the children read descriptions of experimental designs and then they had to choose which design should be implemented to inform about the causal status of a variable of interest. To solve the example item, they had to choose that in order to find out whether the shape of the nose of an airplane influences fuel use, Mister Miller should build two airplanes that differ in the shape of the nose but are similar in all other characteristics, corresponding to correctly applied variable control (Figure 3.1C).

\(^3\)Originally, there were eleven tasks but in preliminary analysis, one non-verbal task turned out not to function well, with very high difficulty and small intercorrelations with the other tasks; this task was removed from the present descriptions and analysis, but for interested readers the task and all related data are included in the online supplementary materials.
3.2.4 Reliability Estimation

To estimate the questionnaire’s overall reliability, Latent reliability estimation was applied, which provides a direct estimate of reliability without taking the common detour of estimating internal consistency (e.g., Cronbach’s Alpha) or factor saturation (Raykov, Dimitrov, & Asparouhov, 2010). Ratings from the verbal knowledge tasks were re-coded from three to two levels by combining the two higher levels, because this reliability estimation method is only valid with dichotomous categorical data (Raykov et al., 2010). The reliability estimate was quite high, \( \nu = 0.83 \).
Table 3.2

*Categorization scheme excerpt for the verbal knowledge assessed in the open answer tasks: Descriptions of ratings and codings, and example answers for all twelve categories*

<table>
<thead>
<tr>
<th>Rating</th>
<th>Coding</th>
<th>Example answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>No rationale/question not answered</td>
<td>That does not interest me</td>
<td></td>
</tr>
<tr>
<td>No knowledge (naïve conception, 0 points)</td>
<td>Irrelevant/wrong rationale</td>
<td>When the heavy ball goes down, it is dangerous</td>
</tr>
<tr>
<td></td>
<td>Tautological reasoning</td>
<td>Because it is so</td>
</tr>
<tr>
<td></td>
<td>Prior content knowledge based reasoning</td>
<td>Both go equally fast</td>
</tr>
<tr>
<td>Partial knowledge (intermediate conception, 1 point)</td>
<td>Focal variable only</td>
<td>Because the engine is different(^a)</td>
</tr>
<tr>
<td></td>
<td>Single confound/controlled variable</td>
<td>Because he took two different balls(^b)</td>
</tr>
<tr>
<td></td>
<td>Single change proposal</td>
<td>He should take two balls of the same size</td>
</tr>
<tr>
<td></td>
<td>General recognition of confounding/controlling</td>
<td>Because there are too many differences</td>
</tr>
<tr>
<td>Full knowledge (scientific concept, 2 points)</td>
<td>More/all confounded/controlled variables</td>
<td>Because the ramps and balls differ(^b)</td>
</tr>
<tr>
<td></td>
<td>More/all change proposals</td>
<td>Because both cars must have the same wheels and shape(^a)</td>
</tr>
<tr>
<td></td>
<td>Generic explanation of experimental design</td>
<td>At a good experiment everything must be the same and only one thing different</td>
</tr>
<tr>
<td></td>
<td>Full design proposal</td>
<td>Only the engine must be different(^a)</td>
</tr>
</tbody>
</table>

\(^a\)Engine was focal variable in respective experiment  \(^b\)Ramp steepness was focal variable in respective experiment.
3.3 Results

Open answers of 3026 students on the five verbal knowledge tasks would yield $N \times 5 = 15130$ coded and rated answers. Only open answers for which students had chosen the correct multiple choice-option (i.e., indicating that a correctly controlled experiment was good, or that a confounded experiment was not good) were coded because the open answers to wrong multiple choice answers could not be reliably interpreted. Open answers connected to a wrong multiple choice answer received a unique code which was handled as missing data in all analysis. Across all students there were 46.76% cases (7075) with wrong multiple choice-answers, in 580 cases (3.83%) students had chosen neither of the two possible answer alternatives, and in another 2010 cases (13.28%), children had not written anything or answers were too fragmentary for interpretation. In all of these cases, answer codings were declared as missing data for all further analysis. The resulting codings included 5465 open answers. An overview of the school grades that the coded open answers stem from is provided in Table 3.3.

Table 3.3

<table>
<thead>
<tr>
<th>Grade</th>
<th>Coded answers (%)</th>
<th>girls (%)</th>
<th>Mdn age ($SD$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>79 (1.5%)</td>
<td>39 (49.4%)</td>
<td>7 (0.49)</td>
</tr>
<tr>
<td>2</td>
<td>846 (15.5%)</td>
<td>396 (46.8%)</td>
<td>8 (0.51)</td>
</tr>
<tr>
<td>3</td>
<td>1802 (33.0%)</td>
<td>914 (50.7%)</td>
<td>9 (0.64)</td>
</tr>
<tr>
<td>4</td>
<td>1629 (29.8%)</td>
<td>740 (45.4%)</td>
<td>10 (0.61)</td>
</tr>
<tr>
<td>5</td>
<td>711 (13.0%)</td>
<td>333 (46.8%)</td>
<td>11 (0.59)</td>
</tr>
<tr>
<td>6</td>
<td>398 (7.3%)</td>
<td>210 (52.8%)</td>
<td>12 (0.57)</td>
</tr>
</tbody>
</table>

3.3.1 Descriptive Statistics

The statistical approach encompassed descriptive statistics and graphical inspection and then three analyses were conducted that cover in turn the three research questions. Children’s mean ratings on the five open answer tasks were rescaled to yield an overall verbal knowledge score ranging from 0 to 1. On these verbal knowledge scores, the overall sample mean was $M = 0.18$ ($SD = 0.10$). On the non-verbal knowledge score, that is, students’ mean across the ten multiple choice tasks, which could range from 0 to 1, the overall sample mean was $M = 0.58$ ($SD$...
= 0.11). In Figure 3.2, children’s scores on both types of knowledge are depicted across school grades. In these *pirate plots* (Phillips, 2016a), for each grade a box plot with mean and 95% shaded highest density interval (Bayesian equivalent to a confidence interval; Hyndman, 1996) is provided, in combination with jittered individual data points and symmetric kernel densities (non-parametric estimate of the data distribution), being more informative than regular bar plots (Phillips, 2016b). Students’ mean scores on both types of knowledge showed a slight increase in third grade, followed by strong increase throughout fourth to sixth grade.

**Figure 3.2.** Scores on (A) verbal knowledge (open answer) tasks and (B) non-verbal knowledge (multiple choice) tasks in different school grades. In each different grade, solid lines indicates mean, shaded area 95% highest density interval, points jittered individual scores, symmetric region kernel density.

A closer look at the pirate plots indicates that in first to third grade, children’s scores were low on both types of knowledge. On verbal knowledge assessed by the open answer tasks (Figure 3.2A), first graders showed little variance. On average, the children in first grade yielded a score of about 0.1, which equals one out of ten maximum points. This average above zero stems from a combination of a small number of students who already received very high scores in first grade, including three students who received the maximum score, and a large number of students who received zero scores. In second and third grade, the average mean score stayed at a similar level, but the distribution was further spread out because a higher ratio of children in these grades managed to yield higher scores. A strong increase in verbal knowledge mean scores is visible.
from fourth to sixth grade. In sixth grade, the average is at about 0.50, and the distribution has its highest density almost at maximum, but there were also still some zero scores.

On non-verbal knowledge assessed by the multiple choice tasks (Figure 3.2B), the overall trend was very similar to that on children’s verbal knowledge. Up to third grade, the median remained between 0.4 and 0.5, indicating that in the lower grades children were not far above guessing on the multiple choice-tasks. Again however, in third grade the variance increased, as visible in the more strongly spread kernel density shape, and in fourth grade a trend of strongly increasing scores was again visible. In sixth grade, the median was above .70, with the largest part of the density shifted towards the maximum score. Thus, similar to the verbal knowledge tasks, on the non-verbal knowledge tasks sixth graders were on a high level.

3.3.2 Relations Between Verbal and Non-Verbal Knowledge

To examine whether students’ verbal and non-verbal knowledge are separable, confirmatory factor analysis models were estimated, a type of latent variable measurement model. Students’ ratings on the five open answer tasks and their scores on the ten multiple choice tasks were modeled as indicating two correlated latent random variables. A method factor was added as another latent variable unrelated to the verbal and non-verbal knowledge variables to reflect structural similarity of the five non-verbal CVS tasks. For the latent variable model, Bayesian estimation was applied (Etz, Gronau, Dablander, Edelsbrunner, & Baribault, 2017), which has favorable characteristics particularly in the smaller samples such as in first grade\(^4\) from first (\(n = 114\) students) and sixth grade (\(n = 110\) students). Technical details are provided in Appendix B.

A depiction of the confirmatory factor analysis model is provided in Figure 3.3. The model was first fit across the whole sample, and then separately for each school grade. All models showed good fit as indicated by posterior predictive p-values between .33 and .54. Values close to .50 indicate that the model predicts values similar to the empirical data (Gelman, 2013).

\(^{4}\)Some researchers deem Bayesian estimation dangerous in small samples (McNeish, 2016), but this point is trivial; ...Nobody said it was easy (Marin, Berryman, Buckland, & Champion, 2002)
Figure 3.3. Bayesian confirmatory factor analysis model: Standardized parameter estimates for whole sample. $VK =$ verbal knowledge factor (open answer tasks); $NVK =$ non-verbal knowledge factor (multiple choice tasks); $MT =$ method factor.

For the question to which degree the two types of knowledge are related or distinct, it is relevant whether the two latent variables representing children’s verbal and non-verbal knowledge are statistically separable, or whether they covary perfectly. Across the whole sample, the analysis revealed a correlation of .83 (90% credible interval = .77 - .90) between the two latent variables (Figure 3.3), and squaring this correlation coefficient indicated that the two latent variables’ variances overlap about 70%. They can thus be distinguished but overlap rather strongly.

Regarding technical specifications, variations in Bayesian prior settings influenced the correlation estimates between verbal and non-verbal knowledge in the higher grades. Under different technical specifications the correlation estimate between the two latent variables pointed towards two different solutions, stemming from bimodality in the parameter distribution. This
technical issue is discussed in Appendix B. Here, the results from the most reliable technical specification are presented, and the implications of this technical issue are taken up in the discussion.

The correlation estimate increased with increasing school grade (Figure 3.4). In first grade, factor loading estimates were low on this factor (average estimate = .28) and one loading was negative, while in all other grades factor loadings were higher and positive. First grade is therefore not interpreted. In the further grades, there was an increase in the correlation from .79 in second grade to .99 in sixth grade. In fourth grade and beyond, increase is visible. While this increase might not seem pronounced, it should be taken into account that correlation is not a linear measure of overlap between constructs; the squared correlation coefficient represents shared variance between the two latent variables. Shared variance thus showed an increase from about 62% (CI87 [.38, .90]) in second and 49% (CI87 [.35, .64]) in third grade to almost 100% in fifth and sixth grade. A substantial increase in shared variance between the two constructs was thus found, and in children from fifth and sixth grade the two types of knowledge are statistically inseparable.

![Figure 3.4](image)

*Figure 3.4. Estimated association between verbal and non-verbal knowledge in different school grades. Estimated correlation parameters $r$ with 87% credible interval error bars in dark blue, associated shared variance parameters $r^2$ in lighter blue.*
3.3.3 Verbal and Non-Verbal Knowledge Patterns

To examine patterns in children’s verbal and non-verbal knowledge, a latent class analysis was conducted across tasks. This allowed examining whether there are answer patterns of children on a low or high level across all tasks, or for example patterns of children who are on a high knowledge level across one type of tasks but rather low on the other type. In this analysis, children’s school grade was additionally taken into account as a covariate, to see how the frequency of different knowledge patterns varied between the different school grades.

Latent class analysis belongs to the family of mixture models, a person-centered analysis method in which groups of persons who share similar data patterns are modeled (Edelsbrunner et al., 2016, in press; Hickendorff et al., in press). In comparison to typical regression based analysis such as correlation and mean comparison analysis (e.g., t-test, ANOVA, multiple regression), latent class analysis does not treat all children as representing a homogeneous population. Different subgroups of children are modeled, which do not have to be specified a priori but are searched for in the data by the estimation algorithm. In this way, patterns of verbal and non-verbal knowledge were extracted based on the ratings of children’s answers on the five verbal knowledge tasks, and their multiple choice answers on the ten non-verbal knowledge tasks. Children’s school grade was included in the latent class analysis as a covariate, to see how the frequency of different knowledge patterns differs in different school grades

For this analysis maximum likelihood estimation in the Mplus software version 7.11 was used, with an expected maximization full information algorithm to be able to use also cases with missing data. Bayesian estimation would provide some advantageous characteristics (Ortega, Wagenmakers, Lee, Markowitsch, & Piefke, 2012) but it was not possible in this software to include covariates directly in a Bayesian latent class analysis, thus maximum likelihood was applied. To support estimation convergence, 1000 initial random start values were used for parameter estimation, of which 400 with the best initial convergence behavior were used for the full estimation process until the convergence criterion of \( \Delta \log l = .01 \) was reached. None of the one- to six-class solutions

\(^5\text{Covariates are often related to latent classes a posteriori by so-called three step methods (Hickendorff et al., in press). The authors of the present study are proponents of the view that it is better to include covariates directly in the model so that they can influence the latent class estimation; eventually, due to the large sample size the covariate did not substantially influence any class estimates.}\)
showed convergence problems.

The latent class analysis revealed four knowledge patterns. The BIC, a reliable criterion for model choice (Nylund, Asparouhov, & Muthen, 2007), pointed towards the four class solution, and in this solution all patterns consisted of reasonable numbers of children and provided the most meaningful interpretations. Relative fit indices and Entropy (class separation) are provided in Table 3.4.

Table 3.4

<table>
<thead>
<tr>
<th># classes</th>
<th>npars</th>
<th>logl</th>
<th>BIC</th>
<th>aBIC</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>-29524</td>
<td>59224</td>
<td>59155</td>
<td>na</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>-22436</td>
<td>45209</td>
<td>45075</td>
<td>.86</td>
</tr>
<tr>
<td>3</td>
<td>64</td>
<td>-22108</td>
<td>44729</td>
<td>44525</td>
<td>.78</td>
</tr>
<tr>
<td>4</td>
<td>86</td>
<td>-21897</td>
<td>44483</td>
<td>44210</td>
<td>.70</td>
</tr>
<tr>
<td>5</td>
<td>108</td>
<td>-21831</td>
<td>44528</td>
<td>44218</td>
<td>.71</td>
</tr>
<tr>
<td>6</td>
<td>130</td>
<td>-21785</td>
<td>44612</td>
<td>44199</td>
<td>.71</td>
</tr>
</tbody>
</table>

Note. Finally chosen was the four class-solution, based on smallest BIC and theoretical considerations. Solutions with more than six classes did not converge and are not displayed. School grade included as predictor variable in all models. npars = number of estimated model parameters; logl = loglikelihood; BIC = Bayesian Information Criterion; aBIC = sample-size adjusted Bayesian Information Criterion; E = Entropy (class separation). Lower values of BIC (mildly balanced with aBIC) indicate better model fit. Models with 7 and more classes did not converge.

The four knowledge patterns are depicted in Figure 3.5. In the figure, from left to right the estimated answer patterns are depicted for the four knowledge patterns across the five open answer tasks, then the five CHT multiple choice tasks, and then the five CVS multiple choice tasks. On the open answer tasks, two lines are visible for each knowledge pattern, indicating in black the probability of children in the respective pattern to yield a receive a rating on the highest level (scientific conception) and in gray the probability to receive a rating on the intermediate level (intermediate conception). On the multiple choice tasks, probabilities indicate whether children in the respective knowledge pattern chose the right answer option.
Across the whole sample, the largest part (47.3%) of children showed a pattern with lack of any knowledge (no knowledge pattern), as indicated by low solution probabilities on the verbal knowledge tasks, and also solution probabilities close to guessing rate on the non-verbal knowledge tasks (most tasks had three answer alternatives). All of the three remaining patterns were similar in size. The second knowledge pattern (20.2% of children) was termed the non-verbal knowledge pattern because these children were able to correctly answer the non-verbal knowledge tasks with rather high solution probabilities, but they received low ratings on their open answers on the verbal knowledge tasks. The third knowledge pattern (17.6% of children) was termed the intermediate knowledge pattern. In this pattern, students solved the verbal knowledge tasks most likely on the intermediate knowledge-level, on which most answers reflect the identification of a single controlled or confounding variable, and they could solve the non-verbal knowledge tasks concerned with CHT. Students in this pattern were however not able to solve the non-verbal CVS tasks. Finally, the smallest pattern, still including 14.9% of children, was termed smooth knowledge because in this pattern, children had knowledge on similarly high levels on all types of tasks; they had high probabilities to have ratings in the third category on their open answers to the verbal knowledge tasks, and also high probabilities to choose the correct answers on the non-verbal knowledge tasks.

On the open answer tasks, a pattern was visible indicating that probabilities of answers on the highest level increased, and probabilities of answers in the intermediate level decreased from tasks one to three (Figure 3.5 OP1-OP3) to tasks four and five (Figure 3.5 OP4, OP5). This can be explained by different structure of tasks four and five: In these tasks, only two design factors were to be considered in the described experimental designs, while in tasks one to three, three design factors had to be considered. This made it easier for children to receive a higher rating on tasks four and five and thus represents a methodological artifact.
Knowledge about Experimental Design

Figure 3.5. Task solution probabilities of children in the four knowledge patterns on all tasks. OP = open answer tasks; these have three rating categories, probability of highest rating category (scientific concepts) in darker colored lines, probability of middle rating category (intermediate conception) additionally depicted in brighter colored lines and labeled with suffix _intlev. CHT = multiple choice tasks focusing on conclusive hypothesis testing; CVS = multiple choice tasks focused on control-of-variables strategy.

The frequencies of the four classes in the different school grades are depicted in Figure 3.6. The figure shows a clear trend that with increasing school grade, the frequency of the no knowledge-class strongly decreased from about 70% of children in first grade, to about 10% of children in sixth grade. Contrary, the frequency of the smooth knowledge-profile increased strongly from almost no children in first grade to almost 60% of children in sixth grade. The intermediate knowledge pattern showed a slight decrease from almost 20% to less than 10% of children across school grades, and the non-verbal knowledge pattern a slight increase from about 10% in first grade to about 20% in the higher grades. As a further observation, fourth grade marks a turning point in knowledge: In fourth grade, the frequencies of all four knowledge patterns are almost equal, and after that in fifth grade, the smooth knowledge-class has overturned the no knowledge-class. Around fourth grade, there is thus the strong heterogeneity in children’ patterns of verbal and non-verbal knowledge about experimental design.
3.3.4 Children’s Reasoning in Different School Grades

To look more deeply at children’s reasoning in the open answer tasks, frequencies of the different types of reasoning according to the categorization scheme were analyzed. Frequencies of the different types of reasoning that the children exhibited in the different school grades are provided in Figure 3.7.

In the categories representing naïve conceptions (Figure 3.7A), the frequency of rationales unrelated to the question children were supposed to answer decreased strongly from about 40% in the lower grades to about 20% in the higher grades.

The frequency of children mentioning only a correctly varied focal variable (conclusiveness), that of tautological answers, and that of prior content knowledge related answers (e.g., engineering model, based on expected outcome) showed slight decrease over school grades.

The least decrease was visible in reasoning based on factors irrelevant to experimental design; it stayed on a level of 6-8% of answers throughout first to sixth grade. In the categories representing intermediate conceptions (Figure 3.7B) similar moderate increase was evident for all categories, and in the categories representing scientific conceptions (Figure 3.7C), strong increase was visible that initiated around fourth grade.

Figure 3.6. Frequencies of knowledge patterns in different school grades.
In a large-scale cross-sectional study, first- to sixth-graders’ verbal and non-verbal knowledge about experimental design were assessed. The aim of the study was to examine whether the two kinds of knowledge could be separated, which patterns of the two types of knowledge students showed, and how the frequency of different types of reasoning about experimental design varied between the different school grades. Results show that the two types of knowledge can be separated in the lower grades, but in the higher grades they are inseparable.
There were four knowledge patterns the frequency of which differed strongly between school grades, with strong heterogeneity of knowledge patterns in fourth and fifth grade. In students’ reasoning about experimental design, typical naïve conceptions and biases from the literature were found, with a breadth of student conceptions that could be represented on three levels.

A latent variable model indicated that while in the upper grades the two kinds of knowledge covary perfectly, in the lower grades children’s level of verbal knowledge does not go directly hand in hand with their non-verbal knowledge. Different kinds of knowledge cannot always be reliably separated, as they usually show a developmental interplay that yields strong interrelations (Schneider, Rittle-Johnson, & Star, 2011; Schneider & Stern, 2010). The present finding shows however that distinguishing verbal and non-verbal knowledge is useful, because children’s level of verbal knowledge is not only overall lower than that of non-verbal knowledge, as indicated by different solution rates in the lower grades, but the two kinds of knowledge also do not perfectly predict from each other. Children who are able to design controlled experimental designs on their might thus particularly in the lower grades not yet be able to verbalize their understanding of the underlying principles.

This finding is in line with descriptions of verbal and non-verbal knowledge by Koedinger et al. (2012), who emphasize that verbal knowledge, which is triggered by asking for verbalization of reasoning processes, does probably not represent the same knowledge as that needed to answer multiple choice tasks. Also, they emphasize that possessing verbal and non-verbal knowledge is more complex than only one kind of knowledge. Thus, knowing and expressing a rationale for a principle of experimental design in addition to being able to apply it is more advanced and complex knowledge than the knowledge required only to apply the principle. This is in line with the present finding that developing more complex knowledge needed to solve both types of tasks takes place later than developing either type of knowledge individually.

The finding that the two types of knowledge can be separated in lower grades was examined in further detail in the latent class analysis. The four extracted knowledge patterns show that there is significant knowledge heterogeneity, with similar parts of children in each of the four knowledge patterns. Fourth grade turned out again as a crucial phase in two regards. First, in fourth grade a strong decrease in the frequency of the least proficient pattern with no knowledge of either kind initiated, and at the same time a strong increase in the frequency of the smooth
pattern with high levels of knowledge of both kinds. In sixth grade, more than 60% of children were estimated to be in the smooth knowledge pattern, in accordance with prior literature indicating that at about age 12 children have undergone a strong developmental phase (Penner & Klahr, 1996; Zimmerman, 2007). Second, in fourth grade, and also still in fifth grade, there was the strongest heterogeneity in children’s knowledge about experimental design. In these two grades, there were almost equal numbers of children who had no knowledge, high levels of both kinds of knowledge, or mixed knowledge patterns. This finding emphasizes the strong developmental changes taking place during this time. In fourth and fifth grade, teachers are thus likely to find a strong mixture of children with different prior conditions for individual and collaborative experimental inquiry in their school classes.

The robust finding of fourth grade as beginning of a transitional phase from the viewpoint of the various analyses bears significance for the Piagetian theory of child and adolescent cognitive development. Piaget and Inhelder (1958) hypothesized that knowledge about abstract principles such as CHT and CVS does develop in adolescents when they enter the formal reasoning stage but that it does neither develop nor can be taught before. On the one hand, the present findings fall in line with various studies rejecting these aspects of Piaget’s theory, because it is well known that children develop some understanding of indeterminacy, and some children even about CVS, already at preschool or early elementary school age (Croker & Buchanan, 2011; Sodian et al., 1991; van der Graaf et al., 2015, 2016). Corroborating these findings, also in the present study even a few first and second graders were able to solve all types of tasks. Descriptive statistics showed that there were three children who even in first grade managed to achieve maximum verbal knowledge scores. While this is a low number (2% of the sample), it shows that even in first grade, knowledge required for identifying and verbally describing variable control in experimental design can already be developed. This is remarkable, given that in first grade children usually just learn how to read and write, and apparently quite immediately some of them develop knowledge about how to verbalize knowledge about experimental design in written rationales. From prior literature it is however evident that in most children, the knowledge about experimental design that develops in early childhood is limited to precursory knowledge on simple choice tasks, as shown by prior evidence that children cannot solve more complex tasks (Kuhn et al., 1988; Kuhn et al., 1995; Schauble, 1996). The present findings strengthen, assessed
on a large-scale, prior findings that age 10 to 12 is a crucial phase for not all, but most children in developing more advanced knowledge about experimental design. Experimental design is generally not taught in Swiss elementary schools. A strong developmental phase thus seems to take place around fourth grade without children receiving any targeted educational intervention. It might thus be doubted that a developmental phase in the classical formal reasoning sense of Piaget exists, but a similar phase, triggered by general education without explicit instruction of experimental design, seems to take place in many children. This finding, suggesting a differentiated picture of children’s development, might be incorporated into newer theories advancing classical Piaget (Demetriou, Shayer, & Efklides, 2016).

A closer look at children’s reasoning underlying their verbal knowledge brought insights into a breadth of conceptions throughout the different school grades. Children’s naïve conceptions encompassed the common finding that prior content knowledge, sometimes in combination with the engineering model-assumption that producing a good outcome is the aim of experimentation, is a major biasing factor influencing children’s reasoning about experimental design, particularly in lower grades (Zimmerman, 2007). The frequency of prior content knowledge bias decreased across school grades, however it stayed at a moderate rate throughout the higher grades: In the higher grades, still more than 10% of all answers were based on children’s prior content knowledge. Thus, this finding is in accordance with prior findings that at about age 10, children are still likely to design experiments such that they would demonstrate the correctness of their beliefs, while 12- to 14-year olds are already more likely to design controlled experiments with the aim to establish causal relations (Penner & Klahr, 1996). It shows however that even in the higher grades, there is still a moderate number of children who are biased by their prior content knowledge or have the wrong assumption that the aim of experimentation is to produce favorable outcomes (Schauble et al., 1991). This finding might bear on the assumption that verbal knowledge is prerequisite for engagement in collaborative experimental inquiry: It is well known that the influence of prior content knowledge does not only influence children’s inquiry when working on their own, but it also influences their collaborative inquiry behavior and resulting learning (Gijlers & De Jong, 2005). In educational interventions it should thus be taken into account that even in higher school grades, a non-negligible number of elementary school children might still consider their prior content knowledge in situations in which it can negatively bias
their experimentation behavior.

The number of answers based on reasoning that was unrelated to the aim of the question (e.g., "I like this experiment") also strongly decreased throughout the school grades. This indicates that in the higher school grades, less children have the assumption that experimentation is done for the sake of entertainment or play. Another observation is the quite constant frequency of about 10% of answers that indicated children understood the question but emphasized factors not relevant to the quality of the experimental design. This often included answers emphasizing environmental or safety concerns (e.g., "It might be dangerous when the ball goes down so fast"); "This is a bad experiment because cars need a lot of gasoline and smell"). In a quite large number of cases, such everyday concerns represented a bias that distracted the children from the relevant aspects and aim of the task. Only in sixth grade, there was a considerable decrease in the frequency of these answers.

On the second level, termed intermediate conceptions, children were able to correctly name a variable that was controlled or confounding, or they gave a general hint at the conclusiveness or confounding of an experimental design. The frequency of reasoning falling in these categories increased slightly across school grades, indicating more advanced knowledge in the older children. The strongest increase, starting in fourth grade, was however visible on the third level, with more than 60% of children’s answers falling in this level in sixth grade. Reasoning categories on the two higher levels thus indicated a general increase in children’s knowledge about experimental design, while the categories on the lowest level provided more detailed insights into the prevalence children’s naïve conceptions in the different school grades.

The well-functioning categorization scheme developed in the present study illustrates the wealth of information that can be extracted from children’s open answers, and it indicates that children’s verbal knowledge about experimental design can be described within the framework of conceptual change theory. How children develop through these three developmental levels could be investigated in longitudinal designs using similar types of tasks. The present study was aimed at providing a cross-sectional overview. While this provided some highly interesting findings, the data are limited in informing about students’ individual development. For example, the frequency of answers reflecting intermediate knowledge increased up to fourth grade but then decreased in fifth and sixth grade. It can only be conjectured that this initial increase and following decrease
reflects that in the highest grades, students were more likely to show the more advanced smooth knowledge pattern. It is however not apparent from the present data whether children develop from the more naïve conceptions-laden patterns through the intermediate knowledge pattern into the smooth knowledge pattern. Generally, children’s development through different knowledge patterns can be quite idiosyncratic, yet often underlain by systematics (Edelsbrunner et al., in press; Schneider & Hardy, 2013; Vosniadou & Brewer, 1992). Longitudinal data, for example assessing students twice or more in subsequent school grades, would allow answering how students transition between these different knowledge patterns and thus help exploring and corroborating the present cross-sectional findings further (Edelsbrunner et al., 2016).

To further disentangle to which extent effects from different task types reflect method variance, and to which extent different tasks trigger different kinds of knowledge, it might be informative to combine elaborate tasks with multiple choice answers, such as those implemented in the present study, and open answers directly within the same task. This would lead to statistical dependence of the two types of answers within the same tasks. Taking into account this within-item dependence in random effects models or testlet models might however enable circumventing the statistical issues arising from such a design (de Boeck & Wilson, 2004; Rijmen, 2010), and allow further disentangling effects of task context from those of different kinds of knowledge.

The present study provides a detailed picture of the heterogeneity in children’s knowledge patterns that teachers can expect in different elementary school grades. Children in the present sample stem from diverse but overall rather high socioeconomic background. In future studies, informative insights might be gained by assessing variables related to children’s socioeconomic background, and taking into account variance on the level of school classes and schools. In the present statistical analysis, dependencies within classrooms, teachers, and schools were not taken into account. Some approaches to do this in multilevel mixture modeling have been recently developed (Fagginger Auer, Hickendorff, Van Putten, Béguin, & Heiser, 2016; Mutz & Daniel, 2013) and applying these in combination with covariates on the level of classrooms, teachers, or schools might help explain part of the heterogeneity in children’s knowledge patterns. Also, in the latent variable model the inspection of the Bayesian parameter estimates indicated that the interrelation between the two latent variables shows a slightly bimodal distribution. In similar
studies, it might be examined for example with mixture models whether this bimodality can be traced back to a mixture of two or more underlying distributions, or whether other statistical factors are related to this finding.

**Conclusion**

A general insight of the present study is that children’s knowledge about experimental design, at least up to fifth grade, is very heterogeneous. This is visible when looking at the difference between students’ verbal and non-verbal knowledge, and the various patterns in these two kinds of knowledge that are present to similar degrees particularly in fourth and fifth grade. This heterogeneity was also evident at a more fine grained level, looking at children’s reasoning when they exhibit verbal knowledge. Knowledge about experimental design thus shows high interindividual variance in its level and also in the underlying reasoning. Teachers should be aware that up to late elementary school, they are likely to encounter children with a number of distinct knowledge patterns and naïve conceptions in their school classes. Children might differ in their capabilities and needs when they engage in experimentation, and not all might yet be able to verbalize their knowledge about experimental design.

**Online supplementary materials**

Publicly stored supplementary materials are available under https://osf.io/qgvfs/.

**Acknowledgments**

The Swiss MINT Study was funded by the Jacobs Foundation from 2010–2014. We are grateful to Nora Müller and Sonja Peteranderl for assistance with data codings.

**References**


Knowledge about Experimental Design


Knowledge about Experimental Design


Phillips, N. (2016b). "yarr!: the pirate’s guide to r".


doi:10.1016/j.learninstruc.2007.07.001
Chapter 4

Variable Control and Conceptual Change: A Large-Scale Quantitative Study in Elementary School

Peter A. Edelsbrunner, Lennart Schalk, Ralph Schumacher, Elsbeth Stern
ETH Zürich

Abstract

We examined the predictive value and interplay of elementary school students’ understanding of the control-of-variables strategy, a domain-general experimentation skill, and their prior content knowledge for subsequent conceptual knowledge acquisition and conceptual change. Trained teachers provided $N = 1809$ first to sixth graders with 15 lessons of guided inquiry-based instruction on the topic floating and sinking. We assessed students’ understanding of variable control before instruction, and their conceptual content knowledge before and after the instruction. Mixture model analyses indicate that the understanding of variable control predicts students’ content knowledge structure before instruction, and their content knowledge development from before to after instruction. Thus, reinstating prior lab-based findings, the understanding of a basic experimentation skill matters in teacher-guided inquiry, in interaction with prior content knowledge. We describe how students’ learning pathways vary depending on their understanding of variable control and prior content knowledge, and discuss implications for
4.1 Introduction

Conceptual change research has yielded many insights into students’ development of conceptual knowledge. These insights have stimulated the generation of elaborate science curricula in kindergarten (Leuchter, Saalbach, & Hardy, 2014), elementary (Hardy, Jonen, Möller, & Stern, 2006), and early secondary school (Smith, 2007). Often, science education in these first stages of schooling is based on inquiry. In groups of students who receive inquiry-based science instruction, inquiry is an effective instructional means for developing conceptual content knowledge (Furtak, Seidel, Iverson, & Briggs, 2012; Hattie, 2009; Slavin, Lake, Hanley, & Thurston, 2014). But not all students advance to the same degree. Where does interindividual variation in content knowledge development originate from? According to conceptual change theory, differences in the prior content knowledge that students bring to class can explain this interindividual variation. Students’ knowledge representations differ because they have different experiences from everyday life and prior education (Carey, 1985, 2000).

Differences exist not only in content knowledge but also in students’ understanding of experimentation. In inquiry-based instruction students often engage in experimentation. Setting up and interpreting experiments requires adequate understanding of domain-general experimentation principles such as the control-of-variables strategy, the understanding of variable control in experimental designs (Kuhn, Black, Keselman, & Kaplan, 2000; Kuhn, Ramsey, & Arvidsson, 2015). Interindividual differences in this domain-general understanding can be expected to influence the acquisition of appropriate scientific concepts. In the present study, we aim at scrutinizing the predictive value and interplay of students’ prior content knowledge and their understanding of variable control for knowledge acquisition and conceptual change in inquiry-based science instruction.
4.1.1 Conceptual Change in Science

When students enter science classrooms, they bring prior conceptions about the instructed topics derived from their everyday experiences (Carey, 1985, 2009; Hardy et al., 2006). Imagine children hanging out at a river. Sitting at the river bank, they note that stones are on the river ground but wood floats by; later they throw small pieces of wood and flip yet another stone and see that the that the light pieces float but the stones sink to the ground. Watching a steamboat entering the port, they admire the captain whom they recognize to be essential for safe ship passage. An anchor is released and sits so firmly on the ground that it prevents any absconding of the massive iron object that floats on the water. Talking about their experiences, they come up with some explanations for their perceptions. They discuss that light things float, heavy things sink, and a captain keeps a ship floating.

The usefulness of conceptions arising from such everyday experiences is obviously often limited. The captain is not the decisive characteristic for a ship’s floating ability and not all wooden things float. But these conceptions are not generally useless. They serve sufficiently well for explaining some occurrences of floating ability. However, these conceptions reach their limits when more and more phenomena are experienced. The conceptions are wrong from a scientific point of view, because they cannot explain all occurrences of floating ability, and therefore they are called misconceptions (Michelene TH Chi & Ohlsson, 2005). The aim of science education is to help students to develop scientifically supported concepts. In the case of floating ability, these are the concepts of object density and buoyancy force. The step from misconceptions and everyday conceptions to scientific conceptions is far. Intermediate conceptions can bridge the gap (Carey, 1992; Hardy et al., 2006). These conceptions typically develop when students blend information given in instruction and their prior conceptions. Intermediate conceptions are also sometimes deliberately introduced by teachers in order to simplify content, but still to prepare their students’ future science learning. For example, when children think about floating and sinking, they often give explanations such as “light things float while heavy ones sink”; a more elaborate but still limited intermediate conception would be “things made of wood float, while stones sink”. When learning science, students show diverse developmental patterns in how they change from misconceptions via intermediate conceptions to scientific concepts. To support this development, it is necessary to understand how these learning patterns are structured and
constrained, and how optimal knowledge development can be supported.

4.1.2 Conceptual Change in the Science Classroom

Powerful processes of knowledge restructuring have to be triggered to enrich students’ initial stock of misconceptions with scientific concepts or first with intermediate conceptions. These processes are referred to as conceptual change (Michelene TH Chi, 2008; Michelene TH Chi & Ohlsson, 2005; Ohlsson, 2009). For example, novices often have difficulties in recognizing meaningful relations on a deep level between prior knowledge and newly acquired knowledge (diSessa, 2008). In such cases, newly acquired knowledge is not connected with prior knowledge, leading to fragmented knowledge elements that are stored independently of each other. Knowledge fragmentation decreases when students gain sufficient conceptual understanding of a domain to integrate knowledge pieces into coherent, more general knowledge structures (Vosniadou & Brewer, 1992). This and similar processes of conceptual change allow integrated knowledge structures to be built up, for example by learning that a single principle, concept, or theory can explain different phenomena (Ohlsson, 2009).

One effective educational intervention for promoting conceptual change in science is inquiry-based learning, where students engage in the thinking processes and activities of scientists (American Association for the Advancement of Science, 1993). This often includes social, procedural, and epistemic activities such as arguing scientific ideas, engaging in experimentation, and interpreting evidence (Furtak et al., 2012). Inquiry-based learning is a successful method for teaching science across various topics and stages of schooling (Anderson, 2002; Bennett, Lubben, & Hogarth, 2007; Flick, 1995; Furtak et al., 2012; Minner, Levy, & Century, 2010; Shymansky, Hedges, & Woodworth, 1990). Particularly in combination with strong teacher guidance, students’ learning benefits in comparison to other traditional instructional methods (Furtak et al., 2012). However, learning differs not only between traditional and inquiry-based instructional conditions. Also within similar inquiry-based instructional settings (e.g., within one classroom), students learn to different degrees. These different learning gains on the one hand reflect differences in students’ prior content knowledge, but it has also been pointed out that specific domain-general experimentation skills influence students’ knowledge development.
4.1.3 Experimentation and Learning from Inquiry

A precondition for beneficial engagement in inquiry is a thorough understanding of experimental designs (Kuhn, 2002; Kuhn et al., 2000). A crucial facet of experimentation concerns varying the focal variable while keeping all other factors constant. Understanding and applying this strategy is referred to as the control-of-variables strategy (CVS), as vary-one-thing-at-a-time (VOTAT). Following this strategy allows making unambiguous causal inferences (Strand-Cary & Klahr, 2008). CVS predicts academic performance and science learning above and beyond general reasoning abilities (Bryant, Nunes, Hillier, Gilroy, & Barros, 2015; Wüstenberg, Greiff, & Funke, 2012). Most but not all children typically develop some understanding of the CVS at ages 6-10, depending on task context and the number of variables that have to be controlled (Sodian & Bullock, 2008; Zimmerman, 2007). Thus, childhood is a sensitive period not only for developing conceptions about scientific phenomena, but also for developing understanding of experimentation.

The development of conceptions about scientific phenomena and understanding of experimentation are probably not independent from each other, but exhibit mutual influence. In observational lab studies, Schauble (1990, 1996) found evidence for this interplay when she studied belief revision about causal mechanisms in observational lab studies. Students’ knowledge about causal relations influenced experimentation strategies, while students’ experimentation strategies in turn influenced the acquisition of content knowledge about causal relations. Based on these studies, it has been widely acknowledged that experimentation skills and content knowledge interplay in inquiry settings (Zimmerman, 2007). Taking these lab-based findings as a starting point, we aimed to scrutinize the generality and potential of this interrelation in classroom education.

4.1.4 The Present Study

We investigated the interrelation between experimentation skills and content knowledge development using a novel methodological approach. As mentioned, teacher-guided, inquiry-based instruction is a fruitful setting for learning science, and in lab-based inquiry settings students’ experimentation skills and their knowledge development are intertwined (Furtak et al., 2012; Schauble, 1990, 1996). It is yet unknown whether and to which degree this relation matters
in classroom-based instruction. In classrooms, inquiry takes manifold forms. We provided a large number of first- to sixth-graders with teacher-guided, inquiry-based instruction on the topic *floating and sinking of objects in water* in their real school environment. In this setting, we examined whether students’ understanding of the CVS predicts their content knowledge development, particularly in interaction with their prior content knowledge.

The type of assessment and the statistical model used to analyze data represent substantial factors in studies on conceptual change (Frède et al., 2011; Straatemeier, van der Maas, & Jansen, 2008). Typical approaches to assessing conceptual change include interviews (Christou & Vosniadou, 2012; Nussbaum & Novak, 1976), drawings (Vosniadou & Brewer, 1992), concept mapping (Liu, 2004), and questionnaires with multiple choice questions (Hardy et al., 2006; Straatemeier et al., 2008) or open questions (Christou & Vosniadou, 2012). These methods can either be interpreted qualitatively, or students’ answers can be quantified and analyzed using statistical models. Qualitative interpretations have the advantage of revealing unexpected aspects of students’ beliefs. They can highlight interindividual differences in children’s content knowledge development. Quantitative analyses with larger sample sizes allow for generalizations beyond the assessed student sample. This advantage of quantification, however, usually requires to treat all students as stemming from the same population. This means that in common statistical analysis, all differences between students are considered random and the same parameter estimates are used to describe the content knowledge structure and its development across all analyzed students. Potential interindividual differences in knowledge structure and development are usually neglected when analyzing scores from assessment instruments, for example by simply comparing pretest and posttest scores across large groups of students. This approach is not in line with conceptual change theory in which qualitative knowledge differences between children are emphasized as explanation for interindividual differences in content knowledge development (Kleickmann, Hardy, Pollmeier, & Möller, 2011; M. Schneider & Hardy, 2013).

It is possible to combine the advantages of qualitative and quantitative methods in the framework of mixture modeling. Mixture modeling is a quantitative approach that allows to model interindividual differences between students. This is possible by modeling that students’ scores stem from a finite number of different populations. In conceptual change research, these populations describe different knowledge states. For example, in mathematics, knowledge states
Variable Control and Conceptual Change

can reveal themselves in different strategies students apply for solving particular mathematics problems (Fagginger Auer, Hickendorff, Van Putten, Béguin, & Heiser, 2016; McMullen, Laakkonen, Hannula-Sormunen, & Lehtinen, 2015). In geography, knowledge states can be represented in the consistency of answers indicating different cognitive models of the shape of the earth (Straatemeier et al., 2008). In physics knowledge, states can refer to groups of students with different numbers of misconceptions, intermediate conceptions, and scientific concepts about the floating ability of objects in water (M. Schneider & Hardy, 2013).

In the mentioned studies, mixture modeling could reveal evidence for prior assumptions and it also pointed towards new insights about students’ knowledge development during classroom instruction. In the present study, we extended this approach. We used large-scale longitudinal mixture modeling with covariates to examine students’ content knowledge development and how it is influenced by their understanding of the CVS. This approach allows to model that students differ in their prior content knowledge and thus combines the strengths of traditional qualitative and quantitative analysis.

We do not know from prior research whether students’ understanding of experimentation will have an impact on the extent to which they improve their domain-specific conceptual knowledge in a teacher-guided inquiry curriculum. Does teacher guidance level out or even enlarge the impact of students’ understanding of experimentation on further learning? There are arguments for both sides. When teachers guide students in setting up experiments and engage them in argumentation about the outcomes, this might sufficiently support inferences and knowledge development even for students who entered the curriculum with poor understanding of experimentation. In that sense, teachers might take the responsibility off students by explicating critical design features of experiments and how these constrain valid inferences. On the other hand, according to the widely accepted cognitive constructivist view of learning, conceptual change requires active learning processes. The role of teacher guidance first and foremost lies in guiding students’ attention towards critical features (Ziegler & Stern, 2016). While teachers can trigger students’ reasoning and point them towards key concepts, students eventually have to figure out and grasp key inferences on their own. In the instructional setting used in the present study, the students for example immerse objects of different shapes, sizes, and materials into water, to examine how these variables influence floating ability. Entering the curriculum with
more advanced understanding of the CVS can be expected to be beneficial for grasping which characteristics determine floating ability. If only one of the object characteristics changes between trials, students who understand the importance of CVS should quickly comprehend that this characteristic is relevant. When none or several main characteristics change, they should accordingly comprehend that a conclusive inference is not possible. Thus, we examined whether in a teacher-guided classroom setting, students’ understanding of the CVS matters, and whether it predicts trajectories of concept development in an inquiry-based curriculum on floating and sinking.

4.2 Method

4.2.1 Participants

The sample consisted of \( N = 1809 \) students from 108 first- to sixth-grade school classes of 50 elementary schools from German speaking cantons of Switzerland. Mean age was 9.29 years (\( SD = 1.36 \)), with an age range of 6 to 13 years. 48.6\% (\( n = 879 \)) of the students were female. The school classes were recruited to participate in the Swiss MINT Study. The Swiss MINT Study is a large-scale longitudinal study in which outcomes of early-initiated cognitively activating science education are investigated in the long term. The study was initiated at ETH Zurich in 2010 and the present sample consists of all school classes who have participated in the relevant parts of the first phase of the study. The study focused on third and fourth grade but there were more than 100 students from each grade. Specifically, there were \( n = 109 \) (6.0 \%) first graders, \( n = 206 \) (11.4 \%) second graders, \( n = 514 \) (28.4 \%) third graders, \( n = 604 \) (33.4 \%) fourth graders, \( n = 258 \) (14.3 \%) fifth graders, and \( n = 118 \) (6.3 \%) sixth graders.

4.2.2 Learning Materials

The students received instruction on the topic floating and sinking of objects in water. The instructional materials for this curriculum were developed and extensively tested at the University of Munster (see Hardy et al., 2006, for details on the materials). The materials comprise 15 lessons of teacher-guided, inquiry-based classroom instruction. In the curriculum, students engage in many hands-on experiments in which for example they compare the floating ability of objects of different size, shape, or kind. In a stepwise manner, more sophisticated explanatory
models are introduced, tested, and discussed as explanations for students’ assumptions and observations. Before and after experiments, the teachers initiate and lead discussions on the students’ prior assumptions and observations, prompting justifications for their assumptions and for their explanations of outcomes. Thus, the instructional principles used in the curriculum encompassed prior knowledge activation, self-explanations, and compare and contrast activities.

Instruction starts with directing learners’ attention to the size of objects as well as to the material they are made of. Concepts of object density, water displacement, and buoyancy force are then introduced by engaging the students in experiments that show the limits of object size, weight, and material as explanatory factors for floating ability. Students experience that object weight and size interact, and how water is displaced and pushes back against objects. The key concepts are then deduced in argumentative discussions together with the teacher, without referring to scientific definitions or more scientific terms such as mass, volume, or density. The concepts are discussed in students’ everyday language (e.g., "the water pushes back where it was displaced from", "it matters how heavy something is in comparison to how big it is") to avoid terms that elementary school students might lack the necessary prior knowledge to comprehend. The curriculum is thus aimed at developing students’ conceptual understanding of the phenomena, rather than building up fact knowledge about scientific definitions, to foster conceptual content knowledge that can be built on in more advanced future education.

The teachers received one day of training in small groups in which the study authors introduced them to the study materials, experiments, and instructional principles. The students typically received either one lesson per week instead of their usual science lessons (which in Swiss elementary schools usually encompass topics from Geography and Biology), or they received the whole curriculum within a week that was devoted to special projects. The curriculum can be used adaptively from first to sixth grade classrooms. In most first and second grade classrooms the most advanced lessons were omitted and instead some basic lessons were treated more intensively. The implementation of the curriculum took the teachers a median time of 2 months.
4.2.3 Assessments

To assess students’ content knowledge development about floating and sinking, they answered a multiple choice questionnaire before and after the instruction. The questionnaire assesses misconceptions (incorrect from a scientific view), intermediate conceptions (partially correct), and scientific concepts (fully correct) about the floating ability of objects in water.

Figure 4.1. An example item from the assessment of students’ knowledge about floating and sinking. The students had to decide whether the metal plate floats or sinks and they could choose one or more rationales for their choice. The first, third, and sixth rationales each yielded one point for their misconceptions-score, the second for their intermediate conceptions-score, and the fourth and fifth for their scientific concepts-score.

For each question, the students could choose multiple answers that represented the different types of conceptions. The questionnaire has been developed in multiple pilot studies and is a reliable indicator of knowledge development triggered by the floating and sinking curriculum (see Hardy et al., 2006). An example item that encompasses answers reflecting all three types of conceptions is provided in Figure 4.1. Students’ answers on the questionnaire yielded three scores, indicating their numbers of misconceptions, intermediate conceptions, and scientific
Variable Control and Conceptual Change

concepts. We used these three scores as indicators of students’ conceptual content knowledge about floating and sinking in our main analysis. The students could obtain maximum scores of 45 misconceptions, 12 intermediate conceptions, and 19 scientific concepts. After instruction, the questionnaire included an additional transfer test with seven questions that were not part of the first assessment. The transfer questions assessed whether the students could apply the instructed concepts to new situations that were not part of the instruction. We used students’ sum score on these seven questions as an indicator of knowledge transfer of the acquired concepts.

Before the instruction and the pretest on floating and sinking, the students answered a questionnaire to assess their understanding of the CVS. The questionnaire consisted of 14 multiple choice questions (latent reliability estimate = .75; see Raykov, Dimitrov, & Asparouhov, 2010). An example item is provided in Figure 4.2. Seven questions dealt with the evaluation or interpretation of experiments, and seven with the creation of experimental designs (Bryant et al., 2015). For six items the correct variation of the focal variable was relevant, and for eight items the control of confounding variables. We included these different types of questions to assess a broad construct representing the varied facets of CVS (Schwichow, Christoph, Boone, & Härtig, 2016). The questions were analogue to classic tasks such as the ramp- (Chen & Klahr, 1999), airplane- (Bullock & Ziegler, 1999) and the mouse-task (Sodian, Zaitchik, & Carey, 1991). All questions treated domains not related to the instructed topic. The students received a score of 1 for each correct answer and a score of 0 for each wrong answer. We used students’ mean score on the questionnaire ranging from 0 to 1 as an indicator of their understanding of the CVS. A bi-factor analysis corroborated the use of a composite score (Reise, 2012). The understanding of the CVS was assessed when the students entered the longitudinal study. For some students, this was directly before the floating and sinking assessment while others received other instructional sessions in between that were not concerned with floating and sinking. Thus, the students took the CVS assessment with a median delay of 9 months before taking the floating and sinking assessment and starting with the instruction.
4.2.4 Statistical Analysis

The student variables encompassed their mean score on the CVS questionnaire, the three scores on the floating and sinking questionnaire representing their numbers of misconceptions, intermediate conceptions, and scientific concepts before and after the instruction, and their age, gender, and school grade. Our statistical approach comprised three analyses. First, we estimated descriptive statistics and intercorrelations of the main study variables. Then, we estimated a basic regression-based model, to compare its results to those from the mixture modeling approach. For the basic regression-based model, we set up a change score model (McArdle, 2009, analytic details provided in the appendix) to estimate how much of the variance in students’ change on the three knowledge indicator variables could be explained by their CVS scores.

In the next step, we conducted a latent transition analysis (LTA), the appropriate type of
mixture model for our data. We first decided on the number of knowledge profiles to estimate based on the BIC criterion (Nylund, Asparouhov, & Muthen, 2007) and theoretical considerations. For the final model, the following parameters were estimated: a) mean and variance patterns capturing students’ knowledge profiles across the three knowledge indicators before and after instruction, b) profile sizes that indicated how many of the children showed each of the knowledge profiles before instruction, c) knowledge profile transition patterns indicating how likely the children were to transition from one to another profile after having received the instruction. Afterwards, we added students’ score on the CVS measure as a covariate to the model. This indicated whether and how students’ understanding of the CVS predicted their content knowledge profiles before instruction, and their transitions between the different content knowledge profiles from before to after instruction. We also controlled for school grade to take into account that the most advanced lessons were omitted in the lower grades.

There were 29 (1.6%) students missing at the first content knowledge assessment, and 40 (2.2%) at the second content knowledge assessment. These students were absent from school for unknown reasons. We applied full information maximum likelihood estimation for all models except for the change score model for which we applied Bayesian estimation (for an overview of Bayesian methods, see Edelsbrunner, 2014; Etz, Gronau, Dablander, Edelsbrunner, & Baribault, 2017; Wagenmakers, Morey, & Lee, 2016). Both estimators handle missing data so that listwise deletion was not required, which means we could use data from all students for the statistical estimations. The Mplus software version 7.11 was used for all analyses (Muthén & Muthén, 2012).

4.3 Results

Tables with complementary descriptive statistics (Table D3) and estimated variance-covariance (Table D2) and correlation matrices (Table D3) are provided in Appendix D.

4.3.1 Regression-based Analysis

A depiction of the change score model and analytic details are provided in Appendix E. The model estimates indicated that students’ CVS score explained 2% of variance in change in their misconceptions-scores from before to after instruction, 1% of variance in change in their
intermediate conceptions-scores, and 2% of variance in change in their scientific concepts-scores. Thus, in a traditional regression-based approach, the estimated explained variance in knowledge development by students’ CVS scores is low. Including students’ school grade or gender in this model did not change the results.

4.3.2 Latent Transition Analysis

Our main analytic approach was the latent transition analysis. We first examined the number of knowledge profiles based on students’ numbers of misconceptions, intermediate conceptions, and scientific concepts before and after instruction. We increased the number of profiles in a stepwise manner up to eight profiles. Best fit was initially obtained with seven profiles before and after instruction. Some profiles were very small at one of the assessment points, including almost no students and thus not much information. We restricted one of these profiles to contain zero students and removed others from the model until we achieved the best fitting model. The model contained four profiles before instruction, and six profiles after instruction (see Table 4.1). We controlled for grade and added students’ scores on the CVS test to examine its interrelations with students’ knowledge profiles before instruction (i.e., their prior knowledge), and with their profile transitions from before to after instruction (i.e., their knowledge development). The exploratory inclusion of gender did not show any substantial relations. Finally, we added students’ score on the transfer questions, to estimate mean values of students in the extracted knowledge profiles on this indicator.
Table 4.1

*Relative fit indices and entropy for the estimated latent transition analysis models.*

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>BIC</th>
<th>aBIC</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_1</td>
<td>60773</td>
<td>60806</td>
<td>60787</td>
<td>na</td>
</tr>
<tr>
<td>2_2</td>
<td>58431</td>
<td>58513</td>
<td>58466</td>
<td>.74</td>
</tr>
<tr>
<td>3_3</td>
<td>57490</td>
<td>57633</td>
<td>57550</td>
<td>.76</td>
</tr>
<tr>
<td>4_4</td>
<td>56921</td>
<td>57136</td>
<td>57012</td>
<td>.75</td>
</tr>
<tr>
<td>5_5</td>
<td>56713</td>
<td>57010</td>
<td>56839</td>
<td>.72</td>
</tr>
<tr>
<td>6_6</td>
<td>56494</td>
<td>56884</td>
<td>56659</td>
<td>.73</td>
</tr>
<tr>
<td>7_7</td>
<td>56358</td>
<td>56853</td>
<td>56567</td>
<td>.75</td>
</tr>
<tr>
<td>8_8</td>
<td>56280</td>
<td>56874</td>
<td>56531</td>
<td>.75</td>
</tr>
<tr>
<td>6_7</td>
<td>56344</td>
<td>56801</td>
<td>56537</td>
<td>.74</td>
</tr>
<tr>
<td>5_7</td>
<td>56339</td>
<td>56757</td>
<td>56516</td>
<td>.72</td>
</tr>
<tr>
<td>5_6</td>
<td>56485</td>
<td>56842</td>
<td>56636</td>
<td>.72</td>
</tr>
<tr>
<td><strong>4_6</strong></td>
<td><strong>56346</strong></td>
<td><strong>56725</strong></td>
<td><strong>56506</strong></td>
<td><strong>.72</strong></td>
</tr>
<tr>
<td>4_5</td>
<td>56705</td>
<td>56974</td>
<td>56818</td>
<td>.70</td>
</tr>
</tbody>
</table>

*Note.* *Model* = estimated model: Number of profiles before (first number) and after (second number) instruction; *AIC* = Akaike information criterion; *BIC* = information criterion; *aBIC* = adjusted Bayesian information criterion; *E* = entropy (degree of profile separation). Lower information criteria indicate better model-data fit, for LTA particularly the BIC. Finally selected model marked in bold.

Our first results concern students’ content knowledge profiles before and after instruction. The estimated profiles are presented in Table 4.2. We classified the profiles according to the level and prominence of the estimated indicator mean values (Marsh, Lüdtke, Trautwein, & Morin, 2009). There were three profiles with a prominent number of misconceptions on an overall low level (low misconceptions profile), on a moderate level (moderate misconceptions profile), and on a high level (high misconceptions profile). There was one profile with an above-average number of all three types of conceptions (fragmented profile, cf. M. Schneider & Hardy, 2013). There was one profile with a high number of intermediate conceptions (intermediate profile), one with high numbers of intermediate conceptions and scientific concepts (prescientific profile, cf. M. Schneider & Hardy, 2013), and one with a moderate but prominent number of scientific

Before instruction, only the three misconceptions-profiles and the fragmented profile were present. After instruction, the moderate misconceptions profile, in which most students (35%) were before instruction, was not present anymore. In addition, the three most proficient profiles, that is, the intermediate, prescientific, and scientific profiles were present only after instruction. Thus, after instruction, the students became more homogeneous in terms of less proficient content knowledge profiles, and at the same time they developed three new, proficient content knowledge profiles. On the transfer test score, the low misconceptions profile had the lowest estimate, followed by the high misconceptions profile, then the fragmented and intermediate profiles, and the prescientific and scientific profiles had the highest estimates.

Table 4.2
Knowledge structure estimates for the seven knowledge profiles.

<table>
<thead>
<tr>
<th></th>
<th>MMC</th>
<th>LMC</th>
<th>HMC</th>
<th>FRA</th>
<th>PRE</th>
<th>INT</th>
<th>SCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>1.19</td>
<td>0.48</td>
<td>2.67</td>
<td>0.92</td>
<td>-1.00</td>
<td>-1.07</td>
<td>-1.18</td>
</tr>
<tr>
<td>IC</td>
<td>-2.33</td>
<td>-0.86</td>
<td>-0.87</td>
<td>0.65</td>
<td>1.40</td>
<td>1.22</td>
<td>-0.77</td>
</tr>
<tr>
<td>SC</td>
<td>-2.11</td>
<td>-1.51</td>
<td>-0.13</td>
<td>0.77</td>
<td>2.22</td>
<td>-0.59</td>
<td>0.36</td>
</tr>
<tr>
<td>TR (SE)</td>
<td>na</td>
<td>1.06 (0.11)</td>
<td>1.86 (0.13)</td>
<td>2.41 (0.17)</td>
<td>4.40 (0.21)</td>
<td>2.82 (0.19)</td>
<td>3.98 (0.24)</td>
</tr>
<tr>
<td>PR</td>
<td>.35</td>
<td>.30</td>
<td>.29</td>
<td>.06</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>POST</td>
<td>na</td>
<td>.28</td>
<td>.05</td>
<td>.23</td>
<td>.11</td>
<td>.11</td>
<td>.21</td>
</tr>
</tbody>
</table>

Note. Standardized estimates of numbers of misconceptions (MC), intermediate conceptions (IC), and scientific concepts (SC). MMC = moderate misconceptions profile; LMC = low misconceptions profile; HMC = high misconceptions profile; FRA = fragmented profile; PRE = prescientific profile; INT = intermediate profile; SCI = scientific profile. TR = estimated score on transfer questions (maximum = 7) for profiles after instruction, with standard error in brackets. PR indicates the estimated probabilities (*100 = percentage of students) to be in the respective profile before instruction, POST after instruction. na indicates a knowledge profile was not present before or after instruction.

In Figure 4.3, interrelations are depicted between students’ CVS score and the four content knowledge profiles at pretest. There was a clear pattern: The higher students’ CVS score was, the less likely they were to begin instruction in the high or moderate misconceptions-profiles. Instead,
with increasing CVS scores, the frequency increased of students entering instruction in the low misconceptions-profile or in the fragmented profile. This result pattern indicates a positive relation between students’ understanding of the CVS and their content knowledge profiles before instruction.

Figure 4.3. Cross-sectional relation of students’ CVS score to knowledge profile frequencies before instruction. Upper panel: The four knowledge profiles before instruction, represented by students’ numbers of misconceptions (MC), intermediate conceptions (IC), and scientific concepts (SC) in each of the profiles. **HMC** = high misconceptions profile; **MMC** = moderate misconceptions profile; **LMC** = low misconceptions profile; **FRA** = fragmented profile. Lower panel: Covariation of estimated profile frequencies (percent of students in the respective profile) before instruction with students’ scores on the CVS assessment, controlling for school grade.

The results regarding our main question, whether students’ understanding of the CVS predicts profile transitions from before to after instruction, are depicted in Figure 4.4. In the figure, all transition paths are depicted the probability of which varied depending on students’ understanding of the CVS. The estimates of all transition paths for students low and high on the CVS score are provided in Appendix D, Table D1.
Figure 4.4. Relations of students’ CVS score to knowledge profile transitions from before to after instruction. Left panel: The four knowledge profiles present before instruction. Right panel: The four profiles after instruction for which transitions were predicted by students’ understanding of the CVS. HMC = high misconceptions profile; MMC = moderate misconceptions profile; LMC = low misconceptions profile; FRA = fragmented profile; INT = intermediate profile; SCI = scientific profile; PRE = prescientific profile. Gray arrows indicate transitions with decreasing probability for students with higher CVS scores, black arrows transitions with increasing probability. In accordance with arrow color, numbers between panels indicate transition probabilities for students with low CVS scores (-1.5SD, first number) and with high CVS scores (+1.5SD, second number). Curved arrow at HMC profile indicates students staying in this profile.
For students starting in the high misconceptions profile (Figure 4.4 A), the probability decreased that they would stay in this profile, while they would more likely transition into the intermediate profile. This indicates a positive predictive value of CVS. For students starting in the moderate misconceptions profile (Figure 4.4 B), with increasing CVS scores the probability decreased that they would transition into the low misconceptions-profile. Instead, the probability increased that they would transition into the scientific profile, indicating again a positive predictive value of the understanding of the CVS. For students starting in the low misconceptions-profile (Figure 4.4 C), the probability decreased that they would transition into the scientific profile, while increasing in the probability to transition into the intermediate profile. Whether this indicates a positive predictive value will be discussed. Finally, for students starting in the fragmented profile (Figure 4.4 D), the probability decreased to transition into the prescientific profile, and the probability increased to transition into the scientific profile. As we will discuss, most of these predictive patterns indicate a positive predictive value of students’ understanding of the CVS for their content knowledge development.

4.4 Discussion

We examined whether and how elementary school students’ understanding of variable control in experimentation, in interplay with their prior content-specific knowledge, predicts their content knowledge development in inquiry-based science instruction. Our findings from a large sample indicate that CVS in fact matters for concept learning in the classroom: Students’ understanding of the CVS is positively related with their prior content knowledge (i.e., more proficient knowledge structures), and with their content knowledge development (i.e., transitions to more proficient knowledge structures) during a fifteen units-curriculum on the topic *floating and sinking of objects in water*.

These findings reinstate earlier lab-based findings (Schauble, 1990, 1996) and extend them to guided inquiry-based instruction in the classroom. Even under teacher guidance, students’ understanding of the CVS matters for the extent to which they gain from a curriculum on floating and sinking.

The relation of students’ understanding of the CVS to their prior content knowledge was positive: Students’ understanding of the CVS was a positive predictor for having one of the two
more proficient knowledge profiles already before instruction. This might be partially explained by students’ broader reasoning abilities that are associated with CVS (Bryant et al., 2015) and by further background variables. For example, parents’ support in students’ everyday inquiry activities and critical thinking might contribute positively to their understanding of the CVS and also to their content knowledge development (Gleason & Schauble, 1999). Thus, students’ socioeconomic background and general reasoning abilities both might positively influence their understanding of the CVS, and also their science content knowledge, explaining the statistical association. Skills based on the CVS nevertheless have unique predictive value beyond that of general reasoning (Wüstenberg et al., 2012). We therefore assume that our results indeed point towards a direct interrelation between the understanding of the CVS and science content knowledge. Crucial for education, the present finding implies that students entering instruction with less proficient prior content knowledge tend to also have a less advanced understanding of the CVS. At the same time, our longitudinal findings indicate that students with more advanced understanding of the CVS have a higher probability to transition and change from less proficient into more proficient content knowledge profiles during the instruction. Thus, students with a good understanding of the CVS, but less advanced prior knowledge still benefit from instruction.

The latent transitional analysis offers information on how to target specific groups of students. The knowledge profile with the highest number of misconceptions (i.e. about three standard deviations above mean) after instruction remained only for students with very low CVS understanding. Consequently, teachers should particularly support students lacking proper understanding of the meaning and significance of experimentation in inquiry-based instruction. Such support might help them get rid of their misconceptions and restructure their relevant concept knowledge towards a more proficient knowledge profile.

Further results illuminate the relation of students’ understanding of the CVS and their content knowledge development from before to after instruction. Most of the predictive estimates point towards a positive predictive value of the understanding of the CVS for content knowledge development: For students starting in the two content knowledge profiles with the highest number of misconceptions, the probabilities to leave their initial profile and transition into a more proficient one increased with increased understanding of the CVS (Figure 4.4 A & Figure 4.4 B). Thus, the understanding of the CVS helped these students restructure their initial knowledge, and
get rid of most of their misconceptions. For students starting in either of the two more proficient profiles, we found an opposed pattern. Students starting in a profile with a low number of misconceptions (Figure 4.4 C) were more likely to pass into a profile dominated by intermediate conceptions when they had a good understanding of the CVS. Students starting in the fragmented profile showed the contrary pattern (Figure 4.4 D): They were more likely to end up in a profile dominated by scientific conceptions instead of both scientific and intermediate conceptions if they had a good understanding of the CVS. Their good understanding of the CVS thus helped them to either build up intermediate conceptions, or, in case they already had more intermediate conceptions initially, to get rid of these and instead develop scientific concepts. In a nutshell, the results can be summarized as follows: The understanding of the CVS helps students with lower-level prior knowledge pass into an intermediate profile. It helps those already in an intermediate profile to pass into a scientific profile. Thus, independent of the content knowledge profile students have in the beginning, a good understanding of experimentation supports their knowledge restructuring and transitioning into a more scientifically valid understanding of floating and sinking.

For students’ content knowledge development in inquiry-based instruction, these results indicate that their understanding of the CVS is a relevant constraint. If students do not understand principles of experimentation to a sufficient degree, they might not be able to grasp the importance and implications of experimentally obtained outcomes. Children often see the meaning of experimentation in producing optimal outcomes, which has been termed the engineering-approach to experimentation (Schauble, Klopfer, & Raghavan, 1991). The controlled comparison between different conditions might be meaningful for students only if they understand its value for gaining conceptual insights about causal relations, beyond the aim of using an engineering-approach for producing optimal outcomes (e.g., a floating state). For instruction, this implies that preparing students with the necessary understanding of the CVS beforehand might support their learning from experimental inquiry. Students’ understanding of the CVS can be trained efficiently with relatively low time investment, especially in teacher-directed training settings (Chen & Klahr, 1999; Lorch Jr et al., 2010; Strand-Cary & Klahr, 2008). We therefore suggest to examine in future studies whether and under which circumstances training students’ understanding of the CVS indeed raises their knowledge gains.
from subsequent inquiry-based instruction in the classroom.

A closer look at the content knowledge profiles and transition patterns offers insight into conceptual knowledge structures. No knowledge profile comprised only one type of conceptions. For example, in the profile with the highest number of misconceptions, students concurrently held moderate numbers of intermediate conceptions and scientific concepts. On the transfer tasks, students in this profile outperformed those in a profile with a lower number of misconceptions. How can this be explained? We assume this resulted from the co-existence of misconceptions with more advanced conceptions: A high number of misconceptions is not necessarily detrimental. Having some misconceptions is not a major problem as long as these do not directly contradict the adequate conceptions. There is some prior research in support of this notion. Some researchers endorse the view that the coexistence of a large number of conceptions is advantageous provided that students know or are taught which conception is useful in which context (Ohlsson, 2009). Thus, students who have misconceptions might be aware that during inquiry activities in school, they have to apply more advanced concepts in order to gain insights. In addition, even if misconceptions might interfere with students’ initial learning, in the long run it can be beneficial to discuss and reflect as many initial conceptions as possible (Kapur & Rummel, 2012). Under these conditions, powerful knowledge changes can take place given a broad conceptual basis that includes misconceptions (Ramsburg & Ohlsson, 2016).

The latent transition analysis yielded informative results in comparison to a traditional regression model. Given that interindividual differences in students’ content knowledge structure exist, results from traditional regression models are flawed because they are based on the assumption of a homogeneous multivariate normal distribution population across all students. Thus, in accordance with conceptual change theory, interindividual differences in students’ knowledge and its development can only be represented adequately in mixture models. This comes at the price of setting up, estimating, and interpreting complex models with many parameters. In our study, this was possible due to the large sample size. With smaller sample sizes, such complex models cannot be estimated because they would not yield reliable information. However, in other studies on conceptual change, moderate sample sizes of a few hundred students were sufficient for estimating and yielding reliable information from similar models (McMullen et al., 2015; M. Schneider & Hardy, 2013; Straatemeier et al., 2008). Thus, in
future quantitative studies on conceptual change, we suggest researchers to consider mixture modeling and to plan sample sizes accordingly. In case of moderate sample sizes, Bayesian estimation offers high model flexibility and ease of parameter estimation by incorporating information from prior studies into the models (see e.g., Chung, Lanza, & Loken, 2008; Etz et al., 2017). Eventually, mixture modeling offers an appropriate option for incorporating the main assumptions of conceptual change theory into statistical models. Thus, it offers a way of yielding the information that conceptual change researchers really want from quantitative data.

In our study, all teachers received the same training but this does not imply that they all delivered the same instruction. It was not our intention to strictly control the instructional characteristics because we aimed at examining learning in realistic classroom situations. Still, in future studies, it might be informative to assess variables on the teacher level to examine how differences in teaching influence students’ knowledge development. For example, the amount of teacher guidance provided during experimentation might influence to which degree students’ own understanding of the CVS matters. From a statistical view, mixture models can be estimated as multilevel models, making it possible to estimate models similar to ours with predictors on the student level and also on the teacher level (Fagginger Auer et al., 2016; Vermunt, 2003).

Our results show that students’ understanding of the CVS is an informative predictor of their content knowledge development. We did not randomly manipulate students’ understanding of the CVS in a training study. Therefore, we do not know whether the present effects are direct and to which extent they are caused by other student characteristics such as intelligence, or further endogeneity factors like feedback effects and common method variance from questionnaires (Antonakis, Bendahan, Jacquart, & Lalive, 2014). We doubt that intelligence can fully account for the present effect estimates because the predictive power of intelligence for learning is limited when students enter learning situations with relevant prior knowledge (Murayama, Pekrun, Lichtenfeld, & Vom Hofe, 2013). Its effect is already soaked up in the prior knowledge (Rütsche & Schalk, submitted; W. Schneider, Körkel, & Weinert, 1989). In addition, CVS has been found to predict science achievement beyond intelligence (Bryant et al., 2015). Furthermore, intelligence can be conceptualized as a network of interacting cognitive abilities (Conway & Kovacs, 2015; Kovacs & Conway, 2016; Van Der Maas et al., 2006). For future studies examining the causal status of the CVS based on longitudinal but non-experimental data, we
suggest to examine the role of the CVS from a network perspective to yield insights into its
dynamic relations with other constructs in the course of knowledge development. To further
scrutinize whether the CVS has a direct causal effect on students’ learning, we also re-emphasize
our proposal to examine whether training students in CVS is indeed beneficial for their content
knowledge development.

4.4.1 Conclusion

Our results indicate that students’ understanding of experimentation is relevant in
teacher-guided inquiry-based instruction in elementary school. This finding opens new prospects
for improving students’ learning. Inquiry and other types of instruction have been researched for
decades in science education. With the present study, we connected two research traditions by
relating the CVS to inquiry-based learning in the classroom: Understanding of CVS is a main
determinant for students’ knowledge and their knowledge development in inquiry-based
education. The latent profile analyses indicate that students’ conceptual change depends on their
knowledge structures. This is a significant finding for theories of conceptual change; even the
most advanced students do hold some misconceptions. We have shown that latent profile analyses
adequately and beneficially connect conceptual development theories and empirical observations.
Future research might benefit from similar approaches to study whether and to what extent
students’ cognitive preconditions shape their learning of basic and advanced science concepts.

Supplementary Materials

The full analytic data and all syntax and output files are stored and publicly available at the
Open Science Framework under https://osf.io/3x5n3/.

Acknowledgments

The Swiss MINT Study was funded by the Jacobs Foundation from 2010-2014. We are
thankful to Mikko Kainulainen and Ansgar Opitz for helpful feedback on a draft of this
manuscript and suggestions for figure design.
References


Variable Control and Conceptual Change


Rütsche, B. & Schalk, L. (submitted). Constructing and refining relational categories: Disentangling the interplay between prior knowledge, reasoning, working memory capacity, and strategy. Manuscript submitted for publication.


Chapter 5

Enhancing Elementary School Students’ Domain-General Experimentation Skills by Inquiry-Based Content-Focused Physics Education

Lennart Schalk, Peter A. Edelsbrunner, Anne Deiglmayr, Ralph Schumacher, Elsbeth Stern
ETH Zürich

Abstract

In a quasi-experimental longitudinal study, it was investigated whether inquiry-based content-focused physics instruction can indirectly promote understanding of the control of variables strategy, a domain-general experimentation skill. Third graders \((Mdn_{\text{age}} = 9 \text{ years}; N = 189)\) either received 60 lessons of basic physics instruction or underwent their regular school curriculum. Students who received the physics curricula strongly increased their conceptual content knowledge. More remarkably, they also showed advantages in transferring the control of variables strategy to novel problems, although this skill was not explicitly instructed. These findings illuminate the intertwined developmental dynamics of content knowledge and experimentation skills. Early inquiry-based physics instruction can exploit these dynamics to prepare students for future science learning.
Keywords: transfer; control of variables strategy; content knowledge

5.1 Introduction

Developing competence in science requires learners to develop domain-specific content knowledge as well as domain-general experimentation skills across educational levels (Council et al., 2012; Sandoval, Sodian, Körber, & Wong, 2014). Laboratory studies indicated that these two competence components do not develop independently, but exert mutual influence (Schauble, 1990, 1996). Can this mutual relation also be exploited in classroom instruction? We implemented intensive basic physics curricula in elementary school classrooms. These curricula were designed to support the acquisition of conceptual content knowledge by employing numerous experimentation activities in a guided inquiry approach, but never directly focused on how experiments have to be designed to allow for valid inferences. We investigated longitudinally whether this content-focused instruction does, by its strong reliance on experiments, also indirectly benefit the ability to use and apply the control of variables strategy (CVS), a domain-general experimentation skill, in novel contexts.

The CVS is a central scientific principle. It specifies that causal inferences from data obtained in an experiment can only then be drawn if only one variable is manipulated at a time (Strand-Cary & Klahr, 2008; Tschirgi, 1980). Understanding the CVS is thus necessary to generate and test causal hypotheses; that is, to design conclusive experiments and to critically evaluate outcomes of experiments (Council et al., 2012; Zimmerman, 2007). A first grasp of the CVS emerges gradually during childhood (Sandoval et al., 2014). Some 1st graders can already recognize confounded hypothesis testing as inappropriate (Sodian, Zaitchik, & Carey, 1991). Across elementary school, the ability to think scientifically, which comprises understanding of the CVS, increases constantly (Körber, Mayer, Osterhaus, Schwippert, & Sodian, 2014). Nevertheless, many secondary school students and adults fail to actively create conclusive experiments (Bullock & Ziegler, 2009; Zimmerman, 2007). This is alarming because understanding of the CVS is an important predictor for competence development in the sciences (Bryant, Nunes, Hillier, Gilroy, & Barros, 2015).

The CVS can be trained directly. Trainings have typically been short term interventions and the majority of training studies have targeted 11-15 year-olds, but even 6-year-olds can benefit
Enhancing Domain-General Experimentation

Direct and explicit teaching of the CVS can enable students to transfer the strategy to new problems and across time (Lorch Jr et al., 2010; Lorch et al., 2014; Strand-Cary & Klahr, 2008). A potential danger of direct trainings is, however, that students do not integrate their understanding of the CVS with their scientific content knowledge (Bao et al., 2009). We assume that integration requires balanced methods of instruction with a close coupling of content and experimentation.

Guided inquiry stands out as an effective instructional approach for directly training the CVS (Chen & Klahr, 1999, 2008; Strand-Cary & Klahr, 2008) but also for helping learners to develop domain-specific conceptual knowledge throughout preschool (Leuchter, Saalbach, & Hardy, 2014), elementary school (Hardy, Jonen, Möller, & Stern, 2006), and secondary school (Hanauer et al., 2006; Linn et al., 2014). Schauble (1990, 1996) demonstrated a mutual relation between the development of domain-specific scientific content knowledge and domain-general experimentation skills in moderately sized, lab-based experiments. We tested whether this finding scales up to real classroom instruction.

5.2 Method

5.2.1 Participants

The participating children represent the full first cohort of 3rd grade classes that joined the Swiss MINT Study. This is a large-scale study in which early science instruction as preparation for future learning is examined longitudinally. The sample comprises students attending elementary schools in Zurich and surrounding German-speaking cantons of Switzerland, a densely populated area in which about 80% of elementary students’ parents have the Swiss nationality. The sample comprises the whole SES range; however, the amount of welfare recipients in Switzerland is generally low (around 3%). Analyses are based on the data of those children for whom we received parental consent to store their children’s test results. Data were collected from October 2011 to July 2013. In total, twelve 3rd grade classes from elementary schools in German-speaking cantons of Switzerland with 189 students participated. Students’ median age was 9 years (range 8-11) at the beginning of the study. Of the 12 participating school classes, 6 randomly selected classes (n = 81, 37 girls) served as intervention group in which we implemented the inquiry-based physics curricula. The other 6 classes (n = 108, 58 girls) served as
a wait list control group who received their usual curriculum during the time of the study. Both intervention and control classes received the same tests assessing students’ physical content knowledge on the four curricula, and their CVS development. The numbers of participants in the two conditions are unequal because not all parents gave consent and because class sizes generally vary in Switzerland. There was no significant correlation of class size with students’ CVS scores, neither at pretest ($r = .05, p = .527$) nor at posttest ($r = .10, p = .190$).

### 5.2.2 Materials

**Learning materials.** In the intervention classes, children were received four early physics curricula that had been developed by a team of science education experts at the University of Munster, Germany (Spectra Materials - KiNT-Boxes 1-4): Floating & Sinking, Air & Atmospheric Pressure, Sound & Spreading of Sound, and Stability of Bridges. These curricula are designed to develop domain-specific content knowledge about basic physics concepts. The Floating & Sinking curriculum introduces the concepts of water displacement, object density, and buoyancy force. The Air & Atmospheric Pressure curriculum introduces air as non-visible matter that has weight and needs space. Additionally, it is demonstrated how air pressure can be used functionally and mechanically. The Sound & Spreading of Sound curriculum introduces the concepts of sound wave pitch and frequency, and wave movement. The Stability of Bridges curriculum introduces basic types of forces and principles of stable construction design. Hardy et al. (2006) provide an extensive exemplary description of one curriculum (Floating & Sinking). Each curriculum comprises all experimentation materials and necessary information (e.g., worksheets, theoretical background information about the content) for implementation through the teacher.

The four curricula employ the same core educational principles. Children frequently engage in guided experimentation to explore the different physics concepts on their own. The inquiry-based lessons are designed with a strong emphasis on instructional guidance and scaffolding. For example, prior knowledge is activated in teacher-led classroom discussions and in paper-based exercises before experimentation. Children write down their assumptions concerning the outcomes of the experiments and have to provide justifications for their expectations in a research notebook. After having conducted an experiment, children write down
the observed outcome and compare it to their expectations in their notebook. The research notebook further contains additional content-related information and exercises. In concluding the lesson, the teacher secures children’s understanding of the underlying physics concept in a teacher-led classroom discussion.

Assessments. We assessed children’s content knowledge with paper-and-pencil tests. The four tests, one for each curriculum, measure children’s domain-specific conceptual understanding of the physics concepts with multiple choice questions (see Figure 5.1 for example questions). As an estimate of reliability, we used McDonald’s omega (ω), a more robust measure than Cronbach’s Alpha (Dunn, Baguley, & Brunsden, 2014). The Floating & Sinking test comprised 11 questions (reliability estimate from posttest: \( \omega_{\text{posttest}} = .70 \)), the Air & Atmospheric Pressure test 15 questions (\( \omega_{\text{posttest}} = .67 \)), the Sound & Spreading of Sound test 17 questions (\( \omega_{\text{posttest}} = .75 \)), the Stability of Bridges test 18 questions (\( \omega_{\text{posttest}} = .62 \)). Each curriculum and its respective test covered various physics concepts, but the reliability estimates indicate that the tests measured students’ knowledge sufficiently reliably for each domain.
Figure 5.1. Example items from the content knowledge assessments. The figure presents one example for each of the assessments for the different curricula: (A) Floating & Sinking, (B) Air & Atmospheric Pressure, (C) Sound & Spreading of Sound, and (D) Stability of Bridges. Items are translated from the original German versions of the assessment.
We constructed two different item orders for each test to prevent children from copying from their neighbors. We summed up the number of correct answers to yield an indicator of children’s content-specific conceptual knowledge of the respective domain. For the Floating & Sinking curriculum, we used a fine grained conceptual understanding score that indicates how often children chose a correct concept to explain whether an object floats or sinks without choosing a complementary misconception (as suggested by Hardy et al., 2006).

We assessed children’s domain-general understanding of the control of variables strategy (CVS) also with a paper-and-pencil test. Importantly, the tasks of the CVS test are not related to any content knowledge (or contexts) provided in the curricula. Thus, this test assessed children’s ability to transfer the CVS to new contexts. The test contained 11 multiple-choice tasks and 5 open answer tasks (see Fig. 2 for examples). These tasks were analogous to well-established CVS tasks from the research literature - for example the mouse task (Sodian et al., 1991), the ramp task (Chen & Klahr, 1999), and the airplane task (Bullock & Ziegler, 2009). Following Bryant et al. (2015), we designed two kinds of multiple choice tasks. In creation tasks, children had to choose which experiment to conduct in order to examine the potential causal influence of a focal variable. In evaluation tasks, children had to value a given experimental design (i.e. whether it is a good experiment or not). The children received 1 point for each correct answer, and 0 points for incorrect answers. In the five open answer tasks, children had to decide whether a given experimental design actually allows drawing a definite conclusion and to write down a justification for their decision. Children received 0 points if their justification did not indicate any understanding of the CVS, 1 point if it referred to a single detail of the experimental design that was or was not properly controlled, and 2 points if it referred to two or more critical design features, or if they explained the rationale underlying the CVS. Two independent raters coded all answers of the ten coded items (5 at pretest, 5 at posttest) with a median interrater reliability of Spearman’s rho = .92 (range across items: .72 - .95). They resolved disagreements by discussion. Overall, the 16 tasks yielded a maximum score of 21 points (reliability at pretest $\omega = .72$, at posttest $\omega = .83$); we used the percentage of points as students’ CVS score for analyses.
Figure 5.2. Example items from the control of variables (CVS) assessment. (A) represents an example of the 11 multiple choice items; (B) represents an example of the 5 open answer items from the CVS assessment. Items were adapted from Bullock and Ziegler (2009), Chen and Klahr (1999). Items are translated from the German versions used in the present study.
5.2.3 Procedure

The four curricula encompass 60 lessons in total. Classroom teachers implemented the curricula in their regular classes. Beforehand, the teachers from the intervention classes received half a day of training for each curriculum provided by the study authors. To ensure high implementation fidelity, teachers learned and practiced all instructional sequences and experiments in the trainings. Following the teacher trainings, teachers in the intervention group implemented the inquiry-based physics curricula within the standard elementary school subject “Human beings and their environment”. In the control classes, students received their standard science instruction. In Swiss elementary schools, science instruction encompasses 4-5 lessons per week in the subject “Human beings and their environment”. This subject comprises, among others, topics from Natural Sciences, History, and Geography. Teachers, however, typically focus on the (social) geographical and biological sciences rather than on topics of physics or chemistry (Metzger & Schär, 2009). After the end of the present study, teachers of the control classes received the same training as the intervention class teachers, so that they could implement the curricula also in their classes.

All children first answered the CVS assessment as a pretest. Afterwards, the teachers in the intervention classes implemented the four physics curricula and assessed the respective content knowledge immediately before and after they instructed each topic. After finishing the fourth curriculum, children answered the CVS assessment again as a posttest. The control classes answered the same assessments in comparable time intervals. The intervention and control classes needed between 8 and 20 months from CVS pre- to posttest with a median of 16 months. At the beginning of the study all students were in 3rd grade classes, at the end all were in 4th grade.

5.3 Results

The intervention and control classes showed comparable pretest performance. Only on the Sound & Spreading of Sound test, students from the intervention classes performed moderately better than students from the control classes (difference: $d = 0.51$, $p = .001$). On the other content knowledge assessment (see Table 5.1), there were no statistical differences between intervention and control classes (all $ds < 0.23$, all $ps > .050$). More importantly, students from the intervention and control classes also did not differ on the CVS pretest score ($d = 0.05$, $p = .733$).
To examine children’s content knowledge gains on the four topics, we conducted ANOVAs and dependent t-tests. Missing values were treated with listwise deletion. Between 2 and 25 of the 189 children missed one of the content knowledge assessments per curriculum (i.e., either the pre- or posttest). All repeated measures ANOVAs comparing children’s performance across conditions showed a significant interaction, indicating that children in the intervention classes had higher learning gains than children in the control classes: Students in the intervention classes showed strong learning gains in all four curricula while students in the control group showed a moderate improvement on one test only, probably reflecting a retest-effect (see Table 1). The early physics curricula thus strongly supported the development of domain-specific content knowledge.

To examine whether children’s understanding of the CVS benefited the physics curricula, we first conducted a repeated measures ANOVA. This is the simplest statistical approach for our design, with time as a within-subjects factor (i.e. pre- and posttest performance) and condition as a between-subjects factor. This analysis yielded significant main effects of time \((F(1, 162) = 68.36, \ p < .001, \ \eta^2 p = .30)\) and intervention \((F(1, 162) = 94.61, \ p = .035, \ \eta^2 p = .03)\), and a significant interaction time x intervention \((F(1, 162) = 9.03, \ p = .003, \ \eta^2 p = .05)\). Students in the control classes scored \(M = 29\% (SD = 20\%)\) at pretest and \(M = 38\% (SD = 22\%)\) at posttest \((d = 0.50)\); students in the intervention classes scored \(M = 30\% (SD = 18\%)\) at pretest and \(M = 47\% (SD = 28\%)\) at posttest \((d = 0.75)\). Thus, supporting our prediction, students in both conditions performed similar in the pretest, but the students in the intervention classes showed stronger gains on the CVS score (see Figure 5.3).
Table 5.1

*Main characteristics and psychometric practices in reviewed studies.*

<table>
<thead>
<tr>
<th>Topic</th>
<th>Condition</th>
<th>Pretest mean (SD)</th>
<th>Posttest mean (SD)</th>
<th>Gain (d)</th>
<th>Interaction (Condition &amp; Pre- vs Posttest)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floating &amp; Sinking</td>
<td>Control</td>
<td>3.79 (1.46)</td>
<td>3.86 (1.38)</td>
<td>0.06</td>
<td>$F(1, 171) = 179.19, p &lt; .001, \eta_p^2 = 0.51$</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>3.43 (1.81)</td>
<td>8.46 (3.25)</td>
<td>1.80***</td>
<td></td>
</tr>
<tr>
<td>Air &amp; Atmospheric Pressure</td>
<td>Control</td>
<td>6.69 (2.07)</td>
<td>6.64 (1.88)</td>
<td>-0.04</td>
<td>$F(1, 181) = 160.18, p &lt; .001, \eta_p^2 = 0.47$</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>6.92 (2.37)</td>
<td>11.08 (2.03)</td>
<td>1.78***</td>
<td></td>
</tr>
<tr>
<td>Sound &amp; Spreading of Sound</td>
<td>Control</td>
<td>7.66 (2.49)</td>
<td>8.14 (2.68)</td>
<td>0.19</td>
<td>$F(1, 180) = 51.92, p &lt; .001, \eta_p^2 = 0.22$</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>9.00 (2.79)</td>
<td>12.50 (2.62)</td>
<td>1.28***</td>
<td></td>
</tr>
<tr>
<td>Stability of Bridges</td>
<td>Control</td>
<td>9.06 (1.80)</td>
<td>10.06 (2.07)</td>
<td>0.41***</td>
<td>$F(1, 162) = 109.64, p &lt; .001, \eta_p^2 = 0.22$</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>9.31 (1.96)</td>
<td>12.47 (1.73)</td>
<td>1.42***</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Gain effect sizes are Cohen’s d from dependent t-tests, with *** indicating $p < .001$. 

Enhancing Domain-General Experimentation
This analysis, however, does not acknowledge that the intervention was implemented on the level of the whole school class. Within school classes, individual development is indistinguishable from the intervention. In a repeated measures ANOVA, individual development is confounded with the development of the whole school class. Despite the limited number of school classes, we therefore modeled the multilevel structure of the data to validate the results of the repeated measures ANOVA. This statistical approach prevents the underestimation of standard errors due to the data dependencies within school classes (McCoach, 2010). In the multilevel model, we regressed the posttest score on the pretest score on the individual level, and on both the average pretest score and the intervention variable on the classroom level (see Figure 5.4).
Figure 5.4. Multilevel regression model predicting the control of variables strategy (CVS) posttest score. On the WITHIN level, the posttest score (L1-CVS_POST) is regressed only on the pretest score (L1-CVS_PRE); on the BETWEEN level, the posttest score (L2-CVS_POST) is regressed on the class average pretest score (L2-CVS_PRE) and on the intervention (INT). The dot in the WITHIN-level model indicates random intercepts across school classes. Raw parameter estimates and significance levels are presented outside brackets, standardized parameter estimates within brackets. Standardized parameters of residuals indicate percentage of nonexplained variance on level of students (L1-ε) and school classes (L2-ε). * indicates $p < .05$, *** $p < .001$.

As for the content knowledge assessments, there were some missing values for the CVS assessment. 11 children (5.8%) were absent from school at pretest, and 14 children (7.4%) at posttest. For the multilevel modeling we therefore chose full information maximum likelihood estimation and correcting for deviations from non-normality with robust estimation (using the Mplus software version 7.1, L. K. Muthén & Muthén, 2012). This approach allowed analysing the whole sample of children in the multilevel model.

On the individual within level, the pretest explained 23% of the posttest variance. Whether the ability to transfer the CVS additionally benefited from the intervention can be examined on the between level. The intraclass correlation coefficient, that is, the variance in the posttest CVS
score explained by classroom differences, was 11%. Differences that already existed between school classes at pretest explained 56% of the variance at posttest. The intervention explained another 40% of the posttest variance. A significant positive regression weight for the intervention variable ($b = 0.09, p < .05$) confirms improved performance of students in the intervention classes on the CVS posttest in comparison to students in the control classes. Essentially, the results of the repeated measures ANOVA and the multilevel modeling converge indicating that children in the intervention classes more strongly increased their ability to transfer the CVS compared to children in the control classes.

5.4 Discussion

Children who received the four inquiry-based physics curricula did not only strongly gain in content knowledge, they also gained in their ability to apply the CVS to novel contents. Their strong advantage over the control classes in content knowledge development is of course rather trivial given that the control classes did not receive instruction on these topics. But the higher gains on the CVS assessment in the intervention classes are remarkable. Even though the CVS was never directly and explicitly instructed, the various guided experimentation activities in the four curricula exemplified the CVS in manifold contexts over a long time. We assume that this high dose of experimentation supported children in developing a domain-general understanding of the CVS, as indicated by their ability to apply this experimentation skill in novel contexts and situations.

We realized this study as a classroom-based quasi-experiment. After receiving an introduction to the learning materials, teachers implemented the curricula in their classes and in their standard school lessons, substituting the topics they usually teach. This experimental design reduces control over how the teachers actually implemented the curricula. However, the curricula, which comprise all necessary teaching materials, provide strong guidance for teachers. The various experiments conducted within the curricula allowed students to inquire phenomena on their own, scaffolded by prompts in their research notebook. Obviously, the content-focused curricula worked: Students strongly gained conceptual content knowledge in every curriculum. Moreover, their intensive engagement in experimentation also helped them to improve their understanding of a domain-general experimentation skill. These findings highlight thus the
potential of early science education. Complementing previous lab-based research (Schauble, 1990, 1996), we show that the mutual relation between content knowledge and domain-general experimentation skills can be used productively in classroom-based science education.

Experimentation skills are not limited to the CVS, they encompass various other facets (Council et al., 2012; Körber, Osterhaus, & Sodian, 2015; Kuhn, Iordanou, Pease, & Wirkala, 2008; Zimmerman, 2007). Nevertheless, children’s ability to use the CVS predicts future learning across scientific domains (Bryant et al., 2015). Hence, early inquiry-based science education with a strong focus on experimentation has not only a great potential to develop children’s content knowledge but also prepares them for future science education.

Supplementary Materials

The full analytic data and all syntax and output files are stored and publicly available at the Open Science Framework under https://osf.io/4b95e/.

References


Enhancing Domain-General Experimentation


Chapter 6

General Discussion

6.1 Synopsis of the Four Studies

The four studies presented in this thesis are aimed at increasing our understanding of the measurement and interrelations of the manifold facets of scientific thinking, which can be subsumed under the two broad facets of domain-general and domain-specific scientific thinking. A unifying characteristic of the studies is that they are settled across the borders of different traditional approaches in this field. The field of research on scientific thinking has cognitive-developmental, educational, and large-scale assessment strands. Some major insights have been developed in lab contexts, while other research has been conducted in the field, that is, in schools.

The research in the current thesis crosses the distinction between qualitative and quantitative research (chapter 3), it brings earlier lab-based findings to real classrooms (chapter 4, chapter 5), it screens the application of techniques from the assessment strand in psychological and educational research (chapter 2), and examines interrelations between development and education (chapter 3, chapter 4, chapter 5).

The scientific contributions of the four studies and future directions are now discussed in turn.

Study 1 (chapter 2) examined current practices in psychometric modeling in research on scientific reasoning (i.e., predominantly domain-general scientific thinking). The study was only concerned with the measurement aspect of one of the two major facets of scientific thinking
More indirectly, in this study also aspects of education and development were involved. In discussions of domain-general measurement, it was discussed, and suggestions provided, for how scores from domain-general measures should be related to external variables representing education and development. The critical reflections undertaken in this study, therefore, could also contribute to further investigations into educational and developmental processes. Finally, insights from this study could also be transferred to domain-specific measurement. Elaborate psychometric models have been applied in prior research on domain-specific scientific thinking (Kleickmann, Hardy, Pollmeier, & Möller, 2011; Straatemeier, van der Maas, & Jansen, 2008). The suggestions from this study can be applied to domain-specific measurement, for example of conceptual knowledge about contents in Physics and other Sciences.

This study first of all brought the insight that Rasch modeling is the major approach to psychometric modeling in research on domain-general scientific thinking. The Rasch model has gained in popularity in recent years, across many fields in Psychology and beyond (Bond & Fox, 2015; Rost, Carstensen, & Von Davier, 1997; Sinkovics, Salzberger, Salzberger, & Sinkovics, 2006). In the present study, it was argued that this approach is not the universal answer to measurement, and corroborated by evidence from data simulations it was shown that it is likely that Rasch modeling might have contributed to some strong but probably biased conclusions about domain-general scientific thinking. The main point of this study is that domain-general scientific thinking (i.e., scientific reasoning) might not be as coherent in a psychological sense as

---

**Figure 6.1.** Embedding of study 1 in the research framework.
researchers believe Rasch modeling indicates.

The contribution of this study is that it revealed these issues, informs researchers and suggests additional psychometric modeling techniques that might yield more valid insights. It is hoped that the automated choice to use the Rasch model will be questioned by researchers in future studies, and that they will inform themselves, with the help of this study, about alternative approaches and the potential and limits of Rasch modeling.

In the present research, the findings from this study were taken up, first and foremost, by not using the Rasch model. This should however not be interpreted as motivated avoidance of this approach; Rasch modeling surely has the potential to yield informative insights in various fields. However, re-thinking what the Rasch model does yielded the conclusion that it is not the right tool for examining knowledge structure and its development in the present studies, which are theorized to be underlain by multidimensional data-patterns and non-linear patterns of change.

Study 2 (chapter 3) was again concerned with domain-general scientific thinking, however in a broader sense (Figure 6.2. This study encompassed aspects of measurement, education, and development of domain-general scientific thinking. The measurement aspect was approached with two specific strategies. First, qualitative data (i.e., open responses on questionnaire tasks) were coded based on a categorization scheme which was developed particularly for this purpose. This mode of assessment enabled yielding insights about a type of knowledge that would not be visible in multiple choice answers. Second, the main scaling approach was that of mixture modeling, in which individual differences in knowledge structure are explicitly acknowledged and modeled by implementing a categorical latent variable. This person-focused perspective enabled revealing non-linear knowledge patterns that would be largely hidden in Rasch modeling and other linear approaches to psychometric modeling. This approach was also combined with a variable-centered analysis (continuous latent variable modeling), and the complementary insights from both approaches show the usefulness of method triangulation in psychometric modeling. The aspects of development and education were included in the modeling of data from a large sample throughout elementary school. The sample was cross-sectional, nevertheless the analysis across school grades yielded various insights into the roles of education and development.

Finally, study 2 also looked into domain-specific scientific thinking, although rather indirectly. In the coding scheme, a frequent category, which was expected based on prior
literature, reflected the biasing influence of children’s prior, domain-specific content knowledge on their applying of variable control. In this study, a cross-sectional view on how this biasing influence develops throughout elementary school was therefore provided.

The main insight from this study is a detailed cross-sectional depiction of interindividual differences, but at the same time also patterns in children’s knowledge about experimental design. Prior research had left a rather ambiguous picture: On the one hand, studies showed that children were unable to design experiments (Penner & Klahr, 1996), on the other hand however even young children are able to choose between determinate and indeterminate designs (Sodian, Zaitchik, & Carey, 1991).

\[\text{Figure 6.2. Embedding of study 2 in the research framework.}\]

This indeterminate situation of evidence was ameliorated in the present study. The findings from this study support the view that one major reason for this contradictory evidence is developmental heterogeneity: There is a limited number of patterns in children’s knowledge, as shown in mixture model analysis, however the number of children in each of the knowledge patterns differs strongly across age, and shows biggest heterogeneity in around fourth grade. A strong developmental phase around fourth grade, which is related to heterogeneity, is thus able to provide an explanation for the prior contradictory findings.

The general limitation of this study is its cross-sectional design. In future studies, it might be highly informative to assess children for example in two encompassing years twice, or even more often, to further examine these patterns. This has been done similarly once in a prior study (Bullock & Ziegler, 2009), however that study was limited to a few tasks and thus not able to provide a detailed look into children’s knowledge structures. A well-chosen mix of tasks mixed in
combination with a longitudinal setting might thus yield additional informative insights.

**Figure 6.3.** Embedding of study 3 in the research framework.

Study 3 (Chapter 4) was the first to focus specifically on the interrelation between domain-general and domain-specific scientific thinking. This study had the broadest scope in the research model (Figure 6.3), by including a developmental aspect in domain-general scientific thinking, and developmental and educational aspects, as well as a specific measurement approach in domain-specific scientific thinking.

Regarding domain-general scientific thinking, this study included a variable representing children’s developmental level of knowledge about CVS. This variable was used as predictor for children’s domain-specific knowledge development. The specific measurement aspect was that domain-specific knowledge was assessed with a detailed questionnaire that encompasses three types of conceptions. In addition, the scaling model was again as in Study 2 a mixture model, however this time applied on developmental data to examine change during an educational intervention.

The results from this study indicate that the particular measurement approach was informative, because knowledge patterns emerged that would not be visible in traditional approaches. This was shown in a comparison with a traditional regression model. Thus, the mixture modeling approach might deal as informative tool also in future studies.
The main insight from this study is that children’s knowledge about CVS can predict details of their knowledge development when they receive inquiry-based Physics instruction. This result brings prior findings to classrooms (Schauble, 1990, 1996), and it offers various directions for future research.

The major limitation of this study is that children’s understanding of CVS was not trained but assessed cross-sectionally. The intriguing perspective for future research is to train CVS and see whether this measure influences children’s learning from inquiry instruction. This was done similarly in a seminal study by Chen and Klahr (1999); however, in that study, transfer measures were similar to the training circumstances; the context there did not include everyday-classroom education. In studies exploring this possibility, it should particularly be taken into account that different types of inquiry instruction exist, which has caused long-standing debates about the role of inquiry and guidance in science education (Kirschner, Sweller, & Clark, 2006). Thus, well-chosen designs might further clarify the role and importance of CVS understanding in inquiry-based instruction. Particularly the role of teacher guidance could be examined, for example by comparing the predictive value of CVS between a setting in which children engage in self-guided experimentation, and a setting in which children receive a form of guidance from their teachers.

Study 4 (Chapter 5) finally examined the interrelation between domain-general and domain-specific scientific thinking from the complementary direction. In this study, measurement approaches might not seem particularly specific. However, the measure for children’s knowledge about CVS in this study was a broad indicator variable that consisted both of information from multiple choice-questions, and also from coded open answers. In addition, a multilevel modeling approach was chosen for statistical analysis. This approach was deemed correct given the quasi-experimental intervention on the level of classrooms. It should however be noted that this approach is in many cases limited with low numbers of school classes.
In the present study, the multilevel model was deemed sufficiently well powered because the intervention was very intensive, with 60 lessons of classroom instruction. This extensive design might represent an issue for the possibility to replicate the present effects. Particularly in educational settings, it is generally difficult to conduct close replications of intervention studies, because contexts and educational systems differ strongly between classes, schools, and countries. Repeating such an intensive design in a similar context might therefore be difficult. The general idea behind this study should however be taken up and extended in conceptual replications. The findings from the study indicate that transfer from domain-specific to domain-general scientific thinking is possible. This supports the assumption that inquiry-based instruction might achieve a big aim if it takes place continuously over a long period of time: That children think, work, and learn like scientists (Furtak, Seidel, Iverson, & Briggs, 2012), and thus become such.

6.2 Outlook

The present research yielded various insights into the measurement and interrelations of domain-general and domain-specific scientific thinking in childhood. A main point that can be taken from this research is that a well-chosen strategy in combining theoretical questions and assumptions with a suitable approach to measurement can yield informative insights that probably go beyond what could be achieved with sticking to rather common or traditional methods. Hopefully, the present research can push other researchers to become fond of creative methodologies, and thus to provide the basis for further development, criticism, and inspiration.
The studies, taken together, cover all the aspects described in the research model for the present research: Measurement, education, and development of both facets of scientific thinking have been examined. Combining the insights from the present research, it can be concluded that measurement of both facets has been investigated, criticized, and particular approaches have been chosen in the present studies that go beyond these reviewed in prior research. These measurement approaches were utilized to yield insights about the interrelations of the two facets of scientific thinking. Thus, overall, a combination and interplay of all these aspects was involved in the present research.

The present studies complement each other, and each can contribute a piece to the bigger puzzle of domain-general and domain-specific scientific thinking. A promising venue would be to combine the various aspects in a future study, such that the dynamics of the different pieces could be examined within the same puzzle. Major insights could be gained from a combined investigation of the dynamic interrelation between domain-general and domain-specific scientific thinking, in which the development of both facets, and their mutual influence are monitored concurrently.

It is planned to conduct further investigations of this kind in the Swiss MINT Study. Currently, many students participating in this study are transitioning from primary school into secondary school. This transition offers the opportunity to compare students who have received the four basic curricula as well as further more advanced curricula within the Swiss MINT Study to their classroom peers in new schools. This situation offers a natural comparison group, and thus the possibility for quasi-experimental examinations into the impact of elaborate science education on children’s and adolescents’ scientific thinking.

References


Appendix

A  Swiss MINT Study Variable Control Test ........................................ 196
B  Technical information on Bayesian Latent Variable Models ............... 219
C  Full Categorization Scheme and Examples for Handling of Unclear Cases . 232
D  Complementary Statistics for Latent Transition Analysis .................. 237
E  Technical Information on Bayesian Latent Change Score Model ........... 241
Wir machen Experimente!
Version A

Vorname:

Nachname:

Klasse und Schule:
Lehrperson:

Datum:

Geburtsdatum:

Mädchen: ☐
Junge: ☐


Versuche, alle Aufgaben zu lösen. Halte dich nicht zu lange bei einer einzelnen Aufgabe auf.

© MINT-Lernzentrum der ETH Zürich
1. **Welche Maus ist im Haus?**

Zwei Brüder wissen, dass eine Maus in ihrem Haus ist. Wenn sie nämlich am Abend ein Stück Käse liegenlassen, dann ist der Käse am Morgen weg.

Sie haben diese Maus aber noch nie gesehen.

Klaus denkt, es ist eine kleine Maus.

Herbert denkt, es ist eine grosse Maus.

Die Brüder wollen nun herausfinden, ob sie eine grosse oder kleine Maus im Haus haben.

Sie bauen zwei Häuschen: eins mit einem grossen und eins mit einem kleinen Loch:

![Häuschen mit großem und kleinem Loch](image)

Sie stellen über Nacht das Häuschen mit dem kleinen Loch in ihre Küche und legen den Käse hinein. Am Morgen fehlt der Käse. Was wissen die Brüder jetzt?

Kreuze die richtige Antwort an:

- [ ] Sie wissen, dass die Maus gross ist.
- [ ] Sie wissen, dass die Maus klein ist.
- [ ] Es kann eine grosse oder eine kleine Maus gewesen sein.
2. Welches Flugzeug verbraucht am wenigsten Treibstoff?

Herr Müller baut Flugzeuge und möchte, dass sie möglichst wenig Treibstoff verbrauchen. Er hat verschiedene Ideen, wovon der Treibstoffverbrauch abhängen könnte:

- Er überlegt sich, dass ein Flugzeug eine spitze oder eine runde Nase haben kann.
- Er überlegt sich, dass die Höhenruder unten oder oben angebracht werden können.
- Er überlegt sich, dass ein Flugzeug doppelte oder einfache Flügel haben kann.

Herr Müller vermutet, dass ein Flugzeug mit einer spitzen Nase weniger Treibstoff verbraucht als ein Flugzeug mit einer runden Nase.

Was soll Herr Müller tun, um herauszufinden, ob die Form der Flugzeugnase für den Treibstoffverbrauch wichtig ist?

Kreuze die richtige Antwort an:

- Herr Müller muss ein paar Flugzeuge bauen und vergleichen, wie viel Treibstoff sie verbrauchen.
- Herr Müller muss zwei Flugzeuge bauen, eines mit runder Nase und eines mit spitzer Nase. Sie müssen aber sonst ganz gleich sein. Dann muss er vergleichen, wie viel Treibstoff sie verbrauchen.
- Herr Müller muss zwei ganz unterschiedliche Flugzeuge bauen, bei denen er die Nase, die Flügel und die Höhenruder unterschiedlich macht. Dann muss er vergleichen, wie viel Treibstoff sie verbrauchen.
3. Welcher Fallschirm bremst am besten?

Peter beobachtet auf einer Wanderung Fallschirmspringer. Er möchte wissen, wann ein Fallschirm am besten bremst.

Er überlegt sich, dass ein Fallschirm breit oder schmal sein kann.

Er überlegt sich, dass es schwere und leichte Fallschirmspringer gibt.

Er überlegt sich, dass ein Fallschirm oben rund oder rechteckig sein kann.

Peter vermutet, dass ein breiter Fallschirm besser bremst als ein schmaler Fallschirm.
Was soll Peter tun, um herauszufinden, ob ein breiter Fallschirm besser bremst als ein schmaler?

Kreuze die richtige Antwort an:

☐ Peter muss zwei Fallschirme beobachten, einen breiten und einen schmalen. Das Gewicht der Springer und die Form der Fallschirme müssen aber ganz gleich sein. Dann muss er schauen, welcher Fallschirm am besten bremst.

☐ Peter muss zwei ganz unterschiedliche Fallschirme beobachten, bei denen sich die Breite und die Form der Fallschirme und auch das Gewicht der Fallschirmspringer unterscheiden. Dann muss er schauen, welcher Fallschirm am besten bremst.

☐ Peter muss ein paar Fallschirmspringer beobachten und schauen, bei welchem der Fall am besten gebremst wird.
4. Welche Kugel rollt am weitesten?


<table>
<thead>
<tr>
<th>Steigung:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Er kann die Rampe steil oder flach machen.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Oberfläche:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Er kann die Oberfläche der Rampe glatt oder rau machen.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Länge:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Er kann die Rampe kurz oder lang machen.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gewicht der Kugel:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Er kann eine leichte oder schwere Kugel herunterrollen lassen.</td>
</tr>
</tbody>
</table>
Robert vermutet, dass eine Kugel weiter rollt, wenn die Rampe steil ist.

Um seine Vermutung zu überprüfen, baut er die beiden unten abgebildeten Rampen, die sich in der Steigung, der Oberfläche und der Länge voneinander unterscheiden. Dann lässt er auf der einen Rampe die leichte und auf der anderen Rampe die schwere Kugel herunterrollen. Anschließend vergleicht er, wie weit die beiden Kugeln bei den verschiedenen Rampen gerollt sind.

Rampe 1:

<table>
<thead>
<tr>
<th>Rampe 1:</th>
</tr>
</thead>
</table>

Rampe 2:

<table>
<thead>
<tr>
<th>Rampe 2:</th>
</tr>
</thead>
</table>

Ist das ein gutes Experiment, um herauszufinden, ob eine Kugel bei einer steilen Rampe weiter rollt als bei einer flachen?

Kreuze die richtige Antwort an:

☐ Ja, das ist ein gutes Experiment.
   Warum ist das ein gutes Experiment? Begründe deine Antwort:

   __________________________________________________________
   __________________________________________________________

☐ Nein, das ist kein gutes Experiment.
   Warum ist das kein gutes Experiment? Begründe deine Antwort:

   __________________________________________________________
   __________________________________________________________
5. Welcher Drachen fliegt am besten?

Anna bastelt gerne Drachen. Sie will, dass ihre Drachen möglichst gut fliegen. Sie hat ein paar Ideen, wovon es abhängen könnte, dass ein Drachen möglichst gut fliegt.

<table>
<thead>
<tr>
<th>Idee</th>
<th>Beispiel 1</th>
<th>Beispiel 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ein Drachen aus Papier oder aus Kunststoff gebastelt werden kann.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ein Drachen einen kurzen oder einen langen Schweif haben kann.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ein Drachen mit Streben aus Holz oder mit Streben aus Metall gebaut werden kann.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Anna vermutet, dass ein Drachen mit einem langen Schweif besser fliegt als ein Drachen mit einem kurzen Schweif.
Was soll Anna tun, um herauszufinden, ob Drachen mit kurzem oder langem Schweif besser fliegen?

Kreuze die richtige Antwort an:

☐ Anna muss zwei ganz unterschiedliche Drachen bauen, bei denen sie die Materialien, die Streben und die Schweiflängen unterschiedlich macht. Dann muss sie vergleichen, wie gut sie fliegen.

☐ Anna muss ein paar Drachen bauen und vergleichen, wie gut sie fliegen.

☐ Anna muss zwei Drachen bauen, einen mit einem kurzen Schweif und einen mit einem langen Schweif. Sie müssen sonstaber ganz gleich sein. Dann muss sie vergleichen, wie gut sie fliegen.
6. Welche Giraffe hat die Karotte gefressen?

Gabi, die Tierpflegerin, weiß, dass Giraffen gerne Karotten essen. Im Zoo gibt es zwei verschieden große Giraffen:

- Die große Giraffe kommt mit ihrem langen Hals an alle Baumspitzen der vier Bäume dran, die im Giraffengehege stehen.
- Die kleine Giraffe kann nur die Baumspitzen der zwei kleineren Bäume im Gehege erreichen.

Am Abend hat Gabi an der zweitkleinsten Tanne eine Karotte oben an der Baumspitze festgemacht. Am Morgen ist die Karotte weg.

Welche der Giraffen hat die Karotte gefressen?

Kreuze die richtige Antwort an:

☐ die kleine Giraffe

☐ die große Giraffe

☐ Es können beide Giraffen gewesen sein.
7. Welcher Nussknacker braucht am wenigsten Kraft?


| Er überlegt, dass ein Nussknacker aus Plastik oder aus Metall sein kann. | ![Nussknacker aus Plastik](image1.png) | ![Nussknacker aus Metall](image2.png) |
| Er überlegt, dass ein Nussknacker kleine oder grosse Zähne haben kann. | ![Nussknacker mit kleinen Zähnen](image3.png) | ![Nussknacker mit großen Zähnen](image4.png) |
| Er überlegt, dass ein Nussknacker lange oder kurze Griffe haben kann. | ![Nussknacker mit langen Griffen](image5.png) | ![Nussknacker mit kurzen Griffen](image6.png) |

Herr Nüssli vermutet, dass man beim Nüsse Knacken mit einem Nussknacker mit langen Griffen weniger Kraft braucht als mit einem Nussknacker mit kurzen Griffen.

Was soll Herr Nüssli tun, um herauszufinden, ob die Länge der Griffe eine Rolle spielt?

Kreuze die richtige Antwort an:

- Herr Nüssli muss ein paar Nussknacker bauen und vergleichen, wie viel Kraft man zum Nüsse Knacken braucht.
- Herr Nüssli muss zwei ganz unterschiedliche Nussknacker bauen, bei denen er das Material, die Grifflänge und die Grösse der Zähne unterschiedlich baut. Dann muss er vergleichen, wie viel Kraft man zum Nüsse Knacken braucht.
8. Welcher Bär hat das Boot versenkt?

Fritz ist Fischer. Er hat drei verschiedenen grosse Boote. Über Nacht kommt jedoch immer wieder ein Bär und steigt in die Boote. Sein Gewicht drückt die Boote herunter und lässt Wasser herein fließen, so dass sie untergehen.

Fritz hat drei Boote: ein kleines, ein mittleres und ein grosses. Er weiss Folgendes:

- Das kleine, das mittlere und das grosse Boot gehen unter, wenn der grosse Bär darauf steht.
- Das kleine und das mittlere Boot gehen unter, wenn der kleine Bär darauf steht. Wenn aber der kleine Bär in das grosse Boot steigt, dann passiert nichts. Das Boot ist nämlich gross genug, um den kleinen Bären zu tragen.

Fritz macht nun einen Versuch, um herauszufinden, ob ein kleiner oder ein grosser Bär nachts seine Boote versenkt. Er stellt eine Nacht lang nur ein Boot raus, um herauszufinden, welcher Bär auf dem Boot war.

**Welches Boot muss er rausstellen, damit er danach sicher weiss, welcher Bär auf seinen Booten war?**

Kreuze die richtige Antwort an:

- das kleinste Boot A
- das mittelgrosse Boot B
- das grosse Boot C
- Er kann es auf diese Weise nicht herausfinden.
9. Welche Rolle spielt die Wassermenge für das Pflanzenwachstum?

Georg ist Gärtner und möchte herausfinden, welche Rolle die Wassermenge für das Wachstum von Pflanzen hat.

Dazu lässt er zwei Pflanzen unter ganz verschiedenen Bedingungen wachsen: Er gibt ihnen unterschiedlich viel Wasser, Dünger und Sonnenlicht.

<table>
<thead>
<tr>
<th>Die eine Pflanze stellt er an einen Ort, wo es viel Sonnenlicht gibt.</th>
<th>Die andere Pflanze stellt er an einen Ort, wo es schattig ist und wenig Sonnenlicht gibt.</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Sonne" /></td>
<td><img src="image2.png" alt="Sonne" /></td>
</tr>
<tr>
<td>Der einen Pflanze gibt er viel Dünger.</td>
<td>Der anderen Pflanze gibt er wenig Dünger.</td>
</tr>
<tr>
<td><img src="image3.png" alt="Dünger" /></td>
<td><img src="image4.png" alt="Dünger" /></td>
</tr>
<tr>
<td>Der einen Pflanze gibt er viel Wasser.</td>
<td>Der anderen Pflanze gibt er wenig Wasser.</td>
</tr>
<tr>
<td><img src="image5.png" alt="Wasser" /></td>
<td><img src="image6.png" alt="Wasser" /></td>
</tr>
</tbody>
</table>

Ist das ein gutes Experiment, um herauszufinden, ob die Wassermenge für das Pflanzenwachstum eine Rolle spielt?

Kreuze die richtige Antwort an:

- ☐ Ja
- ☐ Nein
10. **Welcher Hammer braucht die geringste Kraft?**


| Er überlegt sich, dass ein Hammer einen schweren oder leichten Kopf haben kann. | ![Hammer mit schwerer Kopfart](image1) | ![Hammer mit leichter Kopfart](image2) |
| Er überlegt sich, dass ein Hammer einen flachen oder spitzen Kopf haben kann. | ![Hammer mit flachem Kopf](image3) | ![Hammer mit spitzen Kopf](image4) |
| Er überlegt sich, dass ein Hammer einen langen oder kurzen Griff haben kann. | ![Hammer mit langem Griff](image5) | ![Hammer mit kurzem Griff](image6) |

Herr Hämmerli vermutet, dass man mit einem Hammer mit langem Griff weniger Kraft braucht als mit einem Hammer mit kurzem Griff.
Was soll Herr Hämmerli tun, um herauszufinden, ob die Länge des Griffs des Hammers eine Rolle spielt?

Kreuze die richtige Antwort an:

☐ Herr Hämmerli muss zwei ganz unterschiedliche Hämmer bauen, bei denen das Gewicht der Köpfe, die Form der Köpfe und die Länge der Griffe ganz unterschiedlich sind. Dann muss er vergleichen, wie viel Kraft man zum Nägel Einschlagen braucht.

☐ Herr Hämmerli muss zwei Hämmer bauen, einen mit langem Griff und einen mit kurzem Griff. Sie müssen aber sonst ganz gleich sein. Dann muss er vergleichen, wie viel Kraft man zum Nägel Einschlagen braucht.

☐ Herr Hämmerli muss ein paar Hämmer bauen und vergleichen, wie viel Kraft man zum Nägel Einschlagen braucht.
11. Wovon hängt die Geschwindigkeit von Autos ab?

Frau Schmidt baut Autos und möchte, dass sie möglichst schnell fahren. Sie hat verschiedene Ideen, wovon die Geschwindigkeit abhängen könnte:

<table>
<thead>
<tr>
<th></th>
<th>Auto 1</th>
<th>Auto 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sie überlegt sich, dass ein Auto große oder kleine Räder haben kann.</td>
<td><img src="image1" alt="Große Räder, stark Motor, breite Form" /></td>
<td><img src="image2" alt="Kleine Räder, schwacher Motor, schmale Form" /></td>
</tr>
<tr>
<td>Sie überlegt sich, dass ein Auto einen starken oder schwachen Motor haben kann.</td>
<td><img src="image3" alt="140 PS" /></td>
<td><img src="image4" alt="70 PS" /></td>
</tr>
<tr>
<td>Sie überlegt sich, dass ein Auto eine schmale oder breite Form haben kann.</td>
<td><img src="image5" alt="Schmale Form" /></td>
<td><img src="image6" alt="Breite Form" /></td>
</tr>
</tbody>
</table>

Frau Schmidt vermutet, dass Autos mit einem starken Motor schneller sind als Autos mit einem schwachen Motor.

Sie macht nun verschiedene Experimente, um herauszufinden, ob Autos mit starkem Motor schneller sind als Autos mit schwachem Motor.

**Experiment 1:**

Sie baut zwei ganz unterschiedliche Autos, bei denen die Räder, die Motoren und die Form verschieden sind. Dann vergleicht sie, wie schnell sie fahren.

| Auto 1: große Räder, starker Motor, breite Form | Auto 2: kleine Räder, schwacher Motor, schmale Form |
Ist das ein gutes Experiment, um herauszufinden, ob Autos mit starkem Motor schneller sind als Autos mit schwachem Motor?

Kreuze die richtige Antwort an:

☐ Ja, das ist ein gutes Experiment.
   Warum ist das ein gutes Experiment? Begründe deine Antwort:
   __________________________________________________________
   __________________________________________________________
   __________________________________________________________

☐ Nein, das ist kein gutes Experiment.
   Warum ist das kein gutes Experiment? Begründe deine Antwort:
   __________________________________________________________
   __________________________________________________________
   __________________________________________________________
Experiment 2:

Sie baut zwei Autos, die verschiedene Motoren haben, sonst aber ganz gleich sind. Dann vergleicht sie, wie schnell sie fahren.

| Auto 1: grosse Räder, starker Motor, schmale Form | Auto 2: grosse Räder, schwacher Motor, schmale Form |

Ist das ein gutes Experiment, um herauszufinden, ob Autos mit starkem Motor schneller sind als Autos mit schwachem Motor?

Kreuze die richtige Antwort an:

☐ Ja, das ist ein gutes Experiment.
   Warum ist das ein gutes Experiment? Begründe deine Antwort:
   __________________________________________________________
   __________________________________________________________
   __________________________________________________________

☐ Nein, das ist kein gutes Experiment.
   Warum ist das kein gutes Experiment? Begründe deine Antwort:
   __________________________________________________________
   __________________________________________________________
   __________________________________________________________
12. Fressen Meisen Sonnenblumenkerne?

Anna und Sonja legen Sonnenblumenkerne aufs Fenstersims, um Vögel zu füttern. Das Futter ist jeden Tag weg. Sie haben schon gesehen, dass manchmal Ämseln das Futter wegpicken. In ihrem Garten gibt es aber auch noch Meisen.

- Anna glaubt, dass nur Amseln Sonnenblumenkerne fressen.
- Sonja glaubt, dass Amseln und Meisen Sonnenblumenkerne fressen.

Amseln sind deutlich grösser als Meisen.

Um herauszufinden, wer Recht hat, basteln die beiden ein Vogelhaus und legen die Sonnenblumenkerne hinein.

Sie könnten eine kleine Öffnung in das Vogelhaus machen, so dass nur Meisen ins Haus kommen.

Oder sie könnten eine grosse Öffnung in das Vogelhaus machen, so dass Meisen und Amseln ins Haus kommen können.

Was für ein Haus müssen sie bauen, um herauszufinden, ob nur Amseln oder auch Meisen Sonnenblumenkerne fressen?

Kreuze die richtige Antwort an:

- Sie müssen ein Vogelhaus mit einem grossen Eingang bauen.
- Sie müssen ein Vogelhaus mit einem kleinen Eingang bauen.
- Sie können das auf diese Weise nicht herausfinden.
13. Welcher Ball springt am besten?

Peter will sich einen Ball kaufen.

Es gibt Bälle in verschiedenen Grössen: kleine und grosse.

<table>
<thead>
<tr>
<th>kleine</th>
<th>grosse</th>
</tr>
</thead>
</table>

Es gibt Bälle aus verschiedenen Materialien: weiche Gummibälle und harte Lederbälle.

<table>
<thead>
<tr>
<th>weicher Gummiball</th>
<th>harter Lederball</th>
</tr>
</thead>
</table>

Peter will den Ball kaufen, der am besten springt. Er vermutet, dass das Material wichtig dafür ist, ob ein Ball gut oder schlecht springt. Er ist sich aber nicht ganz sicher.

Peter möchte deshalb mit einem Experiment herausfinden, ob das Material wichtig dafür ist, ob ein Ball gut springt.
Experiment 1:

Um seine Vermutung zu überprüfen, testet Peter die folgenden Bälle:

<table>
<thead>
<tr>
<th>grosser, weicher Gummiball</th>
<th>kleiner, harter Lederball</th>
</tr>
</thead>
</table>

Ist das ein gutes Experiment, um herauszufinden, ob das Material wichtig dafür ist, ob ein Ball gut springt?

Kreuze die richtige Antwort an:

☐ Ja, das ist ein gutes Experiment.
   Warum ist das ein gutes Experiment? Begründe deine Antwort:

   ____________________________________________________________

   ____________________________________________________________

   ____________________________________________________________

☐ Nein, das ist kein gutes Experiment.
   Warum ist das kein gutes Experiment? Begründe deine Antwort:

   ____________________________________________________________

   ____________________________________________________________

   ____________________________________________________________
Experiment 2:

Um seine Vermutung zu überprüfen, testet Peter die folgenden Bälle:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>kleiner, weicher Gummiball</td>
</tr>
<tr>
<td></td>
<td>kleiner, harter Lederball</td>
</tr>
</tbody>
</table>

Ist das ein gutes Experiment, um herauszufinden, ob das Material wichtig dafür ist, ob ein Ball gut springt?

Kreuze die richtige Antwort an:

☐ Ja, das ist ein gutes Experiment.
   Warum ist das ein gutes Experiment? Begründe deine Antwort:
   __________________________________________________________
   __________________________________________________________
   __________________________________________________________

☐ Nein, das ist kein gutes Experiment.
   Warum ist das kein gutes Experiment? Begründe deine Antwort:
   __________________________________________________________
   __________________________________________________________
   __________________________________________________________
14. Wie gut können Ferkel riechen?

Armin und Susanne wollen wissen, wie gut ihr Ferkel riechen kann.

Einige Tiere haben Nasen, mit denen sie viel besser riechen können als Menschen. Sie riechen also Dinge, die Menschen nichtriechen können.

- Armin denkt, dass das Ferkel sehr gut riechen kann.
- Susanne denkt, dass das Ferkel keine gute Nase hat und nur schlecht riechen kann.

Sie wollen rausfinden, ob das Ferkel eine gute oder schlechte Nase hat.

Sie haben eine Idee, wie sie ihre Vermutungen überprüfen können: Sie wollen Essen im Käfig verstecken und schauen, ob das Ferkel das Essen findet. Sie haben zwei verschiedene Nahrungsmittel, die verschieden stark riechen:

- Das Essen mit dem starken Geruch riecht so stark, dass auch Menschen es riechen können. Auch ein Tier mit schlechter Nase kann das riechen.

Sie verstecken das stark riechende Essen im Käfig des Ferkels. Das Ferkel findet das Essen sofort. Was bedeutet das?

Kreuze die richtige Antwort an:

☐ Armin hat Recht: Das Ferkel hat eine gute Nase.
☐ Susanne hat Recht: Das Ferkel hat eine schlechte Nase.
☐ Man weiss noch nicht, wer Recht hat.
Appendix B

Technical information on the Bayesian Latent Variable (Confirmatory Factor Analysis) Models

The Bayesian latent variable models were estimated in the Mplus software, version 7.11 (L. K. Muthén & Muthén, 2012).

**Model specification.** From prior analyses with the assessment instrument, it was known that the five multiple choice tasks assessing the control-of-variables strategy share common method variance. This dependency seems reasonable, as these five items are quite similar in terms of structure and answer options. Consequently, a method factor was added to the model which catches common variance in these five tasks that is not due to what they share with the other tasks. Regarding parameterization, the latent variables’ variances were set to unity for scale identification, and thus all factor loadings were freely estimated.

**Estimation.** For the estimations, MCMC with Gibbs sampling with four chains distributed for parallel computing across four processors, a minimum of 10000 draws after thinning with a factor of ten in each chain for smooth posterior distributions and a maximum of 100000 draws was used. In addition, the first half of MCMC samples were handled as burn-in. Start values were based on ML estimates obtained with numerical integration. As convergence criterion, potential scale reduction (PSR; Brooks & Gelman, 1998) equal or below 1.030 for each parameter after the first 10000 effective draws was applied. Trace plots and autocorrelation plots were inspected visually. Posterior medians were interpreted as parameter point estimates.

**Prior elicitation.** In addition to factor loadings $\lambda_{tf}$ of tasks $t = 1, \ldots, T = 15$ on factors $f = 1, \ldots, F = 3$, there were two other types of parameters: Task thresholds $\xi_{tc}$ representing $C - 1$ task difficulty parameters for up to three categories $C$ on tasks $T$, and one factor covariance $\sigma$ between the verbal and non-verbal knowledge factors. Prior distributions were chosen as follows:

$$\lambda_{tf} \sim N(0, 2.25)$$
$$\xi_{tc} \sim N(0, 4)$$
$$\sigma \sim U(-1, 1)$$

These priors were chosen based on prior literature describing priors for factor analysis with categorical data that possess favorable characteristics in terms of model convergence and parameter bias while being weakly informative (Zitzmann, Lüdtke, & Robitzsch, 2015; Zitzmann, Lüdtke, Robitzsch, & Marsh, 2016). The priors from these references were adapted to be slightly
more informative taking into account the nature and background knowledge concerning the present research design. For the factor loading parameters $\lambda_{tf}$, it was taken into account that the present tasks have shown in earlier frequentist analysis to be quite reliable (loadings well above .20) but none to be overly reliable (no loadings above .90). However, particularly in first grade, it was deemed uncertain whether such high loadings would be obtained, and whether all items would function properly. Priors for loadings were thus set as Gaussian with a location parameter of 0. This might not be the most adequate point estimate in higher grades where tasks were expected to be rather highly reliable. In higher grades however most samples were large so that rather weakly informative priors would likely be overturned by the data. Thus, this location parameter was kept across all grades, to also keep priors constant. The scale parameter (a priori variance estimate) was chosen to represent an expected standard deviation of 1.5. This implies that more than 95% of the prior density lies between factor loadings of -3.0 to +3.0. Models were estimated in standardized metric, thus loadings are usually not below -1.0 or above +1.0. Under this prior choice, the prior density is mildly decreasing within the presumed likely parameter area of about -.90 to +.90, making these factor loading priors potentially more informative than default weakly informative priors from the literature.

Similar rationales were applied in the prior setting for the threshold parameters $x_{tk}$. Threshold parameters usually lie within a range of -4 to +4, and this was reflected in choosing a location parameter of 0 and a scale parameter of 4, which yields a standard deviation of 2 and thus the assumption that these parameters quite certainly lie within a range of -4 and +4. In addition, the verbal knowledge tasks had two thresholds each; the higher threshold distinguishing between the second and third category on these tasks was the with a location parameter of 1 instead of 0, to reflect the prior assumption that these would be higher than the thresholds between the first and second category.

For the covariance between the two factors, $\sigma$, a uniform prior within -1 to +1 was chosen, which has two favorable characteristics. First, estimates are restricted within the admissible range for a correlation, suppressing inadmissible solutions (Hox & Maas, 2001; Zitzmann et al., 2016). Second, within these boundaries, this parameter is flat and thus uninformative, reflecting that particularly in first grade, it was uncertain in which range the intercorrelation of the factors might fall.
Prior robustness: Sensitivity analysis. Prior sensitivity was examined in two ways. First, the prior for the covariance parameter between the verbal knowledge and the non-verbal knowledge factors was changed from the uniform prior $\sigma \sim U(-1, 1)$ used in the main analysis into a Gaussian prior with location parameter 0 and scale parameter 10, $\sigma \sim N(0, 10)$. This prior choice induces slight decrease in prior density within the plausible range of -1 and 1 and should help examining whether the parameter was sensitive towards the uniform versus slightly Gaussian shape within the most likely parameter distribution area. Second, also the default priors in Mplus were examined. These are generally weakly informative, and supposedly slightly less informative than the main set of priors chosen for the present analysis. The default Mplus priors are however also based on simulation studies, technical and substantial considerations (Asparouhov & Muthén, 2010). The default Mplus priors were:

- $\lambda_{tf} \sim N(0, 5)$ for the factor loadings,
- $\xi_t \sim N(0, 5)$ for the 10 thresholds on the non-verbal knowledge tasks,
- $\xi_{tc} \sim N(0, inf)$ for the 10 thresholds on the verbal knowledge tasks,
- $\sigma \sim IW(0, 3)$ for the factor covariance.

The results under the different prior choices for the total sample model are depicted in Figures B1. Across the total sample, the prior choice did not induce any changes in parameter estimates: Under all three prior choices, chains mixed well, autocorrelation was low and posteriors unimodal. However, increasingly across school grades the different prior options revealed bimodality. Under the uniform and Gaussian priors, posteriors had highest density around .98, but also some density around .80. This is visible in the posterior distribution plots in sixth grade presented in Figure B2. Under the Mplus default priors, with the inverse Wishart prior for the factor covariance, the posteriors had most density distributed around .75, and such peak density around .99 is not visible.

An explanation for this influence of priors on the factor covariance posteriors is not apparent. However, the inverse Wishart, which is an inverse Gamma prior for this single parameter, has been criticized for various reasons (McNeish, 2016). First, because particularly in small samples it tends to be overly informative despite setting low degrees of freedom (representing high prior uncertainty Gelman et al., 2006), and second, because under a non-informative prior large standard deviations can be associated with large absolute correlations
Appendix B

(Barthelmé, 2012; Simpson, 2012). In the present case, in which the inverse Wishart is used for a single parameter, it is difficult to know whether these issues might explain the strange behavior of the inverse Wishart observed here. A theoretical explanation for this issue might be that the data are really underlain by a mixture of a distribution in which the location of the factor covariance is at about .75-.80, and another distribution in which the location is at about .98, representing subgroups of students for which the correlation differs. There might however also be various other purely statistical reasons. For the interpretation of the present results, it seems most plausible that the highest density is around .98, which was observed in accordance with both the uniform, and the Gaussian prior. The result from the uniform prior is used in the discussion of the present article, but the observed issue is also described in the discussion in order to make readers aware of its potential impact.

To examine whether autocorrelation might be related to this issue, the model in sixth grade was again estimated with the uniform prior, now with thinning increased to 100. Thinning can help reducing autocorrelation, however this does not necessarily increase posterior precision (Link & Eaton, 2012). The number of effective MCMC samples was also increased to 50000, to see whether this combination of thinning with an increase of posterior samples would change the posterior shape. In Figure B3, it can be seen that this measure indeed got rid of autocorrelation, however this did not lead to substantial changes in the posterior; the slightly bimodal density remains. Therefore, the density might not be an MCMC sampling problem.

Another observation was that under the normal prior and the Mplus default priors, in first grade parameter estimation did not converge for the factor covariance according to potential scale reduction, and visible in the relatively strong variance between chains (Figure B8). The present choice of priors thus seems to have supported convergence in the smaller and more messy sample in first grade, which under less informative priors was not achievable.

Finally, an apparent software error was observed: For many of the estimated models, the variable indicating the chain number of each MCMC draw was not exported. This error could be solved by removing manual seed setting (the BSEED option), after which the chain number variable was correctly exported, which allowed using the estimated for making the present figures in the R software package.

As an overall conclusion, the present choice of priors turned out to be not strongly
informative; apart from the mentioned observations, all posterior parameter point estimates stayed within small range across all prior specifications, and the overall posterior distributions also stayed quite similar. A posteriori, the present prior choice was thus deemed quite favorable. It should be noted at this point that it took some weeks of data simulations and literature consulting until the present prior choices were deemed ready for analysis.

**Figure B1.** Trace plots (left column panel), autocorrelation plots (middle column panel), and posterior density plots (right column panel, median indicated) of \( \sigma \) parameter under uniform prior (upper row), normal prior (middle row), default prior (lower row) for model in total sample. Autocorrelation and posterior plots depicted across all four chains.
Appendix B

Figure B2. Trace plots (left column panel), autocorrelation plots (middle column panel), and posterior density plots (right column panel, median indicated) of \( \sigma \) parameter under uniform prior (upper row), normal prior (middle row), default prior (lower row) for model in grade 6.

Figure B3. Autocorrelation plot (left), and posterior density plot (right, median indicated) of \( \sigma \) parameter under uniform prior grade 6, with thinning modified to 100 and 50000 effective draws.
Figure B4. Trace plots (left column panel), autocorrelation plots (middle column panel), and posterior density plots (right column panel, median indicated) of \( \sigma \) parameter under uniform prior (upper row), normal prior (middle row), default prior (lower row) for model in grade 5.
Figure B5. Trace plots (left column panel), autocorrelation plots (middle column panel), and posterior density plots (right column panel, median indicated) of $\sigma$ parameter under uniform prior (upper row), normal prior (middle row), default prior (lower row) for model in grade 4.
Figure B6. Trace plots (left column panel), autocorrelation plots (middle column panel), and posterior density plots (right column panel, median indicated) of \( \sigma \) parameter under uniform prior (upper row), normal prior (middle row), default prior (lower row) for model in grade 3.
Appendix B

Figure B7. Trace plots (left column panel), autocorrelation plots (middle column panel), and posterior density plots (right column panel, median indicated) of $\sigma$ parameter under uniform prior (upper row), normal prior (middle row), default prior (lower row) for model in grade 2.
Alternative models. Potential alternative factor analytic models were examined and their predictive fit compared with posterior predictive p-values \((ppp)\) (Gelman, 2013; Gelman, Carlin, Stern, & Rubin, 2014; Kaplan & Depaoli, 2013). Relative model fit indexes such as the deviance information criterion (DIC) were not available for models with categorical endogenous variables. The interpreted model estimated across the whole sample with one factor for the multiple choice tasks, one for the open answer tasks, and one for the method dependence between the five CVS tasks showed better predictive model fit \((ppp = .36)\) than three alternative models; one model without the method factor \((ppp = .00)\), another model assuming three correlated factors for the CHT tasks, the CVS tasks, and the open answer-tasks \((ppp = .10)\), and a model in which the
method factor instead of the non-verbal knowledge-factor was correlated with the verbal knowledge factor ($ppp = .00$).

References


Appendix C

Full Categorization Scheme and Examples for Handling of Unclear Cases

This Appendix contains, in order as follows:

Categorization scheme, German version

Coding of unclear cases, German version

Categorization scheme, English translation

Coding of unclear cases, English translation
<table>
<thead>
<tr>
<th>Kodierung</th>
<th>Bewertung</th>
<th>Bezeichnung</th>
<th>Beispiel</th>
</tr>
</thead>
<tbody>
<tr>
<td>91</td>
<td>0</td>
<td>Nicht kategorisierbare Begründung</td>
<td>„er soll einfach alle Bälle ausprobieren“, „weil man immer etwas verändern muss“</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>Falsche Bewertung (bei multiple choice-Antwort konfundiertes Experiment für gut oder unkonfundiertes Experiment für schlecht befunden)</td>
<td>„das interessiert mich nicht.“, „es spielt keine Rolle ob ein auto schneller als ein anderes fahren kann“, „weil ich eine ganz andere Antwort habe“, „weil ich noch nie so etwas gemacht habe“</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>Keine Begründung (Es wurde nichts geschrieben oder zu fragmentarisch um eine Aussage erkennen zu können)</td>
<td>„weil jedes Auto den gleichen Motor hat (FALSCH)“, „weil es nicht im Design des Experiments vorhanden war“</td>
</tr>
<tr>
<td>31</td>
<td>0</td>
<td>Irrelevante oder falsche Begründung (Bezug zu Design des Experiments, jedoch nicht zu konfundierten/unkonfundierten Eigenschaften, oder Begründung enthält Falschaussage)</td>
<td>„weil es nicht mit dem Material geht sondern um die Gestaltung des Experiments (FALSCH)“, „weil die schwere Kugel untergeht, ist es gefährlich“</td>
</tr>
<tr>
<td>32</td>
<td>0</td>
<td>Tautologische Begründung (Zirkulär bestätigende Aussage ohne Rational)</td>
<td>„weil es es ist“, „weil es gut nachvollziehbar ist“, „weil dann weiß man es direkt“, „weil es ein gutes experiment ist“, „weil es gut ist“</td>
</tr>
<tr>
<td>33</td>
<td>0</td>
<td>Ergebnis-basierte Begründung (Vermutung über Ausgang/Vermutung über Einfluss der Fokalvariable)</td>
<td>„ich denke es fahren beide gleich schnell“, „der stärkere Motor ist schneller“, „weil die Autos unterschiedlich sind und nicht nur unterschiedliche Motoren haben, weil alles unterschiedlich ist“</td>
</tr>
<tr>
<td>34</td>
<td>0</td>
<td>Korrekter Hinweis auf variierte Fokalvariable ohne Eingehen auf andere Variablen</td>
<td>„er soll einfach alle Bälle ausprobieren“, „weil man immer etwas verändern muss“</td>
</tr>
<tr>
<td>41</td>
<td>1</td>
<td>Korrekter Hinweis auf einen konfundierten Faktor, unkonfundierten Faktor</td>
<td>„weil es mit einer leichten und einer schweren Kugel unfair ist“, „weil die schwarze Kugel spesso weiter kommt“</td>
</tr>
<tr>
<td>42</td>
<td>1</td>
<td>Korrekter Hinweis auf einen zu verändernden konfundierten Faktor</td>
<td>„weil es bei Auto 1 mehr Luftwiderstand hat“</td>
</tr>
<tr>
<td>43</td>
<td>1</td>
<td>Korrekter allgemeiner Hinweis auf die Konfundierung oder erfolgte Kontrolle</td>
<td>„weil der Motor ist verschieden“, „weil das zu viele Unterschiede sind“, „weil die Autos unterschiedlich sind“</td>
</tr>
</tbody>
</table>

**Teilweise korrekte Begründung (zwischenvorstellung):**

<table>
<thead>
<tr>
<th>Kodierung</th>
<th>Bewertung</th>
<th>Bezeichnung</th>
<th>Beispiel</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>2</td>
<td>Korrekter Hinweis auf mehrere/alle konfundierte Faktoren oder unkonfundierte Faktoren (auch kombiniert mit Vorschlag, s. Beispiele)</td>
<td>„weil sie ganz unterschiedlich sind und nicht nur unterschiedliche Motoren haben“, „weil wir doch nur den Motor untersuchen, nicht die Räder und Form“</td>
</tr>
<tr>
<td>52</td>
<td>2</td>
<td>Korrekter Hinweis auf mehrere/alle zu verändernde Faktoren bei konfundierten Experimenten (Abgrenzung zu 51: Bei 52 nur Gleichhaltung konfundierter Faktoren erwähnt, ohne Variation der Fokalvariable)</td>
<td>„weil sie sonst alles gleich gebaut hat, weil nur die Räder unterschiedlich sind“, „weil alles gleich ist ausser dem Motor“</td>
</tr>
<tr>
<td>53</td>
<td>2</td>
<td>Korrekte allgemeine Erklärung eines unkonfundierten Experiments oder konfundierten Experimenten (Abgrenzung zu 52: Bei 52 nur Gleichhaltung konfundierter Faktoren erwähnt)</td>
<td>„weil die Oberfäche muss gleich sein und die Kugel gleich schwer wie die andere ist und die schwere Kugel weiter rollt als die leichte Kugel“, „weil beide Autos die gleichen Räder und die gleiche Form haben müssen“</td>
</tr>
<tr>
<td>54</td>
<td>2</td>
<td>Korrekter Vorschlag für unkonfundierte Durchführung des Experiments (Abgrenzung zu 52: Bei 52 nur Gleichhaltung konfundierter Faktoren erwähnt)</td>
<td>„weil bei einem guten experiment muss alles gleich sein und nur eine sache muss anders sein“, „es darf nur der Motor unterschiedlich sein“, „sie müsste 2 gleiche Autos nehmen aber unterschiedliche Motoren“, „man sollte nicht alles unterschiedlich machen, nur die Steigung.“</td>
</tr>
<tr>
<td>Kodierung unklarer Fälle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;etwas&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;etwas&quot; im Sinn von eindeutig &quot;eine Sache&quot; interpretierbar, wird kodiert als <strong>43</strong></td>
<td>&quot;weil 2 Sachen anders sind und für ein gutes Experiment darf nur etwas anders sein&quot;, &quot;Man darf nur etwas anders haben und sonst weiss man nicht, was es war.&quot;, &quot;man darf nur etwas anders manchen, sonst kann man es nicht gut vergleichen&quot;, &quot;man darf nur etwas verändern&quot;, &quot;weil nur etwas verschieden ist&quot;, &quot;weil nur etwas verändert ist&quot;, &quot;es wurde etwas verändert&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;etwas&quot; ohne, der eindeutige Sinn ersichtlich ist, wird kodiert als <strong>34</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>der allgemeine Hinweis ist eher unspezifisch als generisch, man weiss nicht genau, ob das Kind wirklich verstanden hat, warum es die Begründung verwendet</td>
<td>&quot;weil alles verschieden ist&quot;, &quot;weil alles gleich ist&quot;, &quot;weil viel zu viel unterschiedlich ist&quot;, &quot;weil zu viel verändert ist&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Der Hinweis hier ist spezifisch auf mehrere Variablen bezogen, der Bezug zur Variablenkontrolle ist eindeutig ersichtlich, man erkennt, dass das Kind die eigene Begründung auch verstanden hat</td>
<td>&quot;weil nur der Motor unterschiedlich ist&quot;, &quot;weil die Grösse und das Material unterschiedlich ist&quot;, &quot;weil die Steilheit und die Länge gleich ist&quot;, &quot;Weil nur das Material unterschiedlich ist&quot;, &quot;weil sie unterschiedlich gross sind&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Abgrenzung 43 - 51</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>43</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>51</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Abgrenzung 51 - 53</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>51</strong></td>
<td>&quot;weil sie zwei Autos gebaut hat, die genau gleich sind&quot;, &quot;weil sie sonst gleich sind&quot;, &quot;weil nur das Material unterschiedlich ist&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>53</strong></td>
<td>&quot;man darf nur etwas anders machen, sonst kann man es nicht gut vergleichen&quot;, &quot;weil es sind mehrere Sachen und man darf nur eine Sache verändern&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Abgrenzung 22 - 31</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>22</strong></td>
<td>&quot;weil man nicht immer zu schnell fahren kann&quot;, &quot;weil das Auto nach vorne zieht&quot;, &quot;weil man es wissen muss, wenn man 18 Jahre alt ist&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Abgrenzung 33 - 41</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vorwissen vs. erkennen einer Störvariable: tendenziell als Erkennen einer Störvariable interpretieren</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code</td>
<td>Rating</td>
<td>Category</td>
<td>Examples</td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>91</td>
<td>0</td>
<td>Not possible to categorize</td>
<td>Because always something should be changed (Rationale shows structure but does not fall in any of the categories)</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>Wrong choice</td>
<td>(Choice of multiple choice-answer judging confounded experiment as good or unconfounded experiment as bad)</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>No rationale</td>
<td>(Nothing written or too fragmentary to code reliably)</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>Rationale does not answer question</td>
<td>(No connection to experimental design)</td>
</tr>
<tr>
<td>31</td>
<td>0</td>
<td>Irrelevant or wrong rationale</td>
<td>Because each car has the same engine, &quot;because it is not about the size but about the material,&quot; when the heavy ball goes down, it is dangerous (Connection to experimental design, but not to confounded/unconfounded characteristics, or rationale includes erroneous statements)</td>
</tr>
<tr>
<td>32</td>
<td>0</td>
<td>Tautological reasoning</td>
<td>Because it is so, &quot;because it can be found out,&quot; because then it is directly known, &quot;because it is a good experiment,&quot; because like this Miss Smith cannot find out whether a car with a stronger engine, or with a weaker engine is faster (Circular confirming statement without rationale)</td>
</tr>
<tr>
<td>33</td>
<td>0</td>
<td>Outcome-based rationale</td>
<td>&quot;I think both go equally fast,&quot; &quot;the stronger engine is faster,&quot; &quot;the soft gumball can jump higher,&quot; &quot;it is known that the heavy ball does not jump,&quot; &quot;because at the higher ramo it goes farther&quot; (Assumption about outcome: assumption about influence of focal variable)</td>
</tr>
<tr>
<td>34</td>
<td>0</td>
<td>Correct hint towards the varied focal variable without mentioning other variables</td>
<td>Unconfounded experiment: &quot;because the engine is different,&quot; &quot;because car 1 has a strong engine and car 2 not,&quot; &quot;because then both is tried out&quot;</td>
</tr>
<tr>
<td>41</td>
<td>1</td>
<td>Correct hint towards confounded or unconfounded factor</td>
<td>Confounded experiment: &quot;because he took two different balls,&quot; &quot;because with a lighter and a heavier ball it is unfair&quot;</td>
</tr>
<tr>
<td>42</td>
<td>1</td>
<td>Correct hint towards factor that has to be changed</td>
<td>Confounded experiment: &quot;both should start from the top,&quot; &quot;he should take two balls of the same size,&quot; &quot;that's right with a steeper one but the balls have to be similar&quot;</td>
</tr>
<tr>
<td>43</td>
<td>1</td>
<td>Correct general hint towards confounding or correct controlling</td>
<td>Confounded experiment: &quot;because there are too many differences,&quot; &quot;that's the wrong combinations,&quot; &quot;because the balls are very different,&quot; &quot;because everything is different&quot;</td>
</tr>
<tr>
<td>51</td>
<td>2</td>
<td>Correct hint towards more/all confounded factors or unconfounded factors</td>
<td>Confounded experiment: &quot;because the balls and ramps differ,&quot; &quot;because they are all different and don't only have different engines,&quot; Unconfounded experiment: &quot;because everything else she built the same,&quot; &quot;because only the balls are different,&quot; &quot;because everything is the same apart from the engine,&quot; &quot;like this it only depends on the engine&quot; (also combined with suggestions, see examples)</td>
</tr>
<tr>
<td>52</td>
<td>2</td>
<td>Correct hint towards more/all adequate changes for confounded experiments</td>
<td>Confounded experiment: Because the surfaces have to be similar and the balls have to weigh the same because the heavier ball rolls farther than the light ball, &quot;because both cars must have the same wheels and shape&quot; (Difference to 54: At 52 only keeping equal of confounded variables is mentioned, not variation of the focal variable)</td>
</tr>
<tr>
<td>53</td>
<td>2</td>
<td>Correct general explanation of confounded or unconfounded experiments</td>
<td>Confounded experiment: &quot;only the engine may be different,&quot; Unconfounded or non-confounded experiment: &quot;because at a good experiment everything must be the same and only one thing must be different&quot;</td>
</tr>
<tr>
<td>54</td>
<td>2</td>
<td>Correct suggestion for unconfounded design for the experiment</td>
<td>Confounded experiment: &quot;only the engine may be different,&quot; &quot;she should take 2 similar cars but different engines,&quot; &quot;it must be two ramps with the same steepness and only the height must differ,&quot; &quot;one shouldn't make everything different, only the steepness&quot; (Difference to 52: At 54 also variation of focal variable is mentioned)</td>
</tr>
</tbody>
</table>
### Coding of unclear cases

**"something"**
- "because it's 2 things and for a good experiment only something may be different",
- "only something may be changed, otherwise one does not know what was the reason",
- "only something may be made different, otherwise it can't be compared",
- "only something may be changed",
- "because only one thing is different",
- "because only something is changed"

**"something" interpretable in the clear sense of "one thing"** is coded as 43
- "something was changed"

**"something" without clear sense how it is used** is coded as 34

**Distinguishing 43 - 51**
- the general hint is rather unspecific than generic; it is therefore not clearly visible whether the child really understood why it used the rationale
- "because everything is different", "because everything is similar",
- "because too much is different", "because too much has been changed"

**43**
- The hint here is specifically towards more variables; connection to variable control is clearly visible, it is obvious that the child has understood the own rationale
- "because only the engine is different", "because the size and the material are different",
- "because the steepness and the length are the same", "because only the material differs",
- "because they are differently big"

**Distinguishing 51 - 53**
- 51
- Here, a hint towards more variables is given that are specifically mentioned
- "because she built two cars that are exactly the same", "because else they are the same",
- "because only the material differs"

- 53
- Here, these variables or the whole experiment are explained in a generic way, but it must be obvious that the child has understood its own rationale and does not just provide a suggestion how the experiment might be done
- "only something may be made different, otherwise one cannot compare it",
- "because there are various things and only one thing may be changed"

**Distinguishing 22 - 31**
- 22
- Here, the child clearly does not answer the question but still provides a rationale
- "because one may not always drive too fast", "because the car pulls towards the front",
- "because one should know it when they are 18 years old"

**Distinguishing 33 - 41**

Prior knowledge vs. recognizing confounding variable:
- Usually coded as recognizing of confounding variable
Appendix D

Complementary Descriptive and Inferential Statistics for Latent Transition Analysis
### Table D1

Transition probabilities between respective profiles.

<table>
<thead>
<tr>
<th></th>
<th>CVS + 1.5SD</th>
<th>CVS - 1.5SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pre</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMC (.47)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>LMC (.16)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>HMC (.36)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>FRA (.01)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>SCI (.14)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

| **Post** |             |             |
| MMC (.00) | 0.00 | 36 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| LMC (.12) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| HMC (.01) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FRA (.16) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCI (.26) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

**Note.**
- MMC = moderate misconceptions profile;
- LMC = low misconceptions profile;
- HMC = high misconceptions profile;
- FRA = fragmented profile;
- PRE = prescientific profile;
- INT = intermediate profile;
- SCI = scientific profile.

Numbers in brackets next to profiles indicate profile size before (Pre) and after (Post) instruction. All other numbers indicate transition probabilities between respective profiles.
Table D2

*Estimates of covariances of the study variables.*

<table>
<thead>
<tr>
<th></th>
<th>gra</th>
<th>gen</th>
<th>CVS</th>
<th>mc_pr</th>
<th>mc_po</th>
<th>ic_pr</th>
<th>ic_po</th>
<th>sc_pr</th>
<th>sc_po</th>
</tr>
</thead>
<tbody>
<tr>
<td>gra</td>
<td>1.509</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gen</td>
<td>0.006</td>
<td>na</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp</td>
<td>0.036</td>
<td>-0.001</td>
<td>0.041</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mc_pr</td>
<td>-1.615</td>
<td>-0.237</td>
<td>-0.163</td>
<td>31.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mc_po</td>
<td>-1.702</td>
<td>-0.128</td>
<td>-0.276</td>
<td>14.024</td>
<td>38.917</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ec_pr</td>
<td>0.335</td>
<td>0.11</td>
<td>0.092</td>
<td>-0.173</td>
<td>-2.336</td>
<td>6.312</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ec-po</td>
<td>-0.026</td>
<td>0.053</td>
<td>0.076</td>
<td>0.767</td>
<td>-1.1</td>
<td>1.928</td>
<td>5.731</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sc_pr</td>
<td>-0.191</td>
<td>0.052</td>
<td>0.046</td>
<td>7.488</td>
<td>2.308</td>
<td>2.928</td>
<td>1.52</td>
<td>11.151</td>
<td></td>
</tr>
<tr>
<td>sc-po</td>
<td>0.925</td>
<td>0.037</td>
<td>0.151</td>
<td>0.354</td>
<td>-2.462</td>
<td>2.944</td>
<td>3.281</td>
<td>4.457</td>
<td>17.275</td>
</tr>
</tbody>
</table>

**Note.** *gra* = grade; *gen* = gender; *CVS* = CVS questionnaire mean score; *mc_pr* = misconceptions score before instruction; *mc_po* = misconceptions score after instruction; *ic_pr* = intermediate conceptions score before instruction; *ic_po* = intermediate conceptions score after instruction; *sc_pr* = scientific concepts score before instruction; *sc_po* = scientific concepts score after instruction. All estimates represent full information maximum likelihood covariance estimates. Covariances of variables with themselves are variances.
Appendix D

Table D3

Estimates of means, standard deviations, and intercorrelations of the study variables.

<table>
<thead>
<tr>
<th></th>
<th>sc-pr</th>
<th>sc-po</th>
<th>ic-pr</th>
<th>ic-po</th>
<th>sc-pr</th>
<th>sc-po</th>
<th>mc-pr</th>
<th>mc-po</th>
<th>CVS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>gra</strong></td>
<td>3.58</td>
<td>1.23</td>
<td>0.15</td>
<td>-0.01</td>
<td>0.47</td>
<td>0.20</td>
<td>-0.23</td>
<td>-0.08</td>
<td>-0.14</td>
</tr>
<tr>
<td><strong>gen</strong></td>
<td>0.01</td>
<td>na</td>
<td>-0.01</td>
<td>-0.15</td>
<td>6.75</td>
<td>2.51</td>
<td>-0.05</td>
<td>0.04</td>
<td>0.15</td>
</tr>
<tr>
<td>CVSm</td>
<td>0.19</td>
<td>0.16</td>
<td>0.40</td>
<td>5.61</td>
<td>6.24</td>
<td>0.32</td>
<td>0.35</td>
<td>3.14</td>
<td></td>
</tr>
<tr>
<td>mc_pr</td>
<td>0.11</td>
<td>0.09</td>
<td>0.18</td>
<td>0.07</td>
<td>0.11</td>
<td>0.35</td>
<td>0.19</td>
<td>6.79</td>
<td>2.23</td>
</tr>
<tr>
<td>mc_po</td>
<td>0.18</td>
<td>0.22</td>
<td>0.18</td>
<td>0.02</td>
<td>0.11</td>
<td>0.28</td>
<td>0.33</td>
<td>0.32</td>
<td>4.76</td>
</tr>
<tr>
<td>ic_pr</td>
<td>-0.05</td>
<td>0.03</td>
<td>0.07</td>
<td>0.40</td>
<td>0.11</td>
<td>0.35</td>
<td>0.19</td>
<td>6.79</td>
<td>3.34</td>
</tr>
<tr>
<td>ic_po</td>
<td>-0.05</td>
<td>0.04</td>
<td>0.15</td>
<td>0.06</td>
<td>0.07</td>
<td>0.32</td>
<td>0.83</td>
<td>8.83</td>
<td>2.39</td>
</tr>
</tbody>
</table>

Note. **gra** = grade; **gen** = gender; **CVS** = CVS questionnaire mean score; **mc_pr** = misconceptions score before instruction; **mc_po** = misconceptions score after instruction; **ic_pr** = intermediate conceptions score before instruction; **ic_po** = intermediate conceptions score after instruction; **sc_pr** = scientific concepts score before instruction; **sc_po** = scientific concepts score after instruction. All estimates represent Pearson correlation coefficient estimates, those including gender point-biserial correlation estimates.

Table D3
Appendix E

Technical Information on Bayesian Latent Change Score Model

In the latent change score model, no knowledge profiles are estimated, instead this is a traditional regression-based model in which variance in students’ knowledge on the three scores after the instruction was predicted by their knowledge before instruction, and additionally by their CVS score (Figure E2). The latent change score model was estimated using Bayesian estimation with Markov Chain Monte Carlo (MCMC) sampling. This estimation method can be more effective than maximum likelihood estimation and overcomes interpretational issues of other estimation methods and of p-value hypothesis testing (Edelsbrunner, 2014; Etz, Gronau, Dablander, Edelsbrunner, & Baribault, 2017; Wagenmakers, 2007; Wagenmakers, Morey, & Lee, 2016). We used default priors in Mplus and MCMC with Gibbs sampling with four chains, each with 10000 draws, 5000 samples burn-in, and a thinning factor of 5. We did not have substantial prior information concerning the predictive strength of CVS measures for conceptual change in regression models. The following priors were applied in Mplus:

\[ N(0, \infty) \] for intercept and regression weight parameters, reflecting large uncertainty about where these parameters might fall (and thus not restricting extreme values).

\[ IW(0, -4) \] for variance and covariance parameters (an improper prior because \( v < p - 1 \), as well reflecting maximum uncertainty).

These priors seemed appropriate because the large dataset in combination with the relatively simple autoregressive, panel regression-like model made it very likely that the data would completely overturn these priors.

There were no convergence problems, potential scale reduction factors were equal or smaller 1.020 for all estimated parameters (Brooks & Gelman, 1998). Trace plots of the main parameters, the regression of the three knowledge indicator change scores on the CVS score, are depicted in Figure E1. The estimated standardized model parameters are presented in Figure E2.
Figure E1. Trace plots (left column panel), autocorrelation plots (middle column panel), and posterior density plots (right column panel, median indicated) from the latent change score-model to predict students’ prior knowledge and knowledge change from their understanding of the control of variables-strategy. The plots stem from the MCMC draws of the three unstandardized parameters for the regression of students’ change on the misconceptions score (upper panel), intermediate conceptions score (middle panel), and scientific concepts score (lower panel) on their CVS score.
Figure E2. The latent change score-model to predict students’ knowledge change from their understanding of the CVS. CVS = CVS questionnaire mean score; C_mc = change on misconceptions score from before mc_pr to after mc_po instruction; C_ic = change on intermediate conceptions score from before (ic_pr) to after (ic_po) instruction; C_sc = change on scientific concepts score from before (sc_pr) to after (sc_po) instruction. Double-headed arrows indicate variance or correlation parameters, single-headed arrows regression parameters. Main parameter estimates are the predictive paths in bold from CVS to C_mc, C_ic, and C_sc. Change score variables represent intercepts and residual variances after instruction, controlling for before instruction values, like in a generic panel regression model. 0s and 1s are fixed parameters, the other values represent standardized Bayesian parameter estimates.
Appendix E

References


Peter Adriaan Edelsbrunner

pety@shitmail.org

Birth date: 15 January 1987
Origin: Graz, Austria.
 orcID: 0000-0001-9102-1090
Hobbies: Running, Skateboarding, Metal
Memberships: EARLI SIG3,
Friends of the Stanford Encyclopedia of Philosophy,
John Dewey Society

SCIENTIFIC INTERESTS
Mixture modeling in educational and developmental research
Scientific thinking throughout the lifespan
Complexity in concept learning
Statistical and psychometric practices in educational and developmental research

EDUCATION
2012 - 2017 Doctoral studies in Learning and Instruction: ETH Zurich, Switzerland.
2007 - 2012 Diploma studies (MSc) in Psychology: University of Graz, Austria.
2011 Semester Abroad: University of Padova, Italy.
2011 Internship: School Psychological Service Leibnitz, Austria.
2009 Internship: Cognitive Science Section, University of Graz, Austria.
2005 High-school diploma: Federal Gymnasium Kirchengasse, Graz, Austria.

WORK AND VOLUNTEERING
Since 2016 Solinetz Zürich.
Since 2012 Research Assistant and Lecturer: ETH Zurich, Switzerland.
2011 - 2012 Student teaching assistant: University of Graz, Department of Psychology.
2009 Social- and learning-caretaker: Association of Foster-Parents Graz, Austria.
2008 – 2010 Freshmen’s Tutor: Department of Psychology, University of Graz.

TEACHING
2015 - 2017 Educational Science: Assessment and Support of Knowledge Development in the Classroom. Seminar at ETH Zurich.
2011 - 2012 Research Methods in Developmental Psychology, Health Psychology, Differential Psychology. Seminars at University of Graz; teaching assistant.
ACADEMIC SERVICES
Since 2016 Associate editor: Journal of European Psychology Students.
Since 2015 Coordinating Board: EARLI Special Interest Group on Conceptual Change.
Since 2014 President: Association of Scientific Staff, ETH Zurich.
2015 - 2016 Director: Peer Mentoring Group for Methods and Statistics, University of Zurich, Switzerland.
2015 - 2016 Working Group: Good Scientific Practice, ETH Zurich.
2015 Appointment committee: Professor of Learning and Technology with a Focus on Higher STEM Education, ETH Zurich.
2011 – 2012 Executive Board: European Federation of Psychology Students’ Associations.

EVENT ORGANIZATION
2018 2nd PhD Summer School on Conceptual Change: Conceptual Change and Epistemic Cognition, Klagenfurt, Austria.
2016 1st PhD Summer School & 10th Conference on Conceptual Change: Conceptual Change meets Other Disciplines, Florina, Greece.
2014 8th EFPSA European Summer School: The Psychology of Work: Research on Occupation & Organisations, Mellau, Austria.

WRITINGS