
**Scientific Reasoning in Preschool:
The Development, Correlates, and Promotion
of Control of Variables Strategy Abilities**

April Christine Moeller Bachhuber



**12. Dezember 2019
München**

Scientific Reasoning in Preschool: The Development, Correlates, and Promotion of Control of Variables Strategy Abilities

Dissertation zum Erwerb des Doctor of Philosophy (Ph.D.)

am Munich Center of the Learning Sciences

der Ludwig-Maximilians-Universität

München



Vorgelegt von

April Christine Moeller Bachhuber

München, 12.12.2019

Erstgutachterin: Prof. Dr. Beate Sodian

Zweitgutachter: Prof. Dr. Heinrich Hußmann

Tag der mündlichen Prüfung: 29.01.2020

Abstract

Scientific reasoning is a critical skill for participating in society and shaping its future. However, scientific reasoning is also difficult and, despite a recent push to promote these skills in education, children and adults still struggle. At the same time, research has shown that very young children and infants have precocious abilities in causal reasoning. For example, they can use covariation information to make causal inferences. These abilities are likely potential precursors for later scientific reasoning abilities. A distinction, however, between causal and scientific reasoning abilities is the metaconceptual understanding of theory and evidence and the metacognitive ability to intentionally seek knowledge, for example through experimentation. So, although young children show sensitivity to the informativeness of evidence, it is unclear if they intentionally act to test hypotheses. The present work investigates the development, structure, and correlates of scientific reasoning in three- to six-year-old preschool children using a knowledge-lean task and also investigates the potential to promote these abilities with digital training tools. In Chapter 1, we reviewed the causal reasoning and scientific reasoning literature. In Chapter 2, we found that preschoolers have a beginning metacognitive understanding of their own ignorance as well as beginning abilities in recognizing and selecting a controlled test of a hypothesis. Older children, in particular, were more successful in providing verbal justifications than younger children. In Chapter 3, we found that scientific reasoning was related to and predicted by inhibition and Theory of Mind and that Theory of Mind seems to be an important prerequisite for developing scientific reasoning abilities. Finally, in Chapter 4, we found potential for training experimentation abilities in preschool children using a tablet application and a video tutorial. The findings of this thesis suggest that preschoolers have early abilities in scientific reasoning and that there is potential to further promote these abilities in early childhood education.

Extended Summary

The ability to reason scientifically is critical not only throughout one's education but also for active participation in modern society (Bromme & Goldman, 2014). However, mature scientific reasoning is difficult, and both children and adults struggle with many aspects of reasoning scientifically, for example, distinguishing between hypotheses and evidence or designing controlled experiments (Kuhn, Amsel, & O'Loughlin, 1988). For these reasons, researchers, educators, and organizations place great emphasis on teaching the skills involved in scientific reasoning throughout education, usually beginning in middle-elementary school (UNESCO, 2014). In contrast to the struggles children and adults have with reasoning scientifically, very young children, and even infants, show precocious causal reasoning abilities (Muentener & Bonawitz, 2018). For example, children can use covariation evidence to make causal inferences (Gopnik, Sobel, Schulz, & Glymour, 2001), they are sensitive to the informativeness of evidence (Cook, Goodman, & Schulz, 2011), and they can intervene on causal systems to gain information (Gweon & Schulz, 2008). Many of these causal reasoning abilities seem related to scientific reasoning abilities and may be possible precursors to the development of scientific reasoning.

However, scientific reasoning requires a metaconceptual understanding of the distinction between theory and evidence in a way that allows children to recognize that a theory can be tested and revised or that evidence can support or undermine a hypothesis (Kuhn, 1989, 2002; Kuhn & Franklin, 2007). Further, it requires the ability to generate hypotheses and then generate evidence to test those hypotheses. Thus, though children may spontaneously isolate or even control variables in causal reasoning assessments (Cook et al., 2011; van Schijndel, Visser, van Bers, & Raijmakers, 2015), it is unclear if they are intentionally seeking knowledge through these behaviors, for example, by

generating and testing particular hypotheses during exploration (Kuhn, 2002). All of these abilities require reflection and thus rely on metacognitive abilities, such as the understanding of one's own knowledge and ignorance, which are developing around five years of age (Bullock, Sodian, & Koerber, 2009; Perner, 1991; Rohwer, Kloo, & Perner, 2012; Sodian & Bullock, 2008; Wimmer & Perner, 1983).

Children's and adults' abilities in scientific reasoning are affected by a number of factors, such as their prior beliefs or prior knowledge about task content, the outcomes of experiments, or the level of difficulty of the tasks (e.g., Kuhn et al., 1988; Tschirgi, 1980). It is possible that studies using such tasks underestimate children's abilities and, instead, highlight their struggles with difficult tasks. This could provide an explanation for the discrepancy between "precocious" causal reasoning ability, which is typically measured through knowledge-lean or decontextualized tasks, and deficient scientific reasoning, which is more commonly measured with tasks using scientific or every day content about which children may have prior beliefs or knowledge and which can be quite complex in terms of task design or requirements to manipulate variables.

To investigate the possibility that children have greater scientific reasoning abilities than shown previously, we developed novel knowledge-lean tasks using the blicket detector paradigm (Gopnik & Sobel, 2000) to limit the influence of prior knowledge or beliefs on preschool children's (three- to six-year-olds) abilities in scientific reasoning. Specifically, we assessed their ability to recognize when evidence is confounded and that, as a result, they cannot know something conclusively, as well as their ability to recognize a controlled test of a hypothesis. Critically, and to distinguish from causal reasoning, we included a task that required children to reflect on their own ignorance resulting from confounded evidence and tasks that required children to specifically test a hypothesis by selecting a controlled test. In this way, we could take

advantage of and build upon children's precocious causal reasoning abilities but target the scientific reasoning abilities that are not typically assessed with knowledge-lean paradigms.

The present work investigated the development, structure, and correlates of scientific reasoning in three- to six-year-old preschool children using these knowledge-lean tasks and also investigated the potential to promote these abilities with digital training tools. In Study 1, we investigated the stability of children's scientific reasoning abilities using the novel knowledge-lean tasks and found that while children's spontaneous responses improve from one session to a second session two weeks later, their ability to provide verbal justifications for their responses was relatively stable. Further, we found that preschoolers have a beginning metacognitive understanding of their own ignorance as well as beginning abilities in recognizing and selecting a controlled test of a hypothesis and that older children, in particular, were more successful in providing verbal justifications.

The wording of the questions used in Study 1 did not emphasize the scientific process of testing a hypothesis, for example, simply asking children to select a choice. In Studies 2a and 2b, we improved the wording to the questions to instead place an emphasis on selecting a test to find out about the hypothesis and also to provide children with clear options for interpreting the outcome of the experiment. In these studies, we performed a robust assessment of children's abilities with two trials of each task in one session. We again found evidence of preschoolers' metacognitive understanding of their own ignorance and ability to recognize and select a controlled test of a hypothesis. Older children in these samples were also more successful in providing justifications, and we found that older children were also better in the task assessing their state of knowledge. In Study 3, we found that adults' initial responses to these same knowledge-lean tasks were

very successful, but that they were limited in their ability to provide robust, elaborated explanations or justifications for their initial responses and often provided justifications similar to those provided by preschoolers. Further, adults were more successful in interpreting the outcome of their experiment than in explaining their choice of experimental design.

In Study 4, we found that scientific reasoning was related to and predicted by inhibition and Theory of Mind. Further, Theory of Mind seems to be an important prerequisite for developing scientific reasoning abilities as few children who were not successful in Theory of Mind were successful in scientific reasoning. In Studies 5a and 5b, we used an iterative design process to develop a tablet application training tool. We successfully reduced usability issues in subsequent iterations, and we found potential for training experimentation abilities in preschool children using direct instruction. In Study 6, we developed a video tutorial for teaching experimentation and investigated the effect of animation in the tutorial on children's scientific reasoning. With this study, we found beneficial effects of animation and potential for promoting scientific reasoning using direct instruction with a video tutorial.

Taken together, the findings of this thesis suggest that preschoolers have early abilities in scientific reasoning when measured with a developmentally appropriate, knowledge-lean task, and that these abilities likely build upon existing causal reasoning abilities. In addition, scientific reasoning abilities are related to the ability to inhibit unwanted responses, for example, choosing the incorrect test of a hypothesis. They are further related to Theory of Mind, for example by having an awareness of one's own state of knowledge or recognizing that different tests of a hypothesis will provide evidence and what that evidence means for the hypothesis. Theory of Mind abilities appear to be a necessary precursor for developing scientific reasoning abilities. Finally, the findings of

this thesis suggest that there is potential for promoting scientific reasoning abilities in early childhood.

Acknowledgments

The research presented here was supported by the Elite Network of Bavaria [project number: K-GS-2012-209] and the German Research Council (Grant SO 213/34-1). I would like to extend my sincere gratitude to the opportunities made possible by the ENB, including conference attendance and an extended research stay at Brown University.

Prof. Dr. Beate Sodian, thank you for your support and guidance throughout the past years. You have helped me to become a better researcher and a more critical thinker and have given me space and encouragement to pursue my research interests. Prof. Dr. Heinrich Hußmann, thank you for welcoming me into the Media Informatics department and allowing me to learn through the supervision of students. I appreciate your support in bringing some interdisciplinarity into my project. Prof. Dr. Markus Paulus, thank you for your insightful feedback in the colloquium and for being available to me throughout the years. Prof. Dr. Dave Sobel, thank you for enthusiastically agreeing to be my international supervisor, for your support in data analysis, writing, and conference preparation, and for welcoming me into the Causality and Mind Lab at Brown University. PD. Dr. Tobias Schuwerk, it was a pleasure to work with you for my master thesis and I appreciate your continued support during the pursuit of my Ph.D. Prof. Dr. Frank Fischer, thank you for the opportunity to be a member of the REASON program and for your support and feedback in courses and in the bi-monthly meetings.

It has been a wonderful experience to be a member of the REASON 2016 cohort; thank you to everyone for contributing to that experience! I would like to especially thank my office mates and friends, Despoina Georgiou and Sarah Bichler, for creating an environment that was productive and enjoyable, and Arianne Herrera-Bennett for always being available for all types of questions. I would also like to thank Irina Ciobanu for the coordination of the program and for always being available for what probably felt like a

never-ending stream of questions. I would like to thank Christopher Osterhaus for the support through the classes you taught, the feedback on writing and data analysis, and for being a great new officemate in the final stages of writing this thesis. Thank you again to Sarah Bichler for the invaluable feedback on this thesis and for encouraging me to collaborate on interesting projects with you.

Özgün Köksal-Tuncer and Andrea Saffran, you have both been incredibly supportive throughout the past years in welcoming me to the department and to the team and in helping me think through every aspect of my research from theory to study design, to data analysis and interpretation. It has been a pleasure to work with you both! Thank you also to Maria Fysaraki for lending a listening ear to the struggles of an interdisciplinary project and for our talks of the future. Thank you to Shalaila Haas for the feedback and support you've provided over the many past years.

Thank you to my family, Mom, Dad, Andrew, and Matthew, for your love, encouragement, and continued interest in my work, even when it was difficult to understand. Thank you to my second family, Karin, Konrad, and Franziska, for your love and support, home-cooked meals, and spa-destination weekend getaways. Thank you to Riley for snuggles and stress-relief. Johannes, thank you for being an amazing partner, husband, and friend. Thank you for being there to listen, to give feedback, to review my writing, and to brighten my days.

To my family.

Table of Contents

Abstract	v
Extended Summary	vii
Acknowledgments	xii
List of Tables	xxii
List of Figures	xxiii
1 Review of the Causal Reasoning and Scientific Reasoning Literature.....	27
1.1 Introduction and Overview of the Thesis.....	27
1.2 Causal Reasoning	31
1.2.1 Categorizing events and objects by causality.....	35
1.2.2 Understanding internal parts as causal.....	38
1.2.3 Using covariation evidence to make accurate causal inferences	39
1.2.4 Reasoning diagnostically about potential uncertain causes	43
1.2.5 Inferring causal relations from evidence.....	44
1.2.6 Intervening on causal systems to disambiguate evidence.....	46
1.2.7 Generating and interpreting causal explanations	50
1.2.8 Similarity of causal reasoning to the process of scientific discovery	53
1.3 Scientific Reasoning	56
1.3.1 The development of Scientific Reasoning	58
1.3.2 Evidence of preadolescent children’s limited Scientific Reasoning abilities	66
1.3.3 Factors influencing children’s Scientific Reasoning and CVS abilities	78
1.3.4 Evidence of preadolescent children’s beginning Scientific Reasoning abilities.....	84
1.4 Chapter Summary.....	100
1.5 The Present Studies.....	103

2	Preschoolers' (and Adults') Scientific Reasoning.....	107
2.1	Study 1: Stability of Preschoolers' Scientific Reasoning Abilities	112
2.1.1	Method.....	112
2.1.2	Results.....	117
2.1.3	Discussion.....	123
2.2	Study 2a: Preschoolers' Scientific Reasoning Abilities.....	126
2.2.1	Method.....	126
2.2.2	Results.....	129
2.2.3	Discussion.....	133
2.3	Study 2b: Replication & Extension of Study 2a	135
2.3.1	Method.....	135
2.3.2	Results.....	137
2.3.3	Discussion.....	143
2.4	Interim Discussion	145
2.5	Study 3: Adults' Scientific Reasoning Abilities	150
2.5.1	Method.....	153
2.5.2	Results.....	157
2.5.3	Discussion.....	162
2.6	General Discussion	165
3	The Structure and Correlates of Scientific Reasoning in Preschool.....	169
3.1	Introduction	169
3.1.1	Metacognition and Theory of Mind.....	169
3.1.2	Executive functioning.....	170
3.1.3	The structure of Scientific Reasoning abilities	172
3.1.4	Correlates of Scientific Reasoning	173
3.1.5	Summary.....	178

3.2	Study 4: The Structure and Correlates of Scientific Reasoning in Preschool	180
	
3.2.1	Method	180
3.2.2	Results	188
3.2.3	Discussion	198
4	The Development of Educational Tools for Assessing and Promoting Control of Variables Strategy Abilities in Preschool	205
	
4.1	Introduction	205
	
4.1.1	Promoting Control of Variables Strategy abilities	208
4.1.2	Children, media, & technology	214
4.1.3	Instructional design	216
4.1.4	Educational products	220
4.1.5	Designing for preschoolers	221
4.1.6	Design and evaluation criteria.....	222
4.1.7	Evaluating products for children with children.....	226
4.1.8	Summary	229
4.2	The Present Studies	230
	
4.3	Study 5a: The Development of a Tablet Application for Training CVS	231
	
4.3.1	Statement of Collaboration	231
4.3.2	Introduction	232
4.3.3	Method	232
4.3.4	Results	242
4.3.5	Discussion	256
4.4	Study 5b: Continued Iterative Development of a Tablet Application for Training CVS	263
	
4.4.1	Statement of Collaboration	263
4.4.2	Introduction	263
4.4.3	Method	265

4.4.4	Results.....	270
4.4.5	Discussion.....	278
4.5	Study 6: The Development of a Video Tutorial for Training CVS.....	282
4.5.1	Statement of Collaboration.....	282
4.5.2	Introduction.....	283
4.5.3	Method.....	285
4.5.4	Results.....	289
4.5.5	Discussion.....	295
4.6	General Discussion & Recommendations	296
4.6.1	Recommendations.....	300
5	General Discussion	303
5.1	Summary of results	304
5.1.1	Study 1: Stability of preschoolers' scientific reasoning abilities.....	305
5.1.2	Studies 2a & 2b: Preschoolers' scientific reasoning abilities.....	307
5.1.3	Study 3: Adults' scientific reasoning abilities	310
5.1.4	Study 4: Structure and correlates of scientific reasoning in preschool.....	312
5.1.5	Studies 5a, 5b, & 6: Promotion of scientific reasoning with digital training tools..	314
5.2	Discussion	317
5.2.1	Preschooler's knowledge-lean scientific reasoning.....	317
5.2.2	Structure and correlates of scientific reasoning.....	322
5.2.3	Adults' scientific reasoning abilities.....	326
5.2.4	Developing training tools for preschoolers.....	331
5.2.5	Assessing and promoting CVS in preschoolers with digital tools.....	335
5.3	Implications.....	337
5.3.1	Theoretical implications	337
5.3.2	Practical implications.....	339
5.4	Future Directions.....	339

5.5 Conclusion.....	343
References.....	345
Appendix.....	383

List of Tables

Table 2.1. Examples of children’s explanations for the ICE task in Study 2.1	118
Table 2.2. Examples of children’s justifications for their test selection in Study 2.1	119
Table 2.3. Examples of children’s interpretations of the experiments in Study 2.3.	137
Table 2.4. Examples of adults’ explanations for the ICE task in Study 2.4	155
Table 2.5. Examples of adults’ justifications for their test selection in Study 2.4	156
Table 2.6. Examples of adults’ interpretations of the experiments in Study 2.4	157
Table 2.7. Descriptives for variables of interest in Study 2.4	158
Table 2.8. Correlations among variables of interest in Study 2.4	159
Table 3.1. Descriptives for variables of interest in Study 3.	193
Table 3.2. Correlations among variables of interest in Study 3	195
Table 3.3. Regression analysis predicting ICE in Study 3	197
Table 3.4. Regression analysis predicting CVS in Study 3	197
Table 3.5. Regression analysis predicting scientific reasoning in Study 3	197
Table 3.6. Cross tabulation of mastery of scientific reasoning and ToM in Study 3	197
Table 4.1. Frequency of usability issues in first iteration of Study 4.1	243
Table 4.2. Frequency of usability issues in second iteration of Study 4.1	247
Table 4.3. Correlations among variables of interest in Study 4.1	254
Table 4.4. Duration of task phases for each interface and iteration in Study 4.2	271
Table 4.5. Frequency of usability issues in first and second iteration of Study 4.2	275

List of Figures

Figure 2.1. Materials and an example of the testing procedure for Study 2.1 113

Figure 2.2. Children’s performance on the ICE task in Study 2.1..... 120

Figure 2.3. Children’s performance on the CVS tasks in Study 2.1 121

Figure 2.4. Materials and an example of the testing procedure for Study 2.2 127

Figure 2.5. Children’s performance on the ICE task in Study 2.2..... 130

Figure 2.6. Children’s performance on the CVS tasks in Study 2.2..... 131

Figure 2.7. Children’s performance on the ICE task in Study 2.3..... 138

Figure 2.8. Children’s performance on the CVS tasks in Study 2.3..... 140

Figure 2.9. Children’s interpretations of the experiment outcome in Study 2.3 142

Figure 3.1. Children’s performance on the ICE task in Study 3 189

Figure 3.2. Children’s performance on the CVS tasks in Study 3..... 190

Figure 3.3. Children’s performance on the CVS tasks in Study 3 by gender and trial..... 191

Figure 4.1. Warm up materials for Study 4.1 233

Figure 4.2. Storybook and application for Study 4.1 234

Figure 4.3. Procedure of the task phases in Study 4.1 235

Figure 4.4. Training phase for Study 4.1 236

Figure 4.5. Initial state of CVS task in the storybook and application in Study 4.1 237

Figure 4.6. Interaction for variable selection in Study 4.1 239

Figure 4.7. Frequency of usability issues in Study 4.1 251

Figure 4.8. Frequency of behavioral issues in Study 4.1 251

Figure 4.9. Prototype of near transfer task for Study 4.1 262

Figure 4.10. Character design for Study 4.2..... 266

Figure 4.11. Interface design for Study 4.2 266

Figures

Figure 4.12. Procedure of the task phases in Study 4.2.....	268
Figure 4.13. Screenshots of important frames in the video tutorial in Study 4.3	286
Figure 4.14. Procedure of the task phases in Study 4.3.....	288
Figure 4.15. Children’s attention throughout tutorial in Study 4.3	291
Figure 4.16. Children’s reactions throughout tutorial in Study 4.3.....	292

1 Review of the Causal Reasoning and Scientific Reasoning Literature

1.1 Introduction and Overview of the Thesis

To participate in the global knowledge society, citizens must be well-informed and also possess 21st-century skills (Dede, 2010; Gilbert, 2005; Partnership for 21st Century Skills, 2009; Prenzel, Rost, Senkbeil, Häußler, & Klopp, 2001). Twenty-first-century skills include (among many others) identifying and asking significant questions, analyzing and evaluating evidence and claims, interpreting information and drawing conclusions, and reflecting critically on this process. These skills might sound rather scientific, and indeed, they are many of the same required for engaging in scientific reasoning and argumentation (SRA): problem identification, questioning, hypothesis generation, construction of artifacts, evidence generation, evidence evaluation, drawing conclusions, and communicating and scrutinizing scientific reasoning and its results (Fischer et al., 2014). But even though they sound scientific, scientific reasoning skills are both relevant to and critical for everyone who wants to actively participate in society and its future. Being able to reason scientifically allows citizens to investigate, evaluate, and understand scientific topics relevant to society, such as climate change or vaccination (Bromme & Goldman, 2014; Trilling & Fadel, 2009; Zimmerman & Klahr, 2018). Consequently, there has recently been a push in research, education, and by national and international organizations to focus on promoting SRA (e.g., Chinn, Buckland, & Samarapungavan, 2011; Duschl, 2008; Iordanou, 2016; Kuhn & Crowell, 2011; Sandoval, Sodian, Koerber, & Wong, 2014; American Association for the Advancement of Science [AAAS], 2009; European Commission, 2015; National Research Council [NRC], 2010, 2012; Federal Trade Commission [FTC], 2011, 2015; Next Generation Science Standards [NGSS], 2013; United Nations Educational, Scientific, and Cultural Organisation, 2005/2014). Some

researchers even argue that SRA abilities are as important as the traditional basic skills of reading, writing, and arithmetic (Tolmie, Ghazali, & Morris, 2016) and, as such, they should receive as much attention starting from early childhood and throughout education.

Early theories of child development suggested, however, that young children are not capable of complex scientific reasoning. Complex reasoning abilities were thought to only develop in adolescence in what Piaget termed the “formal operations” stage of development (Inhelder & Piaget, 1958). In line with Piaget’s theory, children’s scientific reasoning abilities have been described as limited, in part because children struggle to differentiate between hypotheses and evidence (Kuhn, 2002; Kuhn & Franklin, 2007). This metaconceptual understanding of the difference between theory and evidence is critical to scientific reasoning, which is defined as the intentional search for knowledge (Kuhn, 2002) by generating, testing, and revising hypotheses and includes the ability to reflect on this process of knowledge acquisition and change (Morris, Croker, Masnick, & Zimmerman, 2012; Wilkening & Sodian, 2005). Kuhn and colleagues (1988) have shown that children do not systematically test hypotheses and often try to produce an effect rather than determine its cause. They also fail to control variables and ignore or distort evidence that does not support their prior beliefs (Kuhn, Amsel, & O’Loughlin, 1988). But more recent research suggests that young children may have much greater scientific reasoning abilities than previously thought (Zimmerman, 2007). Children in early elementary school can differentiate between hypotheses and evidence by selecting a conclusive test of a simple hypothesis (Sodian, Zaitchik, & Carey, 1991), and in recognizing and selecting a controlled experiment (Bullock & Ziegler, 1999). Even preschoolers have been shown to produce controlled tests when they are provided with support and feedback (van der Graaf, Segers, & Verhoeven, 2015).

Indeed, research from an up-until-recently separate literature on causal inference has shown that young children show precocious causal reasoning abilities. For example, they use covariation evidence (an event followed by an outcome) to make accurate causal inferences (Gopnik, Sobel, Schulz, & Glymour, 2001; Schulz & Gopnik, 2004) and they infer causal relations according to evidence even when those relations conflict with their prior beliefs (Kushnir & Gopnik, 2007). The ability of young children to evaluate evidence, learn from covariation data, and infer causal relations seems likely related to scientific reasoning abilities.

Scientific reasoning abilities are frequently investigated through one of its key components: the Control of Variables Strategy (CVS). CVS is a method for designing unconfounded experiments from which valid causal inferences can be made (Chen & Klahr, 1999). It is the understanding that to determine the effect of a variable, one must manipulate the variable in question while keeping all other variables constant. Similarly to the literature on scientific reasoning abilities in general, the literature on CVS abilities has shown limited abilities in both younger and older children, and even adults (Bullock & Ziegler, 1999; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Schauble, 1996), but also that even preschoolers have some understanding of the strategy when provided with feedback and support (van der Graaf et al., 2015). In fact, a large proportion of the CVS literature has focused on promoting these abilities through intervention studies (Schwichow, Croker, Zimmerman, Höffler, & Härtig, 2016). Training has been shown to be effective across age groups and can be performed both with physical hands-on tasks as well as virtual tasks (Schwichow et al., 2016). In addition, experimentation skills are related to science learning, for example, the ability to design controlled experiments using CVS at age 11 has been shown to be predictive of later achievement in science courses at age 14 (Bryant, Nunes, Hillier, Gilroy, & Barros, 2013).

In light of the importance of scientific reasoning skills and the discrepancy between early causal reasoning and limited scientific reasoning, we sought to investigate the development of scientific reasoning abilities in preschool children, the relation between scientific reasoning abilities and other cognitive factors, and the potential for promoting scientific reasoning abilities in early childhood. Those aims are reflected in the organization of this thesis. In Chapter 1, we aimed to bring together the literature on causal reasoning and scientific reasoning in children. We conducted a brief review of the causal reasoning literature, the literature on scientific reasoning, and specifically on the development of an understanding of the Control of Variables Strategy. We further examined aspects of task design that can influence both the assessment and promotion of CVS abilities.

In Chapter 2, we had the goal of investigating preschool children's and adults' abilities in scientific reasoning, specifically, abilities in understanding and using the Control of Variables Strategy. To this end, we developed a novel, knowledge-lean task to assess preschoolers' understanding of confounded evidence, their ability to recognize and select a controlled test, their ability to verbalize their reasoning behind the selection of a test, and their ability to interpret the outcome of experiments. We used this task to assess the stability of the above abilities in a test-retest study with preschoolers (Study 1). Based on the findings of Study 1 and observations of preschoolers' experiences with the task, we made some adaptations to the task and performed a more robust assessment of preschooler's abilities in scientific reasoning and CVS (Studies 2a & 2b). At the same time, we used this knowledge-lean task to investigate adults' abilities in scientific reasoning and CVS (Study 3), both to validate the use of the task as a measure of CVS ability and to investigate how adults perform on a task that is not influenced by any prior knowledge or beliefs about the task content.

In Chapter 3, we sought to investigate the relation between scientific reasoning and other cognitive abilities. We reviewed the literature on the structure and correlates of scientific reasoning abilities in children. We further examined how scientific reasoning abilities in preschool relate to other developing cognitive abilities by investigating the relation between four-year-olds' abilities in CVS, as measured by the novel CVS task from Chapter 2, and their executive functioning and Theory of Mind (Study 4).

In Chapter 4, we aimed to design tools for promoting CVS abilities in preschoolers. We reviewed the literature on the promotion of CVS abilities, as well as design and usability factors for assessment tasks. We iteratively designed and developed an interactive tablet application for assessing and promoting CVS abilities (Studies 5a & 5b). We also created a video tutorial for explicitly instructing children in CVS (Study 6) and assessed its effectiveness in improving children's abilities on the novel CVS task from Chapter 2. We discuss design and usability factors specific to these materials that could influence the assessment and promotion of CVS.

Finally, in Chapter 5, we summarize and discuss the studies presented in this thesis and their theoretical implications for the causal and scientific reasoning literatures, as well as their practical implications for early childhood education.

1.2 Causal Reasoning

Understanding causality, the relation between cause and effect, is a critical skill for navigating a world of uncertainty. Perhaps because it is so important, it appears to develop very early, such that even very young children have precocious causal reasoning capacities. Indeed, the growing literature on early causal reasoning suggests that preschoolers are “sophisticated” and “intuitive” causal reasoners (Muentener & Bonawitz, 2018, p. 43). This first section presents a summary of the causal reasoning literature and outlines several ways in which children's early reasoning abilities may lay the groundwork

for later, more complex, scientific reasoning abilities. Causal reasoning has been shown to play a central role in early learning across a number of different domains. In the physical domain, infants show understanding of spatiotemporal relations: for example, they understand events such as objects being occluded or hidden by another object and can use those events to predict outcomes (Baillargeon, 2004). They also show an understanding of objects in motion and collisions between objects (Baillargeon, 2002; Carey, 2009; Cohen, Amsel, Redford, & Casasola, 1998; Spelke, 1990). In the biological domain, for example, young children understand that animals grow, while inanimate objects do not (Rosengren, Gelman, Kalish, & McCormick, 1991), and in the psychological domain, young children recognize beliefs and desires as causal mechanisms and use them to explain the behaviors of others (Gopnik & Wellman, 1992). Three mechanisms have been proposed to explain the development of early causal reasoning: caused-motion interactions, agents' goal-directed actions, and covariation information (Muentener & Bonawitz, 2018).

The first of these, *caused-motion interactions*, is the transfer of physical force between objects. For example, if one observes an object, A, approach a second object, B, and upon contact, object B begins to move, one could conclude that object A caused object B to move (scenario 1). Research has shown that, like adults, infants are sensitive to the spatial and temporal features of this interaction. For example, if A approaches B, but stops before contacting B, the subsequent motion of B could not be due to A (scenario 2), or if A contacts B, but B does not begin to move until after a delay, A is not thought to have caused B (scenario 3) (Cohen et al., 1998; Cohen & Oakes, 1993; Leslie & Keeble, 1987; Mascialzoni, Regolin, Vallortigara, & Simion, 2013; Newman, Choi, Wynn, & Scholl, 2008). Most studies investigating causal reasoning with infants use looking-time paradigms, in which infants are shown some stimuli until they become habituated to them. Then they are shown new stimuli and their looking behavior is measured. Longer looking

times to certain new stimuli are interpreted as events in the new stimuli being surprising or unexpected (Leslie & Keeble, 1987; Saxe & Carey, 2006; Spelke, Breinlinger, Macomber, & Jacobson, 1992). For example, infants who were habituated to a causal event like scenario 1, looked longer at the noncausal events (scenarios 2 and 3), suggesting that those events were surprising or unexpected (Cohen & Amsel, 1998). Thus, infants as young as six months show an understanding of the causal nature of motion events.

The second proposed mechanism for the emergence of causal reasoning, *representations of agents' goal-directed actions*, suggests that young children recognize when they or others (or other non-human agents) take actions to perform causal events to reach a goal. For example, scenario 3, as described in the above paragraph, in which object A contacts object B but object B does not immediately move, is surprising when children view object B as an inert object. In that case, object B should have no reason not to move. But in a scenario in which object B was previously a self-propelling object, moving around on its own, it does not surprise children that object B does not move after object A contacts it. In this case, object B was capable of moving itself and thus also capable of resisting the force of object A (Luo, Kaufman, & Baillargeon, 2009). Such studies suggest that the spatial and temporal features described above cannot be solely responsible for the emergence of early causal reasoning since infants' understanding of agents and their actions also influences their causal reasoning.

The third mechanism, *interpreting covariation information*, suggests that young children engage in causal reasoning as a result of tracking covariation relations between events in their environment. For example, Sobel and Kirkham (2006) investigated infants' causal inferences by showing eight-month-old children a sequence of events that led to a music event; specifically, A and B together predict C. In one condition, they observed A by itself followed by a second different music event D, suggesting that A by itself does not

predict C. In the second condition, children observed that A by itself was followed by C, suggesting that A predicts C, but that B could also predict C. Then, in both conditions, they saw B by itself followed by a blank screen and the music that accompanied events C and D. In the first condition, children should look more at where event C occurs, because they expect B to predict C. In the second condition, B may or may not predict C, so infants may look equally to the locations where events C and D occur. The children did indeed look more to the C event location than the D event location in the first condition and looked more to the C event location in the first condition than in the second condition. The results of this study suggest that infants can track statistical information and use that information to predict dependent events.

However, simply tracking covariation information also does not seem to be solely responsible for the emergence of causal reasoning abilities, because even when events are presented with equally predictive relations, children only represent the events initiated by an agent as causal. For example, eight-month-old infants observed either a human hand or a toy train approach a box that is partially hidden behind an occluder (Muentener & Carey, 2010). The box would then break apart into pieces. The infants did not see the event that led to the box being broken, just the events of the hand or train approaching the box and then the box being broken. In test trials, the occluder was removed and the infants saw the hand or train either make contact with the box or stop just before, leaving a gap. Then, they either saw the box break apart or not break apart.

If infants thought the agent was causal, then they should be surprised (and look longer) at the event in which the agent contacted the box but it did not break apart, or at the event in which the agent did not make contact with the box but it did break apart. Critically, the covariation information the infants observed at the beginning, an agent approaching the box and the box falling apart, should be interpreted as causal if children

are only making use of the information that event A precedes (and predicts) event B. But this was not the case. The infants were indeed surprised by the unexpected events of contact-but-not-broken and no-contact-but-broken when the human hand was the agent, but not when the toy train was the agent (Muentener & Carey, 2010). These results suggest that children show a bias towards reasoning about agents and their actions over the spatial and temporal features of causal motion events or covariation information.

The following sections describe a number of studies on young children's causal reasoning abilities in different areas, showing that children use causal information to categorize events and objects (Gopnik & Sobel, 2000; Nazzi & Gopnik, 2003; Schulz, Standing, & Bonawitz, 2008), diagnose if objects have hidden causal properties, such as internal parts (Sobel, Yoachim, Gopnik, Meltzoff, & Blumenthal, 2007), reason about counterfactual events (Harris, German, & Mills, 1996), register conditional independence among events and use covariation information to make accurate inferences (Gopnik et al., 2001; Schulz & Gopnik, 2004; Sobel & Kirkham, 2006), appreciate the ambiguity of confounded evidence (Cook, Goodman, & Schulz, 2011; Schulz & Bonawitz, 2007; Sodian et al., 1991), and can intervene on causal systems (Cook et al., 2011; Gopnik et al., 2001; Gweon & Schulz, 2008).

1.2.1 Categorizing events and objects by causality

A common task paradigm used in causal reasoning research with preschoolers is the blicket detector paradigm (Gopnik & Sobel, 2000). Blickets are typically novel objects, such as blocks, boxes, or bricks, with particular properties that "cause" a blicket detector machine to work (light up or play music). The blicket detector is often secretly controlled by the experimenter.

Gopnik and Sobel (2000) developed and used the blicket detector paradigm to investigate young children's understanding of causality and their ability to categorize

causal objects. In a categorization condition, three- to four-year-old children observed objects placed on the blicket detector and the resulting effect, either turning the box on or not. The experimenter then identified one of the objects that had turned the box on as a blicket and asked children to identify the other blicket. In other words, children have seen objects make the box light up and then one of those objects was identified as a blicket. In an induction condition, the experimenter held up two objects and identified them as blickets, then held up the remaining two objects and identified them as not blickets. The experimenter then placed one blicket on the box (which turned on) and asked children to identify the other blicket. In this case, children first see objects labeled as blickets and then see that a blicket makes the box light up.

Children completed seven trials, four “neutral” trials, in which there was no relation between the causal properties and the perceptual features of the objects, and three “conflict” trials, in which the perceptual features conflicted with the causal properties, e.g., of two identical objects, only one of them had causal properties. In the categorization condition, children correctly identified the object with the same causal powers more often for the neutral tasks (74%) than for the conflict (40%) tasks. In the induction condition, children correctly identified the object with the same causal powers based on the common name equally well in both task types (82% in neutral and 73% in conflict tasks). These results suggest that a conflict between causal and perceptual features makes it more difficult for children to categorize a causal object as causal. In other words, the perceptual features are still very salient, and children seem to be swayed to pick the perceptually identical object. This bias was not an issue in the induction tasks; children could reliably identify objects as causal based on the name the object was given.

In a second experiment, Gopnik and Sobel (2000) used the same procedure but did not allow the objects to touch the box, thus introducing a spatial gap and removing the

causal property from the objects. In this case, children did not use the association between the object and the box being turned on to identify the other blicket. They chose at random in the neutral tasks and chose the perceptually similar object in the conflict tasks. The results of this experiment suggest that children in Experiment 1 did, in fact, identify objects as causal based on the contact between the object and the box and that their decreased performance on the conflict categorization tasks is not simply due to confusion or chance responding.

In a third experiment, Gopnik and Sobel (2000) investigated even younger children's causal reasoning abilities. In the categorization condition, 2½-year-olds showed a pattern of performance similar to that of older children, identifying the causal object as the blicket more often in the neutral tasks (55%) than in the conflict tasks (31%). However, they correctly identified the blicket less often than four-year-olds in the neutral tasks in the categorization condition. In the induction condition, the younger children were less likely to correctly identify the blicket and more likely to select perceptually similar objects. In the association condition, younger children performed no differently than older children. With these three experiments, Gopnik and Sobel (2000) showed that very young children can use causal information to both name and categorize objects and make inductive inferences about causal properties on the basis of object names and that this ability is likely developing in these early years from two to four years of age.

Similarly to Gopnik and Sobel (2000), Nazzi and Gopnik (2003) showed that infants can use causal information to categorize objects, rather than relying solely on perceptual cues, such as color, shape, or parts (Imai, Gentner, & Uchida, 1994). However, Nazzi and Gopnik wanted to investigate these abilities without using names or labels for the objects, to control for the influence of language. In several studies, they showed that 2½-year-olds could sort objects based on their causal properties, similarly to the first

experiment of Gopnik and Sobel (2000), and they do not categorize objects when there is a temporal but non-causal association, similarly to the second experiment of Gopnik and Sobel.

In a third set of experiments, Schulz and colleagues (2008) also investigated children's understanding of causality and object categorization. They found that three- to four-year-olds were more likely to explore objects when they observed evidence that conflicted with their expectations based on inductive generalizations from an object's categorization to its causal properties. For example, when children were shown objects that stuck to a board and then received more objects, they explored more when those new objects had the same name but did not stick to the board than when they had a different name and did not stick to the board.

These studies show that young children are sensitive to causal information and use this information to make sense of the world, by categorizing objects on the basis of their causal properties.

1.2.2 Understanding internal parts as causal

The next section discusses children's understanding of how objects are causal; for example, what are the properties that give objects their causal power and can those properties be hidden or internal. Sobel and colleagues (2007) presented three- to four-year-old children with three objects, which each had holes drilled into their center to hold internal parts. Externally, two objects were identical (A & B) and the third was different (C). Internally, one of the identical objects (A) and the different object (C) contained the same internal part, while B was empty. Children then saw that A activated a blicket detector and were asked to choose which other object (B or C) would also activate the machine. The choice was between the object that was externally identical but contained no internal parts (B) and the object that was externally different but contained the same

internal parts (C). Sobel and colleagues (2007) found that children infer that an object's internal parts are related to its causal properties. Children who observe an object with particular internal parts (A) activate a blicket detector select a different-looking object with the same internal parts (C) to activate the blicket detector, and they prefer the internal parts over the external features as the causal mechanism (B) (Sobel et al., 2007). They can diagnose whether objects have hidden (internal) features based on the objects' causal properties and can do this over the lure of perceptual similarity. These results also suggest that children have a preference for causal mechanisms they consider to be plausible, for example, the internal parts over the superficial external properties.

1.2.3 Using covariation evidence to make accurate causal inferences

The next section outlines ways in which young children use observed covariation evidence to make causal inferences. Gopnik and colleagues (2001) presented three- and four-year-old children with two types of tasks: one-cause or two-cause tasks. In the one-cause tasks, the first object, A, activated the blicket detector. The second object, B, did not activate the detector. When both objects were placed on the detector at the same time, the detector activated. The simultaneous presentation was repeated a second time. Thus, children observed the following pattern: A✓, B✗, A+B✓, A+B✓. The experimenter then asked children whether each object, individually, was a blicket or not. In this condition, it was possible that children correctly identified the blicket simply because it was the object that activated the detector more often. Thus, a control two-cause condition was used. In the two-cause tasks, the first object, A, activated the blicket detector. This was repeated three times. The second object, B, did not activate the detector the first time it was placed on it but did activate the detector the following two times. Thus, children observed the following pattern: A✓, A✓, A✓, B✗, B✓, B✓. Again, the experimenter asked the children whether each object was a blicket.

In both conditions, the frequency of activations of the objects was the same: children observed object A activate the detector three times and object B activate the detector twice. In this way, the two-cause condition controls for the possibility that children simply identify the blicket as the one that activated the detector more often. Indeed, children identified object A as a blicket more often than object B in the one-cause task and identified both A and B as blickets in the two-cause task. In a similar experiment, Gopnik and colleagues (2001) also showed that even younger children (30-month-olds) also made these causal inferences. These results suggest that even very young children are sensitive to patterns of dependent and independent probability and use this information to draw accurate causal inferences about what is or is not causal.

Another study investigated 19- and 24-month-olds' causal inferences with a similar procedure to that described above (Sobel & Kirkham, 2006). In a first trial, children observed that object A activated the machine by itself and object B did not activate the machine by itself. They then observed that objects A and B activated the machine together (A✓, B✗, A+B✓; screening off condition). The children were then presented with both objects and, instead of having to identify which object was a blicket, children were told to make the machine go. Almost three-quarters of the children placed object A on the box by itself, with similar performance for the older and younger children, replicating the findings of Gopnik and colleagues (2001) described above, in even younger children.

In a second trial, children observed objects A and B activate the machine together. They then saw that object B by itself did not activate the machine (A+B✓, B✗; indirect screening off condition). The children were then presented with both objects and told to make the machine go. Three-quarters of the older children placed object A on the box, while the younger children performed at chance level.

In a third trial, children observed objects A and C activate the machine together. They then saw that object C activated the machine by itself (A+C✓, C✓; backwards-blocking condition). However, children were then given object A and a third object that had not been placed on the box, B, and were asked to make the machine go.

Object A was associated with making the box light up equally in both the indirect screening off and the backwards-blocking condition, thus there should be no difference in children's selection of object A between the two trials if they are reasoning only based on the associative power. However, overall, children's performance was not different from chance on this trial. Older children selected object A more often in the indirect screening off condition than in the backwards-blocking condition, however, the younger children did not perform differently in these two trials. These results suggest that the 24-month-olds were perhaps using similar causal reasoning mechanisms as the older children (30-month-olds and three- to four-year-olds) in other studies, but that there were developmental differences, such that the 19-month-old infants were not able to succeed in these causal reasoning tasks. However, the authors discussed some limitations, such as the need to inhibit a response that imitated what they had seen the experimenter do and instead complete a novel action. Thus, children may have been reasoning correctly, but failing in the behavioral action required to correctly show that reasoning.

To further investigate this possibility, Sobel and Kirkham (2006) adapted the indirect screening off and backwards-blocking conditions to a looking-time paradigm using eye-tracking methods to avoid the need for a behavioral response and investigated even younger infants' causal inferences. This study was described in the introduction to discuss children's ability to interpret covariation information as one mechanism for the development of causal reasoning. We will describe it again here as it relates to the previous studies in this section. Eight-month-old infants observed a sequence of events

that revealed that A and B together predict C ($A+B \rightarrow C$). In one condition, they observed A by itself followed by a second event D ($A \rightarrow D$; indirect screening off), suggesting that A by itself does not predict C. In the second condition, children observed that A by itself was followed by C ($A \rightarrow C$; backwards blocking), suggesting that A predicts C, but that B could also predict C. Then, in both conditions, they saw B by itself followed by a blank screen and the music that accompanied events C and D.

In the indirect screening off condition, children should look more at where event C occurs, because they expect B to predict C. In the backwards-blocking condition, B may or may not predict C, so infants may look equally to the locations where events C and D occur. The children did indeed look more to the C event location than the D event location in the indirect screening off condition and looked more to the C event location in the indirect screening off condition than in the backwards-blocking condition. The results of these studies suggest that very young children and infants can recognize conditional dependencies between events, for example, determining when events are dependent or independent based on a third event.

Finally, in a series of experiments similar to Gopnik and colleagues' (2001), Schulz and Gopnik (2004) investigated three- to four-year-old children's abilities to make causal inferences using patterns of covariance across biological and psychological domains, as opposed to the physical domain typical of blicket detector tasks. Their findings revealed that preschoolers can also learn the causal structure of biological events (monkey sneezing at particular flowers) and psychological events (bunny being scared of particular animals) and did so consistently.

In summary, this section presented studies that revealed that young children and infants are capable of using covariation evidence to make accurate causal inferences even under circumstances of uncertainty and across different domains.

1.2.4 Reasoning diagnostically about potential uncertain causes

The following section discusses a study that investigated children's ability to reason when there were a number of potential causes, though children could not be certain which were or were not efficacious. Sobel and colleagues (2017) investigated young children's (three- to seven-year-olds) ability to reason about potential causes with unknown efficacy. Using a blinket detector paradigm and four blocks, children observed the effects (light and music) of either all four (all-known), only three (1 unknown), or only two of the blocks (2 unknown). After observing the effects, the box was occluded and then activated which resulted in music playing. Children were asked to select which object had made the box activate. Regardless of their selection, children were told they were incorrect and to pick again, twice. Children's error-free performance was assessed, which meant not choosing the block that they had seen not make the box light up. In other words, picking blocks of unknown efficacy over a block that they knew did not make the box light up.

Four-year-olds performed at chance in all conditions, five-year-olds performed at chance in the unknown conditions, and the two older groups (six- and seven-year-olds) performed above chance in all conditions. These results showed that, around the age of five, children can correctly recall the objects which they had seen have an effect on the box, however, they struggled to diagnose the causes of the effect when they were uncertain about the efficacy of some of the objects. The ability to select objects of uncertain efficacy over objects known to not have an effect seems to be developing between five and six years of age, as the two groups of older children could engage in causal reasoning to do so.

In a second study, Sobel and colleagues (2017) investigated children's ability to diagnose causes when the effect produced was a result of additive causes. They showed children (five- to eight-year-olds) that four blocks placed on the box made it light up green

and play music ($A+B+C+D \rightarrow \text{green, music}$). They were then shown that three blocks, A, B, and C, made the box light up red; blocks A, B, and D also made the box light up red; and block A did not activate the box ($A+B+C \rightarrow \text{red}$; $A+B+D \rightarrow \text{red}$; $A \times$). This information indicated that block C and block D individually make the box light up red and when combined, they make the box light up green and play music ($C \rightarrow \text{red}$; $D \rightarrow \text{red}$; $C+D \rightarrow \text{green, music}$). The box was then occluded and activated such that music played. Children were asked to choose which set of blocks had just made the box play music: B+C, B+D, or C+D. The correct choice is C+D because this combination makes the box light up green and play music. The younger children (five- to six-year-olds) performed no differently than expected due to chance and the older children (seven- to eight-year-olds) performed better than expected due to chance. Sobel and colleagues (2017) concluded that between the ages of six and seven, children are developing diagnostic reasoning abilities about additive effects. These studies show that some aspects of causal reasoning, for example, uncertainty and additive effects, are more difficult and are developing later in early childhood.

1.2.5 Inferring causal relations from evidence

The vast majority of experiments with blinket detectors rely on the principle of spatial contiguity, such that contact between an object and the detector is how an effect occurs, likely because this principle often occurs in real life. However, there are also many cases in which this principle is violated, such as flipping a light switch and having the light turn on. There is no contact between the switch and the light, yet there is a causal relation.

Kushnir and Gopnik (2007) investigated three- to four-year-olds' understanding of causality and causal relations with and without contact. They used a blinket detector paradigm and in some cases, an object would activate the machine when it came into

contact with it, while in other cases, an object would activate the machine simply by hovering over the machine. Initially, children preferred a contact hypothesis, that the object would have to come in contact with the machine to activate it, however, when presented with statistical evidence showing that the machine activated when objects did not touch it, children were able to learn this relation and use this information to perform informative interventions. Kushnir and Gopnik concluded, first, that children can use probability to make sophisticated causal inferences, and second, that they can overcome prior beliefs to do so. The findings of this experiment support the findings of much earlier research on young children's understanding of the role of spatial contiguity in causal relations (Cohen et al., 1998; Cohen & Oakes, 1993; Leslie & Keeble, 1987; Mascialzoni et al., 2013; Newman et al., 2008).

In another investigation of children's preference for spatial contiguity, Schultz (1982) taught children the mechanism of a tuning fork: that they can cause a box to make a ringing noise by holding a tuning fork in front of the box's opening without touching the box. Then Schultz showed children a box that was ringing, with one fork touching the top of the box and another fork held in front of the box and asked the children to identify the cause of the sound. Children aged two to four years could correctly identify the fork held in front of the opening as the cause, preferring the mechanism they had previously seen demonstrated over a spatial contiguity mechanism. In a different scenario, children observed the beam from a flashlight shining on a wall from a distance, as well as a second flashlight placed on the light spot of the first flashlight on the wall and pointing away from the wall. In such a scenario, five-year-olds could correctly identify the first (non-spatially-contiguous) flashlight as the cause of the light, but three-year-olds claimed that the second flashlight on the wall was the source of the light on the wall. These results suggest that young children have a preference for a spatial contiguity mechanism for causal relations,

but that this bias can be overcome and the ability to do so without having previously observed the causal relation is developing between three and five years of age.

Preschoolers may even be better at learning and generalizing causal relations than adults, perhaps because they are less biased by prior knowledge or beliefs about common causal relations. Lucas and colleagues (2014) showed that four- and five-year-olds were better able than adults to learn a less common conjunctive causal relation (e.g., $A \times, B \times, A+B \checkmark$) and to design interventions to prevent or produce an effect based on that relation. Thus, these results suggest that children can learn causal relations, even less common ones, on the basis of evidence. The authors concluded that children seem to pay more attention to current evidence than do adults and adults may not learn less-common relations as easily as do children because they are influenced by prior knowledge and beliefs about more common causal relations (Lucas, Bridgers, Griffiths, & Gopnik, 2014).

1.2.6 Intervening on causal systems to disambiguate evidence

The following section describes a number of studies in which children use their knowledge of causal properties and their ability to reason about causal systems to perform novel interventions on those systems. Gopnik and colleagues (2001) presented three- and four-year-old children with two types of tasks: one-cause or two-cause tasks. In the one-cause tasks, the first object, B, was placed on the detector and did not activate it. It was then removed. Object A was placed on the detector and did activate it. With Object A still on the detector, which was still activated, Object B was again placed on the detector. Thus, children observed the following pattern: $B \times, A \checkmark, A+B \checkmark$. Children were then asked to make the machine stop. In the two-cause condition, the first object, B, was placed on the detector and activated it. It was then removed. Object A was placed on the detector and activated it. With Object A still on the detector, which was still activated, Object B was again placed on the detector. Thus, children observed the following pattern: $B \checkmark,$

$A\checkmark$, $A+B\checkmark$. Again, children were asked to make the machine stop. In the one-cause condition, children selectively removed the causal object, A. In the two-cause condition, children removed both object A and B simultaneously. This experiment shows that children are capable of causal reasoning and use that information to perform appropriate interventions on causal systems. Similar results were found by Sobel, Tenenbaum, and Gopnik (2004).

Cook, Goodman, and Schulz (2011), also used the blicket detector paradigm to investigate young children's (four- to five-year-olds) interventions on causal systems. In this study, one group of children were shown that four beads (All Beads) placed individually on the blicket detector made it light up and play music. The second group of children was shown that two out of the four beads (Some Beads) made the blicket detector light up and play music. Then, each group was shown that two beads stuck together and placed on the box made it light up and play music. The experimenter said, "Wow, look at that. I wonder what makes the machine go. Go ahead and play," and then walked away and out of children's line of sight. Children were allowed to play freely with the beads while experimenters observed if they spontaneously attempted to isolate the variables to determine which of the two beads make the box work. The children in the All Beads group did not isolate and test the beads individually. Half of the children in the Some Beads group did isolate and test the beads individually.

In a second experiment in which the bead pairs were glued together and could not be separated and pulled apart, again about half of the children in the Some Beads condition performed an informative intervention by rotating the pair vertically, to touch just one end of the bead-pair to the box at a time. This finding suggests that children not only process covariation evidence to learn about causal relations, but that they also have a beginning understanding of the need to isolate variables, an important step in valid

experimentation processes, and can perform novel, informative interventions to learn about causal systems (Cook et al., 2011).

In a study with four- to five-year-olds, Gweon and Schulz (2008) showed that children distinguish between confounded and unconfounded evidence and that their exploratory play reflects this distinction. In a confounded condition, children observed that a blue block placed on the closer, black side of a mat made a red box light up, and a yellow block placed on the farther, white side of a mat made a green box light up. Thus, children could not know if it was the block (blue or yellow) or the color of the mat (black or white), which resulted in the effect of the red and green boxes lighting up. In the unconfounded condition, children observed the blue block placed first on the black side and then on the white side and saw that the red box lit up in both cases. They also saw the yellow block placed on both sides and that the green box lit up in both cases. Thus, children could know that the color of the mat was not relevant, but the block (blue or yellow) was important to whether the red or green box lit up. In the unconfounded condition, children spent more time playing on the closer, more convenient side of the mat, and more children preferentially played with that side, suggesting that when there was no information to be gained, it did not make sense to put in extra effort to play on the farther side of the mat. In the confounded condition, almost half of the children performed informative interventions which controlled variables, by placing each block on each side of the mat separately.

Schulz and Bonawitz (2007) claimed that exploratory play is affected by whether observed evidence is confounded or not and expected that children will isolate relevant variables in exploratory play to generate evidence to support causal learning. Children (four- to five-year-olds) played together with the experimenter with a box with two levers. When a lever was pressed, a puppet popped out of the box. In a confounded condition, the

levers were pressed simultaneously, such that the causes of the puppets popping up were confounded. In an unconfounded condition, the levers were pressed one at a time, so that the causes were unconfounded. After this, children were given a choice between the box with which they had just interacted and a novel box. The researchers observed play time and which toy children reached toward first. Children explored the familiar toy more in the confounded condition than in the unconfounded condition and three-quarters of children manipulated the levers separately, disambiguating the evidence. Schulz and Bonawitz concluded that children's exploratory play is sensitive to confounded evidence and that children are motivated to explore stimuli where the causal structure is ambiguous and perform interventions to disambiguate the causal structure.

Interestingly, children prefer evidence from their own interventions over evidence from the interventions of others. Kushnir and Gopnik (2018) investigated if children (four-year-olds) differentiate between the evidence generated from others' and their own interventions. Using the blicket detector paradigm, Kushnir and Gopnik manipulated whether children only observed a sequence of events on the detector or if they intervened themselves on the last two events of the sequence. Specifically, children observed that block A made the box light up twice and block B did not make the box light up twice. Then the children either observed that block A did not make the box light up and then block B did make the box light up, or they placed the blocks on the detector themselves and observed those same effects (A✓, A✓, BX, BX -- AX, B✓). Based on this evidence, block A had a probability of 2/3 to make the box light up and block B had a probability of 1/3. Children were asked to pick the best block to make the machine go. When children observed this whole procedure, they selected block A (81%), the block with the higher probability of making the box light up, but when they intervened themselves, they were more likely to select block B (66%). Even though block B was

associated with a lower probability of activating the box, children themselves had used it to activate the box and preferred this evidence over the evidence they had previously observed. However, when this same intervention condition was used, but at the critical moment when the child placed block B on the box, the experimenter obviously flipped a switch at the same time, children no longer preferred the evidence from their own intervention and instead selected block A (69%).

The results of this study show that young children not only use deterministic covariation information to make causal inferences, but they can also use probabilistic covariation information to determine causal strength. On top of this, they differentiate between their own and others' interventions and prefer evidence from their own actions. Finally, they are sensitive to the confounding of their own actions and no longer prefer evidence from their own interventions when that evidence is confounded. Together, the findings presented in this section show that young children are sensitive to the quality of evidence, whether it is confounded or unambiguous, and that their exploratory play is affected by this information, such that they selectively perform informative interventions in the case of ambiguous evidence.

1.2.7 Generating and interpreting causal explanations

Another important aspect of young children's causal reasoning is the ability to explain causal evidence and events. Children's explanations reflect their knowledge about causality and can be used to facilitate their causal learning. For example, Schult and Wellman (1997) revealed that three- to four-year-olds can provide different types of verbal causal explanations for events in psychological, biological, or physical domains. Further, children distinguish between possible and impossible events in the physical and biological domains as measured by their explanations (Schult & Wellman, 1997). In addition to this, three- to four-year-olds can generate appropriate counterfactual alternatives for possible

events and their explanations of impossible events are related to their ability to correctly claim that no alternative actions can be generated for impossible events (Sobel, 2004).

To investigate two- to six-year-olds' causal explanations and the relation between these and children's subsequent exploratory behaviors, Legare (2012) used a blicket detector paradigm. She presented children with evidence that was either consistent or inconsistent with previous evidence they had observed (e.g., an object that looked like a blicket and behaved like a blicket vs. an object that looked like a blicket but that did not behave like a blicket). After observing the consistent or inconsistent situation, children were asked why that happened. Children's explanations were coded as causal function explanations (e.g., the blicket is broken), causal action explanations (e.g., it was not placed on the box correctly), and category explanations (e.g., that is (not) a blicket). When children observed inconsistent outcomes, and when they provided causal function explanations, they played longer with the blicket detector, suggesting that children were more motivated to explore the objects when the evidence was inconsistent with what should be expected and when they thought there was something wrong with the function of the object, that perhaps could be fixed. Further, in the inconsistent condition, children spontaneously generated new explanatory hypotheses.

Using a different paradigm, Legare and Lombrozo (2014) further investigated children's explanations and their effect on causal learning. They presented children with a gear toy consisting of a baseboard with pegs and a number of gears of different sizes and colors. Children who were prompted to explain how a gear toy worked ("Can you tell me how this works?") or to "explain the machine" performed better on measures of causal mechanism learning, selecting the correct causal piece to complete the gear set-up and correctly reconstructing the gear set-up to function as before, than children who either simply observed the gear toy in use or were prompted to describe the toy, rather than

explain it. Interestingly, the children in the explain conditions performed worse than children in the other conditions on measures of causally irrelevant features such as recognizing the color of the gears or correctly reconstructing spinning tops on the gears which did not have any causal role in the function of the toy. These results suggest that the process of generating causal explanations affects children's causal learning, focusing their attention on the causal features of causal relations. At the same time, perhaps because their attention is focused on the causal features, they are less able to recall other salient but non-causal features. In this way, children prioritize certain information or hypotheses that are likely to support their causal learning.

Not only can children's own explanations affect their causal learning, but the explanations of others can also influence children's ability to learn causal relations. For example, Sobel and Somerville (2009) investigated four-year-olds' ability to correctly identify causal structures and the influence of rationales for an action which revealed the causal structure. Children observed that when a light A was activated by a button, two other lights, B and C, were also activated. This pattern of evidence could be explained by two different causal structures, a common cause model and a chain model. In the common cause model, A activates both B and C. In the chain model, A activates B, which in turn activates C, but there is no direct relation between A and C. The experimenter explicitly outlined these two possibilities to the children. To find out which is the case, the experimenter suggested covering B (because the lights only activate according to their causal relation if they can "see" each other). The reasoning for why the experimenter suggested this was different in three conditions. In a baseline condition, the experimenter did not provide a rationale for covering B. In the second condition, the experimenter gave the rationale that they should cover B to see what happens when A is pressed while B is not visible. This was called the appropriate rationale condition because it explained a valid

reason for covering B in order to discover the causal relation between the lights. In the third condition, the experimenter used an inappropriate rationale, saying that he would cover B because he does not like B and will press A because A is pretty.

Children then observed evidence of a common cause and a chain model (counterbalanced) as the underlying causal structure. They were then asked to select a set of pictures that represented a chain or common cause model as the explanation for the light effects. In all conditions, children observed the same actions and outcomes, yet in the appropriate rationale condition, children were better able to identify the underlying causal structure than in the inappropriate or baseline conditions. These results suggest that children incorporate contextual information, such as others' explanations, into processing conditional probability information and learning about causal relations (Sobel & Somerville, 2009).

1.2.8 Similarity of causal reasoning to the process of scientific discovery

The development of children's causal reasoning has been likened to the process of scientific discovery (e.g., Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994). Drawing an analogy between conceptual structures and every day or scientific theories, the *theory theory* proposed that the cognitive development of conceptual structures was similar to revising theories (e.g., Carey, 1987; Gopnik & Meltzoff, 1997; Gopnik & Schulz, 2007; Keil, 1989; Wellman & Gelman, 1992). Children's knowledge about the world and scientific theories share the same structural, functional, and dynamic properties. Structurally, they are both abstract, coherent, and causal and they can have a hierarchical structure, such that some theories may describe very specific phenomena, but also be contained within an overarching theory. Functionally, they both facilitate the making of predictions, the generation of explanations and inferences, and the production of appropriate interventions on the world. Finally, they both have dynamic features, such as

the coordination of beliefs or hypotheses and evidence or data. Further, they are changing and adapting, based on new evidence (Gopnik & Meltzoff, 1997). In other words, children develop intuitive theories of the world based on their observations and interaction with their environment, but also adapt or revise those theories based on new evidence.

However, it is not clear if young children have explicit awareness of those theories or their revision (Kuhn & Pearsall, 2000).

More recent research on the theory theory has revealed the important role of statistical information and probabilities of events in children's causal learning, as well as the power of informal experimentation through exploratory play. This research has also shown that the process of theory revision is a gradual one, such that children adjust the probabilities of multiple different hypotheses based on the evidence they observe and favor the more probable hypothesis (e.g., Bonawitz et al., 2011; Cook et al., 2011; Gopnik et al., 2004; Gopnik & Wellman, 2012; Gweon, Tenenbaum, & Schulz, 2010).

The analogy between the development of conceptual structures and revision of theories is not the only way in which causal reasoning is similar to scientific processes. Recently there has been growing research interest in the comparison between causal reasoning and scientific reasoning processes. There is clearly some overlap in the types of abilities we see in very young children and those we see in mature scientific reasoners. For example, we have already outlined some studies showing that preschoolers are sensitive to the informativeness of evidence, recognizing when evidence is confounded and that, as a result, there is the potential for information gain (Cook et al., 2011; Gweon & Schulz, 2008; Schulz & Bonawitz, 2007). In those studies, children played and explored more with toys that generated confounded evidence and also showed novel information-seeking behaviors such as attempts to isolate variables. In this way, children show precocious

reasoning abilities and a tendency to search for information, which are both implicated in later scientific reasoning processes.

However, there is also a gap between children's early causal reasoning abilities and scientific reasoning abilities. Indeed, children have traditionally been shown to struggle with scientific reasoning, showing an inability to differentiate between hypotheses and evidence (Kuhn, 1989, 2002; Kuhn & Franklin, 2007). A mature ability to coordinate theory and evidence requires the recognition that the theory may be wrong and can or should be revised, and that evidence may support or help falsify a theory. This process also requires explicit reflection and thus relies on metacognitive abilities (Bullock, Sodian, & Koerber, 2009; Sodian & Bullock, 2008). Further, children often fail to control variables and ignore or distort evidence that does not support their prior beliefs (Amsel & Brock, 1996; Croker & Buchanan, 2011; Kuhn et al., 1988; Schauble, Glaser, Duschl, Schulze, & John, 1995), which is notable considering studies that showed that four-year-olds can infer causal relations that conflict with their prior beliefs about contact causality (Kushnir & Gopnik, 2007; Shultz, 1982).

The coordination of theory and evidence, the processes of generating, testing, and revising hypotheses, and the metacognitive process that are required to support explicit scientific reasoning are likely all implicated in this divide between causal reasoning and scientific reasoning abilities in young children. In the next sections, we will focus our review on the literature on scientific reasoning in early childhood before returning to this discussion on the relation between causal reasoning and scientific reasoning in the chapter summary.

1.3 Scientific Reasoning

Scientific reasoning has been a topic of research for decades and, as a result, there are numerous definitions and conceptualizations, and even names, of scientific reasoning. For example, some researchers use “scientific reasoning” and “scientific thinking” interchangeably (e.g., Koerber, Mayer, Osterhaus, Schwippert, & Sodian, 2015; Osterhaus, Koerber, & Sodian, 2017; Tolmie et al., 2016), while others make a distinction, suggesting that scientific reasoning is just one part of scientific thinking (e.g., Dunbar & Klahr, 2012; Klahr, Zimmerman, & Jirout, 2011; van der Graaf, van de Sande, Gijssels, & Segers, 2019). The latter group argues that scientific thinking consists of scientific reasoning (domain-general scientific process skills) and scientific knowledge (domain-specific content knowledge) (Klahr et al., 2011; Strand-Cary & Klahr, 2008). Further, this group argues that the first component may be a prerequisite for the second (Zimmerman, 2007). However, even this definition distinguishing scientific reasoning from scientific thinking and claiming scientific reasoning as domain-general processes is mired in ongoing discussion about whether scientific reasoning is domain-general or domain-specific (e.g., Fischer, Chinn, Engelmann, & Osborne, 2018; Schunn & Anderson, 1999; Tricot & Sweller, 2014).

On the one hand, as described in the above definition, scientific reasoning is thought to be a domain-general process or set of skills (Inhelder & Piaget, 1958). In other words, domain-general means that scientific reasoning can be applied across any number of domains and is generalizable; though it is “essential to science,” it is “not specific to it” (Kuhn, 2002, p. 498). The processes involved in scientific reasoning can also be applied to everyday contexts. This idea of transfer, that scientific reasoning learned and applied in one context can also be applied in different contexts, is essential to the hypothesis of domain-generality of scientific reasoning. Empirically, there is evidence for the domain-

generality of scientific reasoning, showing that children who learn to apply the control of variables strategy (CVS) to investigate race cars can also transfer and apply those skills to investigations of boats (Kuhn, Schauble, & Garcia-Mila, 1992). Similarly, Chen and Klahr (1999) showed that 4th graders can transfer CVS skills learned in the hands-on context of ramps to paper-and-pencil tests about plants.

On the other hand, some researchers claim that scientific reasoning must be domain-specific and does not represent a single scientific method applicable across even different sciences (H. H. Bauer, 1994), let alone other contexts. The role of content knowledge is important in the argument of domain-specificity; some argue that content knowledge is a requirement for reasoning (Sinatra & Chinn, 2012) and even that all relevant knowledge is domain-specific (Tricot & Sweller, 2014). This group suggests that it is not possible to apply domain-general skills without content knowledge.

Some researchers suggest a middle ground, that there are domain-general scientific reasoning skills, but acknowledging that there is an influence of content knowledge both in acquiring scientific reasoning skills and in applying them to contexts different from those in which they were learned (e.g., Carver & Shrager, 2012; Daxenberger, Csanadi, Ghanem, Kollar, & Gurevych, 2018; Erduran, 2007; Samarapungavan, 2018; Schauble, 2018). In this thesis, we will use the conceptualization of scientific reasoning as domain-general process skills, keeping in mind the influence of the context in which those skills are acquired and used.

In the early days of research on scientific reasoning, scientific reasoning was formulated as a process of problem-solving (Simon & Newell, 1970), which occurs as a search between the space of instances and the space of rules (Simon & Lea, 1974). Building upon this conceptualization, Klahr and Dunbar (1988) developed a framework, the Scientific Discovery as Dual Search (SDDS), to explain scientific reasoning as the

search between two spaces: the hypothesis space and the experiment space. The hypothesis space represents all possible hypotheses for a phenomenon and the experiment space represents all possible experiments that could be conducted to test the hypotheses. The process of scientific reasoning cycles between these two spaces, narrowing down the possible hypotheses as a result of the experiments (Klahr & Dunbar, 1988). This ability to differentiate and coordinate theories or hypotheses and evidence is critical to scientific reasoning (Kuhn, 1989, 2002; Kuhn & Franklin, 2007; Kuhn & Pearsall, 2000).

In the developmental literature, scientific reasoning has been defined as involving the skills needed for inquiry, experimentation, evidence evaluation, and inference, which are used for achieving scientific understanding (Zimmerman, 2007, p. 172). More specifically, it has been described as the reasoning and problem-solving skills involved in generating, testing, and revising hypotheses or theories (Morris et al., 2012, p. 61). Finally, it has also been claimed that the ability to reflect metacognitively on the process of knowledge acquisition and the process of change resulting from the above actions is a requirement of mature scientific reasoning (Kuhn & Dean, 2005). Importantly, a scientific reasoner uses these skills to intentionally seek knowledge (Kuhn, 2010).

1.3.1 The development of Scientific Reasoning

Conceptual models of the development of scientific reasoning have presented two possible pathways of development: (1) scientific reasoning is influenced by the development of general information-processing skills and (2) scientific reasoning is influenced by the development of a metaconceptual understanding of the distinction between hypotheses and evidence (Osterhaus, Koerber, & Sodian, 2015; 2017; Koerber & Osterhaus, 2019). In support of the first model, numerous studies show evidence of relations between scientific reasoning and intelligence, language abilities, and executive functioning (e.g., Koerber & Osterhaus, 2019; van der Graaf et al., 2016, 2018; refer to

Chapter 3 for more detail). However, research has also shown that general information-processing skills do not fully explain children's developing scientific reasoning skills. In addition, children's understanding of false beliefs, as well as the understanding of the nature of science, are related to their scientific reasoning abilities, for example in experimentation (Osterhaus et al., 2017) and understanding evidence (Astington, Pelletier & Homer, 2002). In particular, Osterhaus and colleagues (2017) have tested a model in elementary school children proposing that children's advanced Theory of Mind is a precursor for their epistemological understanding of the nature of science, which in turn is related to their abilities in experimentation.

In the following sections, we will describe in more detail some of these potential precursors of scientific reasoning, namely, metacognition, (advanced) Theory of Mind and False Belief understanding, and the understanding of the nature of science.

1.3.1.1 Metacognition

Metacognition is defined as “knowledge and cognition about cognitive phenomena” (Flavell, 1979, p. 906). In other words, it is the process of knowing or thinking about one's own or others' thinking (Kuhn, 2000). Metacognition can be further broken down into two components (Schneider, 2008). First, knowledge about the mental world and understanding beliefs (including false beliefs), desires, and mental verbs such as want, expect, believe, or think. This first component of metacognition has been the focus of research on how knowledge and understanding about the mental world develops in early childhood under the term Theory of Mind (Wellman, 1992). The second component is knowledge about memory, both declarative and procedural, including the processes of monitoring and self-regulation, for example, reflecting on what one does or does not know or how effectively one feels they have learned something, or allocating study time based on those reflections (Schneider, 2008). Research in this area is typically conducted with

older children or adults because of the focus on task-related problem-solving strategies (Flavell, 2000).

It is also important for metacognition and consequently scientific reasoning, that children understand from where or from whom evidence comes, or how they know something. Some studies have shown that children do not pay attention to or remember how they learn of evidence. For example, Gopnik and Graf (1988) found that three year old children could not say if they had learned about the contents of a drawer from seeing it themselves or simply being told about it, but five-year-olds did not have this issue. Kuhn and Pearsall (2000) claim that sensitivity to the origins of knowledge is developing around the same time as the understanding of false belief. However, more recent studies, some of which were discussed in the causal reasoning section, show that four-year-old children are sensitive to the source of evidence, for example, if they generate it themselves or observe another person generate evidence (Kushnir & Gopnik, 2018). Further, children seem to have a metacognitive understanding of what they know. For instance, Rohwer, Kloo, and Perner (2012) found that even three-year-olds could correctly report the state of their knowledge in the cases in which they had complete knowledge or complete ignorance.

Studies examining both metacognition and scientific reasoning have highlighted this important relation. For example, in a study of university students' susceptibility to ratio-bias, Amsel and colleagues (2008) found that students whom they categorized as being "competent" at metacognitive evaluation were more likely to recognize that when choosing between a ratio of 1:10 or 10:100, they should have no preference between the two options. In a study on scientific argumentation skills with 6th graders, Kuhn and colleagues (Kuhn, Goh, Iordanou, & Shaenfield, 2008) found that emphasizing metacognitive reflection improved children's scientific reasoning and argumentation abilities, suggesting a relation between metacognition and scientific reasoning. In

preschoolers, metacognitive abilities, such as cognitive monitoring and self-regulation of cognitive strategies, have been shown to predict success in solving problems across many different tasks (León, 2015; Maric & Sakac, 2018; Wang, 2015).

1.3.1.2 Theory of Mind and False Belief Understanding

Young children's Theory of Mind is developing in the time around three to five years of age (Perner, 1991; Wellman, 1985). By the age of three, children begin to use mental-state concepts such as desire, belief, or intention to explain both their own and other's behavior (Wellman, 1988). However, at three, children typically fail tasks that assess their understanding that beliefs do not necessarily correspond to reality. Commonly used False Belief tasks test whether children understand that another person can hold a belief that the children themselves know to be false (i.e., she believes that...; first-order false belief). For example, if a child is presented with a candy container, they will think it contains candy, but when shown that, in fact, there are pencils inside the container, they can adjust their belief about what is inside the candy container. However, they do not recognize that another person would initially hold the same false belief that they previously had - that there is candy inside the container. Instead, they believe the person would think there are pencils inside, an unlikely belief without having seen evidence of that case (Perner, 1991; Wimmer & Perner, 1983).

Other tasks go a step further and assess (false) beliefs about beliefs (i.e., he (falsely) believes that she believes that...; second-order false belief). This recursive process of reasoning about second and higher-order beliefs has been termed Advanced Theory of Mind (AToM; S.A. Miller, 2012). For example, in a classic task (Perner & Wimmer, 1985), two characters, Mary and John, know that an ice-cream truck is parked in a particular location. They are each independently told that the truck has moved location. However, John (falsely) believes that Mary still believes that the truck is in the original

location, because he does not know that she has also been told that it moved. The understanding of second-order false belief is developing slightly later between the ages of four and seven years (Coull, Leekam, & Bennett, 2006; Wellman, Cross, & Watson, 2001).

Research has shown that Theory of Mind, and False Belief understanding in particular, is related to children's developing scientific reasoning abilities. For example, Astington et al. (2002) found that second-order false belief was related to five- to seven-year-olds' ability to distinguish between causes of a situation and reasons for believing the situation, after controlling for language and nonverbal reasoning abilities. Piekny and colleagues (2013) found that understanding of false belief at age four predicted experimentation abilities at age five, after controlling for intelligence, language, executive functioning, and working memory. Sodian and colleagues (2016) found that both first- and second-order false belief understanding at five years predicted experimentation skills at eight years, independently of intelligence and executive functions. These findings suggest that children's developing Theory of Mind and understanding that beliefs can differ from reality may be important for distinguishing between beliefs (or theories or hypotheses) and evidence, which may in turn be critical for developing scientific reasoning abilities.

1.3.1.3 Nature of Science Understanding

The metaconceptual understanding of the distinction between theories and evidence and the ability to coordinate the two has been investigated through children's intuitive understanding of the nature of scientific knowledge (Driver, Leach, Millar, & Scott, 1996) and "how science functions" (McComas, Clough, & Almazroa, 1998, p. 5). McComas and colleagues (1998) summarized the nature of science by gathering common descriptions and science objectives from international science education standards. For example, an understanding of the nature of science includes understanding that scientific

knowledge is constructed and can be durable but is also subject to change; that the construction of scientific knowledge depends upon rational argumentation and skepticism; that scientific knowledge must be supported by evidence and justifications; that science attempts to explain natural phenomena using a variety of scientific methods, such as observation or experimentation; that theories play a role in constructing scientific knowledge; and that scientific knowledge is the result of global efforts, is integral to social and cultural tradition, and is affected by history (McComas et al., 1998).

The understanding of the nature of science has been described by Carey and colleagues (1989) as a progression from a naive understanding (Level 1), which is characterized by objectivism and a lack of a distinction between theories, experiments, and evidence, to an intermediate understanding (Level 1.5), which is characterized by some implicit understanding of knowledge construction, but no explicit notion of testing ideas. This progression continues to a basic understanding (Level 2), which is characterized by recognizing science as a search for explanations and of the need to test hypotheses, before reaching a mature understanding (Level 3), which is characterized by the recognition of science as a cyclical and cumulative process of knowledge construction for explaining natural phenomena (Carey, Evans, Honda, Unger, & Jay, 1989).

Research has shown that 4th and 7th graders' spontaneous responses to interview questions, such as what is science about, what are scientists' goals and how do they reach them, what is an experiment, and what are theories and hypotheses, fail to conceptualize theories or distinguish between theories and evidence and do not often exceed a Level 1 understanding of the nature of science (Carey et al., 1989; Grygier, 2008; Sodian, Thoermer, Kircher, Grygier, & Günther, 2002; Sodian, Jonen, Thoermer, & Kircher, 2006). One issue, however, with such a measurement is the demanding nature of the questions and the interview-style instrument. Children must be able to verbalize their

understanding and also spontaneously produce responses about topics on which they may never have reflected and may also not have the appropriate vocabulary to discuss in a sophisticated manner.

However, children's responses can be improved with short-term curricula targeting nature of science understanding, specifically with units focusing on perspective differences in perception or scientific exploration to determine what makes bread dough rise, and by contextualizing the outcome measure for younger children (Grygier, 2008; Sodian, et al., 2002; Sodian, et al., 2006). For example, in a pre-test, 21% of 4th graders were categorized as having an understanding of Level 1.5 or higher, while by the post-test, 40% of children were so categorized. In addition, children also greatly improved in their ability to produce a controlled test (from 11% to 69%), even though the curriculum did not specifically target this ability (Sodian, 2018).

To address the difficulty of providing responses to interview questions, researchers have developed and used different, possibly more developmentally-appropriate instruments to assess children's understanding of the nature of science (citation). For example, Koerber and colleagues (2015) developed a 66-item scale using different response formats such as forced choice, multiple select, multiple choice, and open-ended to assess elementary school children's (2nd - 4th grade) scientific thinking. For items measuring children's conception of the goals of science as well as theories and interpretive frameworks, 4th graders' performance ranged from 38% to 80%, better than 4th graders' performance on the interview-style nature of science questions in the pre-test of the study described above (21%).

In another study, Osterhaus and colleagues (2015) developed multiple-select items to more appropriately capture children's understanding of experimental design, as interview questions may be too difficult and multiple-choice items may overestimate

abilities due to the higher probability of selecting correctly by chance. With multiple-select items, the chance of correctly identifying both correct and incorrect statements is much lower (for example, with three items the chance of guessing correctly in a multiple-choice task is 33%, while with a multiple-select task, the chance of guessing correctly is 12.5%). Osterhaus and colleagues found that multiple-choice items likely overestimate performance on high difficulty items and multiple-select items may be particularly strict, only measuring competence when children can also overcome conflicting naive views. They concluded that multiple-select is preferred over multiple-choice for investigating advanced competencies. These studies highlight the importance of considering task difficulty for measuring scientific reasoning in children of different ages, a topic which we will return to in more detail later in this Chapter.

The progressive model of understanding the nature of scientific knowledge can also be applied to the development of scientific thinking and reasoning (Sodian, 2018). Looking at strategies of experimentation, for example, Sodian suggests that a Level 1 naive conception of something would manifest as simply reproducing an effect in response to a hypothesis, without even contrasting conditions; a child with a Level 1.5 intermediate conception would produce a contrastive test but fail to control potentially confounding variables; and a child with a Level 2 basic conception who is able to differentiate between hypotheses and evidence could produce a controlled experiment (Sodian, 2018).

In summary, children's ability to reflect on alternative possibilities (i.e., different hypotheses), their recognition of what they know or do not know (and can find out through testing hypotheses), their ability to distinguish between belief and reality, and their metaconceptual understanding of the nature of science are all potentially foundational for their developing scientific reasoning capacities.

In the following sections we will, first, examine evidence suggesting that children's scientific reasoning abilities are indeed limited, but will also discuss this evidence in light of a number of factors that influence the measurement of scientific reasoning abilities in children. Second, we will examine evidence suggesting that children's scientific reasoning abilities may not be quite as limited as initially believed.

1.3.2 Evidence of preadolescent children's limited Scientific Reasoning abilities

The purpose of experimentation is to investigate and determine cause-effect relations (McLeod, 2019). To do so, one must consider three different types of variables: the independent or focal variable (the thing you want to find out about and will manipulate, the potential cause), the dependent variable (the outcome measure or effect), and the extraneous or control variable(s) (other potential causes, but not the one you are interested in) (McLeod, 2019). A valid experiment must manipulate the independent variable and control for other extraneous variables to observe the effect of the independent variable on the dependent variable. Only by controlling extraneous variables, can one avoid that they interfere with or confound the effect one observes (McLeod, 2019). A strategy for controlling variables and producing controlled experiments is called the Control of Variables Strategy (CVS) (Chen & Klahr, 1999).

The process of experimentation and controlling variables has often been used to assess scientific reasoning abilities. For example, in one of the first studies to investigate scientific reasoning abilities, Inhelder and Piaget (1958) conducted a number of experiments to illuminate the development of reasoning abilities throughout childhood. Using a pendulum task, children had to determine whether the length of the string, the weight of the pendulum, or the strength of the push to the pendulum was the most important factor for how fast the pendulum would swing. They observed that children conducted confounded experiments, manipulating multiple variables at once, drew

inferences from those confounded and thus inconclusive tests, and preferred to produce effects, e.g., fast swings, rather than to test hypotheses. Only the children who had reached what Inhelder and Piaget termed the “formal operational stage” of development, around 12 years of age, were able to succeed at this task and control variables to conduct unconfounded experiments.

Inhelder and Piaget’s early research (1958) inspired scientific reasoning research for years, and indeed, many studies found evidence to support the idea that preadolescent children show limited abilities to engage in such complex reasoning in first assessments (i.e., without training) (Amsel & Brock, 1996; Bullock, 1991; Klahr & Dunbar, 1988; Klahr, Fay, & Dunbar, 1993; Klahr & Nigam, 2004; Koslowski, 1996; Kuhn, 1989; Kuhn et al., 1988, 1992; Kuhn & Phelps, 1982; Masnick & Klahr, 2003; Schauble, 1990, 1996; Tschirgi, 1980). In particular, research has shown that individuals are biased to produce effects rather than test hypotheses, they are influenced by prior beliefs, and they struggle to design controlled experiments. In the following sections, we describe a number of these studies presenting such evidence.

1.3.2.1 Desire to produce a positive effect

Tschirgi (1980) presented children (2nd, 4th, 6th grade) and adults with a story problem task in which a character attempted to make or do something in a multivariate scenario. There could be two or three variables, with two levels each, and the outcome of the event could be good or bad. The following describes one version of the task with three variables and a good outcome. The character baked a cake with three ingredients (honey, margarine, and whole wheat flour) and the cake turned out great. The character hypothesized that the cake was great because of one particular ingredient (honey). To test this hypothesis (or “to prove this point”), he can bake a second cake. Participants were presented with three options for the ingredients of the second cake. One option, the CVS

option, varied the focal variable (to use sugar instead of honey) and kept the other variable constant (margarine and whole-wheat flour). This is also called the Vary-One-Thing-At-a-Time strategy (VOTAT). The second option did not vary the focal variable but instead varied the other two variables (keeping honey but changing to butter and white flour). This strategy is called Hold-One-Thing-At-a-Time (HOTAT). The third option changed all three ingredients (sugar, butter, white flour; Change-All (CA)). Participants were presented with eight different versions of the task (four good-outcome and four bad-outcome) and asked to pick the answer that made the most sense to them, but that would still prove the point that the cake was good or bad because of a particular ingredient (Tschirgi, 1980).

Overall, the younger children (2nd and 4th grade) used the VOTAT strategy 35% of the time, the HOTAT strategy about 40% of the time, and the CA strategy about 25% of the time. Although the older children and adults performed better, they were still far from fully competent. Sixth graders used the VOTAT strategy 50% of the time, the HOTAT strategy about 34% of the time, and the CA strategy about 16% of the time. Adults used the VOTAT strategy 55% of the time, the HOTAT strategy about 38% of the time, and the CA strategy about 7% of the time. Further, Tschirgi (1980) found that the outcome of the story influenced the strategy that was used to test the hypothesis across all age groups. When the outcome of the story was good, participants selected the HOTAT strategy more frequently than VOTAT or CA. This meant keeping the presumed “good” variable and changing the other variables. When the outcome of the story was bad, participants selected the VOTAT strategy more frequently than HOTAT or CA. In this case, participants changed the “bad” variable to eliminate the negative outcome. Overall, younger children chose the CA strategy more often than adults, mostly in the bad-outcome stories. Tschirgi (1980) concluded that individuals are sensitive to the outcome of an event when testing a

hypothesis. In the case of a negative outcome, individuals logically attempt to produce disconfirming evidence through controlled tests, essentially trying to eliminate the negative effect. In the case of a positive outcome, individuals rather seek confirming evidence, keeping the “good” causal variable, and trying to maintain the positive effect (Tschirgi, 1980).

Kuhn and Phelps (1982) also found that 4th and 5th graders were motivated to produce effects rather than determine causes in a study investigating children’s experimentation strategies. They presented children with four beakers, which held four different colorless liquids. When a fifth clear liquid was added to the four beakers, the liquid in one of the beakers became cloudy. The students had to determine which liquid or combinations of liquids resulted in this effect by isolating and controlling variables. The students were not given any feedback or support other than what they observed through their investigations. Over 11 to 13 weeks, about half of the students were able to succeed in the task and used both valid experimentation and inference strategies. The other half, who were never able to identify the combination that produced the cloudy liquid, did not consistently use valid experimentation strategies, ranging from 9-45% of the time. Kuhn and Phelps identified “planfulness” as playing a key role in whether or not students were successful. Planfulness meant having a goal or purpose for the experiments, having thought of the potential outcomes, and eventually replacing invalid strategies with valid strategies (Kuhn & Phelps, 1982).

Finally, Zimmerman and Glaser (2001) also investigated 6th graders ability to design experiments depending on the hypothesized outcome. In a negative outcome condition, children were asked to design an experiment to test if tap water was bad for plants. All of these students suggested a controlled design, with almost 80% correctly manipulating the focal variable (water type). In a positive outcome condition, children

were asked to design an experiment to test if coffee grounds were good for plants. In this condition, 23% of students correctly used CVS to design a controlled experiment. Instead, students tended to test the generality of the claim, trying to determine for which plants coffee grounds were good. This study replicated the effect found by Tschirgi (1980) that children can use CVS in a negative outcome situation, in this case to avoid that plant health is affected by tap water, but fail to use CVS in a positive outcome situation, instead wanting to produce that positive effect (good plant health) in as many cases as possible.

These three studies present evidence that the outcome of an experiment has an influence on children's goals for experimentation and on their ability to use valid experimentation strategies, such that they are more likely to use valid strategies to avoid negative outcomes and invalid strategies to produce positive outcomes. The next section presents three studies that investigated the influence of prior beliefs on experimentation strategies and one study that investigated the influence of both prior beliefs and outcomes on experimentation strategies.

1.3.2.2 Testing hypotheses or evaluating evidence inconsistent with prior beliefs

Kuhn and colleagues (1988) investigated whether children distinguish between theories and evidence using content about which children had prior beliefs. They told children that the type of cake people ate (chocolate or carrot) determined whether or not they caught a cold. Children saw covariation evidence of characters that ate one type of cake and whether or not they then caught a cold. Based on the evidence, children were asked to explain how the type of cake made a difference, or to identify which variable was causal. Their explanations were coded as evidence-based if participants referred to the data presented to them and as theory-based if participants referred to their own beliefs or theories about what could make a difference. Children often ignored the evidence or distorted it to make it consistent with their prior beliefs (e.g., that sugar is bad for you). In

addition to these strategies, some participants would also adjust their theories to match the evidence available to them, but this seemed to happen without their noticing that they had adjusted their theory at all. Kuhn and colleagues took this as evidence that participants do not differentiate between theories and evidence and thus cannot recognize and reflect on the relation between them (Kuhn et al., 1988).

In another study investigating the coordination of theory and evidence, Kuhn and Pearsall (1998) showed that four- to six-year-old children do not use existing evidence to confirm or refute an assertion. In this study, children were shown sequences of pictures; for example, runners racing against each other. One of the runners is wearing fancy running shoes. The final picture showed the outcome of the race: the runner with the fancy shoes is holding the trophy. Children should explain the outcome (he won) and provide evidence for that outcome (he is holding the trophy). But children did not distinguish between a theory and evidence: Instead of providing evidence of the outcome (he is holding the trophy), they would provide a theory for why that was the case (his fancy shoes made him run faster). Kuhn and Pearsall took this pattern of responding as further evidence that children do not differentiate between theory and evidence and thus cannot coordinate the two in a “consciously controlled manner” (Kuhn & Pearsall, 2000, p. 114).

Amsel and Brock (1996) specifically investigated the role of strong prior beliefs on evidence evaluation abilities. Participants were selected if they believed in a relation between a plant being healthy and the presence or absence of sunlight and if they did not believe in a relation between a plant being healthy and the presence or absence of a magic charm. Participants observed four sets of evidence: perfect positive correlation between sunlight and plant health (confirming prior belief); zero correlation between charm and plant health (confirming prior belief); perfect positive correlation between charm and plant health (disconfirming prior belief); and zero correlation between sunlight and plant health

(disconfirming prior belief). Based only on the evidence presented to them (and not on what they knew about plants) participants should state whether each of the variables was causal or not. Two groups of children, 2nd to 3rd graders and 6th to 7th graders, tended to make judgements consistent with their prior beliefs rather than based on the evidence, even when the evidence did not support their beliefs. Children also made few evidence-based justifications and did so mostly when the evidence confirmed their beliefs (Amsel & Brock, 1996). This pattern further supported the claim that children do not distinguish between beliefs and evidence.

Crocker and Buchanan (2011) manipulated both the outcome of an experiment and the content about which children had strong prior beliefs to investigate children's (four- to ten-year-olds) hypothesis testing strategies. In this case, the good outcome was healthy teeth, and the bad outcome was rotting teeth. The causes were milk or soda and children had strong beliefs about their effects on teeth (i.e., milk is good for teeth and soda is bad for teeth). Thus, a belief-consistent positive outcome would be that milk causes healthy teeth and a belief-inconsistent negative outcome would be that milk causes rotting teeth. Children were able to select a valid test of the hypothesis when it was consistent with their beliefs and when the outcome was positive. They could also isolate variables when the hypothesis was inconsistent with their beliefs and the outcome was negative. However, in the other conditions (inconsistent-positive and consistent-negative), children did not test the hypotheses, instead, they tended to use strategies that produced positive outcomes or avoided negative outcomes. This study brought together both claims from the previous studies that young children do not distinguish between hypothesis testing versus producing a positive effect, nor do they distinguish between beliefs and evidence (Crocker & Buchanan, 2011).

The previous sections have investigated children's scientific reasoning abilities with tasks assessing their ability to select a valid controlled test of a hypothesis, to design a controlled test, to use experimentation to determine a cause, to evaluate evidence, and to use evidence to support claims. These studies found that children are biased to produce positive effects rather than test hypotheses and are influenced by their beliefs in testing hypotheses and evaluating evidence. The next section will further investigate children's experimentation abilities with hands-on tasks as well as simulations, focusing mostly on using CVS to produce controlled tests.

1.3.2.3 Designing controlled tests of a hypothesis

Chen and Klahr (1999) investigated young children's ability to use the Control of Variables Strategy with three different hands-on tasks: a sinking task, a springs task, and a slopes task. Each of these tasks consisted of four variables of two levels each. For example, the slope task consisted of a wooden ramp and ball, which could be manipulated in the following ways: the steepness of the slope of the ramp could be adjusted to be steep or gradual, the surface of the ramp could be smooth or rough, there were two types of balls, and the starting location of the ball could be adjusted to be at the top or middle of the ramp. Children in 2nd, 3rd, and 4th grade (seven to ten-year-olds) were asked to make a comparison to find out if a particular variable affected the outcome. In other words, they had to design an experiment that only manipulated the variable in question and kept all other variables constant. In the first phase, when children had to construct these comparisons without any training or support, 2nd graders constructed controlled comparisons 26% of the time, 3rd graders did so 34% of the time, and 4th graders did so 48% of the time.

Similarly, and using the same ramps task, Klahr and Nigam (2004) assessed 3rd and 4th graders' skills in designing controlled experiments and saw that, on average,

children could design less than one controlled experiment out of four trials. Masnick and Klahr (2003) found that 2nd graders designed controlled experiments with the ramps tasks 16% of the time and 4th graders did so 40% of the time. Similarly, Toth and colleagues (2000) found that, before instruction, 4th graders designed controlled experiments 30% of the time (Toth, Klahr, & Chen, 2000). These studies show that young elementary school children struggle to produce controlled tests in multivariate systems and that this ability appears to be developing throughout the early elementary school years.

In a less structured environment, Dunbar and Klahr (2013) investigated 3rd to 6th graders' ability to conduct experiments to determine how a robot truck works. In free play, children could press buttons that controlled the truck's behavior, e.g., forward, backward, turning, firing, pausing. There was also a repeat button, and children were told they should try to figure out how the repeat button works. Only two of 22 children were able to discover the correct rule, though 14 children were certain they had. Further, children often only conducted one experiment and they tended to ignore evidence that was inconsistent with their hypothesis (Dunbar & Klahr, 2013). This study showed that spontaneous experimentation behavior in unstructured environment is extremely limited, and children do not seem to have a clear understanding of when they had successfully figured something out. This stands in contrast to the findings from the causal reasoning literature, that preschoolers can spontaneously generate informative interventions in exploratory play.

The next few studies investigated children's experimentation strategies within computer-based programs. Kuhn (2007b) investigated 4th graders' use of CVS in multivariable systems. The Earthquake Forecaster asks students to consider five different variables, each with two levels, that might be related to the risk of an earthquake event and to determine which of the variables are causal. A second program, Ocean Voyage,

similarly asks students to consider variables that might influence ships' movement across the ocean. Kuhn had students interact with the Earthquake Forecaster program as a pretest and measured their use of CVS during their investigations. She defined use of CVS as consisting of the intention to find out about a particular variable, generating two tests that only changed the variable in question, and reaching the correct conclusion about the causal effect of the variable based on the outcome of their experiment. On the pretest, none of the children showed competence under those requirements (Kuhn, 2007b). Over a few weeks, students interacted with the Ocean Voyage program in the same manner and were asked to predict the outcome of specific variable combinations (prediction phase) and to identify which of the variables were causal (exploration phase). As a posttest, they returned to the Earthquake Forecaster program and 63% of the students were considered successful in using CVS (Kuhn, 2007b).

However, looking more closely at children's experiences with the Ocean Voyage program, their performance was inconsistent. For example, variables that students had identified as causal during their exploration phase were not necessarily considered causal during the prediction phase. Students also identified fewer variables as causal in the prediction phase compared to the exploration phase. Thus, successful use of CVS in the posttest (by 63% of children) was not necessarily indicative of a deeper understanding of CVS, as illustrated by inconsistent use during the Ocean Voyage interactions. Further, children seemed to lack an understanding that causal variables should always be causal or that multiple variables may combine to produce additive effects (Kuhn, 2007b). Other studies have found similar results, that children and adults do not have a full understanding of additive effects or interaction effects in multivariate systems (Kuhn & Dean, 2005; Kuhn, Iordanou, Pease, & Wirkala, 2008; Zohar, 1994).

Klahr, Fay, and Dunbar (1993) used a simulated version of the robot truck task described above (Dunbar & Klahr, 2013), in which children had to program a sequence of commands for a spaceship to move around on a computer screen. In this study, children were better able to discover the correct rule when it was plausible (75%; e.g., the repeat function repeats the entire program N times) than when it was not (43%; e.g., the repeat function repeats the N th step once). When given an implausible hypothesis, children tended to propose a different, plausible hypothesis, and ignored the implausible one. Children again performed very few experiments, representing a tiny proportion of the full experiment space. Finally, they designed uncontrolled experiments that were, consequently, difficult to interpret and, thus, could not draw valid conclusions from them (Klahr et al., 1993).

Schauble (1990) investigated 5th to 6th graders' ability to determine the relation between features of racecar design and speed using a microworld paradigm. There were five different features which could be varied, and children could design up to three cars at a time to compare. On average, children constructed valid experiments only 22% of the time. Children were also more likely to test features they believed to be causal and ignored features they thought were not causal, perhaps because they wanted to produce effects (faster cars) rather than determine causal relations, a pattern described in a number of studies earlier (e.g., Tschirgi, 1980). Performance did improve after repeated interactions with the microworld over eight weeks, increasing to about half of the experiments being valid, with the only feedback being the outcomes of their design actions on the speed of the cars.

Schauble (1996) found similar results in a second study that 5th and 6th graders used the VOTAT strategy about one-third of the time, but, in this case, children did not increase their use with repeated sessions. However, they did improve their ability to

provide valid justifications from about 25% of the time to approximately 50% over time. Finally, children investigated around 60% of the full experiment space, frequently conducting experiments they had already performed. These studies reveal that late-elementary school children struggle to conduct controlled experiments, though there appears to be some improvement from repeated interactions, and especially improvement in making valid inferences over time. Further, they do not explore the full experiment space in order to reach a conclusion and tend to test variables they believe to be causal often resulting in repeated experiments (Schauble, 1990).

Bullock and Ziegler (1999) investigated the development of scientific reasoning abilities from 3rd to 6th grade in a longitudinal study. They presented children with a story problem task in which the main character had to design a product and test whether a particular variable was important for producing successful outcomes. One version of the task gave the character the goal to produce a fuel-efficient airplane. There were always three variables, with two levels each, that might make a difference to the outcome. The variables were named, pictured, and grouped by dimension.

The story problem explained that the character wants to test one variable, the focal variable, to see if it makes a difference. The task included two response measures. The first measure asked participants to produce a test, ideally a controlled test, by selecting each of the features for two planes. The second measure asked participants to choose an appropriate test from eight cards illustrating eight different airplanes (all possible combinations of the 2 x 2 x 2 design). First, there were no constraints on how many cards a participant selected for the test. Then, participants were instructed to pick only two cards. A correct, controlled test would vary the focal variable while holding the other variables constant (Bullock & Ziegler, 1999).

Neither children nor many adults were likely to *produce* controlled tests, with only 30-35% of 6th graders and adults producing a controlled test and even fewer of the younger children succeeding. When asked to *choose* a controlled test, however, more than a third of the 3rd graders, 60% of 4th and 5th graders, and close to 80% of 6th graders and adults succeeded in selecting cards that represented a controlled test. Further, when asked for a justification of their controlled test, more than 50% of 4th graders, 80% of 5th graders, and almost all of the 6th graders justified this choice in terms of controlling variables (Bullock & Ziegler, 1999), suggesting an explicit understanding of CVS.

A number of conclusions can be drawn from the studies reviewed in this last section. First, that elementary school-aged children struggle to produce controlled experiments without instruction. Though there appears to be development of this ability throughout elementary school, a minority of children can successfully produce controlled experiments. Second, repeated practice, even without instruction, seems to have some benefit for improving these abilities.

Looking at task features, additive and interaction effects in multivariable systems present even more difficulty and there is a difference in difficulty between producing and selecting controlled tests. This finding, in addition to those from the previous sections regarding the outcomes of experiments and the influence of prior beliefs, necessitates a deeper examination of factors that could influence the assessment of children's scientific reasoning abilities. We will outline these factors in the next section.

1.3.3 Factors influencing children's Scientific Reasoning and CVS abilities

The studies described above present a picture of children's CVS and scientific reasoning abilities as quite limited. However, other patterns also emerge that might provide some explanation for why children (and adults) seem to struggle so much with

these tasks, possibly as a result of the assessment features. We will briefly re-summarize the findings above in the light of task features.

1.3.3.1 Prior beliefs

A number of studies described above investigated the impact of prior beliefs on children's ability to interpret evidence or conduct experiments. When strong prior beliefs were held about task content, children were less able to reason scientifically. Instead, they ignored evidence that conflicted with and selectively attended to evidence that supported prior beliefs, they distorted the evidence to match their theories, or they adapted their theories to match the evidence, but without realizing they had done so (Kuhn et al., 1988). They were unlikely to consider hypotheses that seemed implausible (e.g., Klahr et al., 1993). They tended to use experimentation to find support for prior beliefs (Croker & Buchanan, 2011) or to produce effects rather than to illuminate causal relations (Carey, Evans, Honda, Jay, & Unger, 1989; Schauble, Glaser, Raghavan, & Reiner, 1991; Tschirgi, 1980). The ability to overcome prior beliefs seemed to be present starting in middle childhood and developing throughout elementary school (Amsel & Brock, 1996; Kuhn et al., 1988). Thus, assessments that include content about which individuals have strong prior beliefs may both over- and underestimate scientific reasoning abilities.

1.3.3.2 Outcomes of experimentation

Additionally, the outcome of experimentation seems to also influence children's strategies. When an outcome is negative, children tend to be more likely to use CVS to explain it or to determine the cause, however, when the outcome is positive, children instead prefer to keep the focal variable constant, assuming it is the cause of the good outcome, and change the other variables (Siler & Klahr, 2012; Tschirgi, 1980; Zimmerman & Glaser, 2001). Other studies show that children often mistake the goal of experimentation as trying to produce an effect, such as a fast car (Schauble, 1990), a

change in liquid color (Kuhn & Phelps, 1982), or a bubbling effect from yeast mixtures (Carey et al., 1989), and that they tend to attempt this “rather haphazardly” (Carey et al., 1989, p. 518).

Further, when both prior belief and type of outcome are considered, an interaction emerges: when an outcome was negative, children were more likely to use CVS in the belief-inconsistent condition, but when an outcome was positive, young children were more likely to use CVS in the belief-consistent condition (Croker & Buchanan, 2011). These findings suggest that children focus especially on the outcomes of experimentation, with the goal of producing positive effects rather than determining cause-effect relations. It is also possible that the domain knowledge about a task further draws attention to the outcome of an experiment and potentially influences the design or selection of an experiment. In this case, it would be important to consider how potential outcomes of an experiment are perceived and to prefer neutral outcomes over positive or negative outcomes for assessing scientific reasoning abilities.

1.3.3.3 Design and difficulty of tasks

Another factor that can affect performance and, thus, the assessment of scientific reasoning abilities is the design of the task. Saffran and colleagues (2016) investigated the effect of the symmetry of variables on children’s ability to evaluate covariation evidence as presented in a summary table. In a symmetrical condition, the variable was presented as two levels, A or B (e.g., fertilizer A or fertilizer B). In the asymmetrical condition, the variable was also presented as two levels, the presence or absence of the fertilizer. Children (2nd and 4th graders) were affected by the symmetry manipulation: they were less likely to provide a correct judgement of causality in the asymmetrical condition (46%) as opposed to the symmetrical condition (69%). In the asymmetrical condition, it is possible that the absence of the fertilizer results in reduced attention or in children ignoring those

cells, while in the symmetrical condition, with the presence of two types of fertilizer, all the cells are equally relevant and thus children pay attention to and use all of the information to make a judgement (Saffran, Barchfeld, Sodian, & Alibali, 2016).

Further, the requirements of the task itself should be considered in the assessment of scientific reasoning abilities. There is a clear difference in the requirement to recognize whether a test is controlled or confounded and the requirement to produce a controlled test oneself. As we saw in the study by Bullock and Ziegler (1999), across children from 3rd to 6th grade and including adults, it was easier to recognize and select a controlled test than to produce a controlled test. Almost 40% of 3rd graders, over 60% of 4th and 5th graders, and over 80% of 6th graders and adults could select a controlled test, while less than 5% of 3rd graders could produce a controlled test and less than 40% of 6th graders and adults could do so. This pattern of performance was discussed by Bloom (1956), in which he labeled “evaluation” as making judgements about others’ solutions and “synthesis” as generating new solutions. Synthesis was later relabeled “creating” by Krathwohl and Anderson (2001) and they further argued that creating is more advanced (i.e., more difficult) than evaluating. A study of physics teachers’ understanding of scientific evidence also found that they were better at identifying flaws in others’ experimental designs or methods than in designing valid experiments themselves (Taylor & Dana, 2003). This pattern of performance may reflect a difference in implicitly understanding experimentation and the importance of controlling variables without the ability to explicitly explain why it is important or how to do it (Sodian & Mayer, 2013). The level of difficulty, whether in having children select or produce controlled tests, or in the number of variables children must consider (Siegler, 1976), is an important factor to keep in mind when interpreting findings of children’s scientific reasoning abilities, as it is clear that the difficulty of tasks affects the detection of abilities (Crocker & Buchanan, 2011; Koerber & Osterhaus, 2019).

1.3.3.4 Content of tasks

Also likely influencing children's abilities to engage in scientific reasoning is the content of the tasks which are used to measure scientific reasoning. As seen in the studies described above, the content can vary greatly. For example, tasks have included content on pendulums, ramps, springs, sinking or floating, chemistry, earthquakes, robots, planes, ships, race cars, baking cakes, dental health, caring for plants, catching colds, etc. These also represent a distinction between tasks focused on scientific content or tasks focused on everyday contexts.

Some studies have investigated the influence of task content on reasoning scientifically. For example, Linn, Clement, and Pulos (1983) investigated what they called laboratory tasks and naturalistic tasks. They made the distinction that laboratory tasks typically involved apparatus, such as ramps, they represented closed systems where the variables to be investigated are limited as are the possibilities for investigation, and often variables had effects which were already well understood. Naturalistic tasks, however, were often paper-and-pencil tasks or verbal tasks, the effects of variables were less well known, and they represented an open system where the possible variables and manipulations of the variables were not limited. For example, what affects weight loss or the gas mileage of cars. Linn and colleagues found that a significant part of the variance in adolescents' performance on their tasks was associated with prior domain knowledge of the variables.

Further, children tended to only pay attention to variables they thought "made a difference," and ignored other variables rather than controlling for them. They were more likely to engage in this behavior in the naturalistic tasks than in the laboratory tasks. Consequently, children were better able to engage in scientific reasoning and use CVS when there were constraints on the tasks. Contrary to the findings by Linn and colleagues

(1983), that children performed better on the laboratory tasks as opposed to the naturalistic tasks, other studies have found that students perform better on tasks with everyday contexts rather than scientific content (Song & Black, 1992). These results may suggest that constrained tasks in which the number and effects of variables are limited and everyday content with which children have a connection are both important factors in assessing their scientific reasoning performance.

Returning to the discussion of domain-general versus domain-specific scientific reasoning, this distinction can also be applied to the tasks used to measure scientific reasoning. The goal of domain-general tasks is to minimize the influence of content knowledge on the measure of scientific reasoning or CVS. Tasks that use abstract, knowledge-lean contexts or everyday contexts (often with fictional data) fall into this category. Because participants do not have any prior knowledge about the task content, this also limits the effect of prior beliefs on performance. For example, one should have no reason to believe that the color of chewing gum affects dental health (Koerber, Sodian, Thoermer, & Nett, 2005). Domain-general tasks are typically used to investigate pre- or elementary school children's scientific reasoning abilities to avoid the influence of content knowledge or prior beliefs on the assessment of scientific reasoning abilities.

Domain-specific tasks typically use scientific content, for example, in the domains of physics, chemistry, or biology. In a review of scientific reasoning tasks, Opitz and colleagues (2017) found that almost three-quarters of the tasks focused on biology (34%), physics (20%), and chemistry (20%) content (Opitz, Heene, & Fischer, 2017). As such, domain-specific tasks are more commonly used to assess scientific reasoning in older children, students, and adults. Domain-general tasks may be more successful in assessing both young children's and adults' scientific reasoning abilities by avoiding the influence of prior beliefs and prior knowledge about task content. However, there is also discussion

about whether or not performance on domain-general tasks is predictive of performance on other, domain-specific tasks, or in other words, if domain-general abilities transfer to other tasks (Kind, 2013; Millar & Driver, 1987; Osborne, 2013). We will return to the discussion of domain-generality versus domain-specificity when we review the literature on training CVS abilities in Chapter 4. Regardless, it is clear that the content of tasks used to measure scientific reasoning must be considered when interpreting the findings of scientific reasoning abilities on such tasks (Kuhn et al., 1992).

In summary, this section has presented a number of studies showing limited scientific reasoning abilities in both children and adults, without instruction or training. It also discussed factors that contribute to some of the difficulties that arise when assessing scientific reasoning abilities, including the influence of prior beliefs and prior content knowledge about the tasks, the lure of producing positive effects, and the design and level of difficulty of task requirements. In the next section, we will take a look at studies that present more optimistic findings regarding young children's scientific reasoning abilities.

1.3.4 Evidence of preadolescent children's beginning Scientific Reasoning abilities

The previous section painted a dismal picture of children's scientific reasoning abilities. Studies showed that, before adolescence, children struggle with many scientific reasoning tasks, and that even into adulthood full competence remains elusive. However, the previous section also discussed a number of factors that may contribute to poor performance in assessments. In this section, we will take a second look at some of the studies reviewed above and how, in some cases, they show beginning scientific reasoning abilities in children. We will also review studies that attempted to address some of those task factors and, as a result, found signs of scientific reasoning abilities even in very young children, suggesting that young children's abilities may have been underestimated

(Wilkening & Sodian, 2005; Zimmerman, 2007). The following section is structured by focusing on three subcomponents of scientific reasoning: evidence evaluation, experimentation, and explanation and argumentation.

1.3.4.1 Evidence evaluation

The ability to evaluate and interpret evidence is an important component of scientific reasoning. Children must recognize whether evidence is the result of confounded or controlled experiments and, thus, whether or not it can be interpreted. They must also understand how evidence relates to theories or hypotheses. This section will describe a number of studies that suggest young children have beginning abilities in evaluating evidence.

In a study described in the previous section, we saw that when children's beliefs are in line with hypotheses, they perform on the same level as adults in making judgements about the causality of variables from covariation data (Amsel & Brock, 1996). When they are not influenced by prior beliefs or prior knowledge, children as young as five years of age understand that beliefs or inferences can be based on evidence. Ruffman, Perner, Olson, and Doherty (1993) showed that five-year-olds, but not four-year-olds, understand that a person (or puppet character) will reach a conclusion based on the evidence available to them. For example, children observed covariation evidence that all boys who ate green food lost their teeth, while all boys who ate red food did not. Then, with children watching, the experimenter switched the evidence around so that it looked like all the boys who ate the red food lost teeth and the boys who ate green food did not. When a puppet character, Sally, arrived and observed this "Faked Evidence," children were asked to predict what conclusion Sally would reach based on this evidence, i.e., which food would Sally say makes the boys' teeth fall out? Five-year-old children could report that Sally would generate a hypothesis based on the falsified evidence that the red

food causes tooth loss, even though they knew that it was the green food that did so. This task also represents the importance of Theory of Mind for such scientific reasoning tasks, for example, that children realize that others can have different beliefs and that there is a distinction between beliefs and evidence.

Waters, Siegal, and Slaughter (2000) used the same task but adjusted the wording of the questions to more specifically advise children to use the evidence in front of them to answer the questions. Children in the specific questioning condition showed better performance, with over half of four-year-olds succeeding on the faked evidence task. This change in wording to highlight the need to use evidence made it possible for even four-year-olds to succeed in the task, whereas before they did not. Koerber and colleagues (2005) replicated these findings with four- to six-year-olds, showing that, with perfect covariation, even four-year-olds can correctly evaluate evidence and attribute different conclusions to a puppet according to the evidence available to them. Another study showed that when children cannot rely on prior knowledge, for example when they are asked to make evidence-based decisions about whether a novel animal is real or fantastical on the basis of evidence left behind by that animal, four-year-olds show beginning abilities to make a decision based on the evidence, but there is a clear development from four to six years of age (Tullos & Woolley, 2009). From these studies, it is evident that even young children are successful in coordinating theory and evidence: they understand that evidence can be used to draw conclusions or generate hypotheses. This stands in contrast with the conclusion that children do not have these abilities and highlights that the tasks used to assess scientific reasoning can affect the assessment of performance.

Piekny and Maehler (2013) investigated children's (preschool, 1st, 3rd, 5th grade) evidence evaluation abilities using the gum/dental health task (Koerber et al., 2005). They found that for perfect covariation, a majority of preschoolers could correctly identify the

relation, starting at 70% of four-year-old preschoolers and increasing to 100% of 5th graders. In the case of imperfect covariation, the results were less clear, with 40% of four-year-old preschoolers, 60% of five-year-old preschoolers, 33% of 1st graders, 54% of 3rd graders, and 90% of 5th graders succeeding on the task. The task with imperfect covariation was more difficult than perfect covariation, but it is unclear why the 1st graders performed so much worse than both the four- and five-year-old preschoolers. The authors suggest that an incomplete shift from base-rate reasoning to calculating probabilities could be the cause of this drop in performance.

Using the same gum/dental health task as above (Koerber et al., 2005), Piekny, Grube, and Maehler (2014) investigated preschool children's evidence evaluation abilities longitudinally from four to six years of age. They found that when evidence represented a perfect pattern of covariation, even four-year-olds (68%) succeed in identifying causal relations, though there is still improvement at age five (85%) and six (97%). Over half of the children showed stable, perfect performance across three time-points and another third showed patterns of performance indicating improvement. In the case of imperfect covariation however, only 20% of four-year-olds succeeded in the task, while over 80% of five- and six-year-olds did. A third of the children showed stable, perfect performance across three time points, and 42% showed patterns of performance indicating improvement. The results of these two studies support previous research showing that preschool children can interpret perfect covariation data (Koerber et al., 2005). When it comes to imperfect covariation data, however, they are less able to interpret the evidence, and there seem to be discrepancies in how well preschool children can interpret this data, ranging from 20-68% of four-year-olds succeeding on such a task.

Saffran and colleagues (2016) showed that when evidence is presented in a way that highlights the symmetry of variables (e.g., different types of fertilizers as opposed to

the presence or absence of fertilizers) 2nd and 4th graders could make correct judgements of causality on the basis of covariation data almost 70% of the time.

Taken together, these studies show that even young children can successfully coordinate theory and evidence under certain conditions: they understand that evidence can be used to draw conclusions or generate hypotheses and they are sensitive to the quality of evidence. These abilities are also undergoing development from preschool through elementary school.

1.3.4.2 Experimentation

In addition to showing an understanding of evidence and its relation to theory in evidence evaluation tasks, children also show beginning abilities to seek information through investigation and experimentation in simple tasks. For example, Sodian and colleagues (1991) investigated young children's (1st and 2nd grade) scientific reasoning abilities using a simple task about which children had no prior knowledge or beliefs. A story-problem task presented a scenario in which two brothers knew there was a mouse in their house, but they disagreed about its size; one brother thought it was a big mouse, the other thought it was small. There were two boxes with food inside, one with a small opening and one with a large opening, that the brothers could leave out for the mouse. Children were asked which box should be put out if the brothers wanted to tell for sure whether the mouse is big or small (hypothesis testing/ Find Out). In a second scenario, they were asked which house should be put out if they wanted to make sure the mouse got some food (effect production/ Feed). These two scenarios differentiated between testing a hypothesis, finding out the size of the mouse, and producing an effect, feeding the mouse.

Children were then presented with solutions that resulted either in an inconclusive or a conclusive test and asked if they could know if the mouse is big or small (Conclusive Test). Results showed that 55% of 1st graders and 86% of 2nd graders succeeded in this

task by answering all three questions correctly (hypothesis testing, effect production, and conclusive test). Sodian and colleagues (1991) concluded that these results show that young children differentiate between beliefs and evidence, because they understand the difference between testing a hypothesis and producing an effect, they can distinguish between conclusive and inconclusive tests, and they understand what inferences can be made from a conclusive test.

Piekny and Maehler (2013) used the same Mouse House task described above with four- to 12-year-old children. They found that a majority of preschoolers could correctly produce an effect, while about a third of the preschoolers could correctly test a hypothesis. Approximately 18% of preschoolers could distinguish between those two tasks, responding correctly on both. About 30% of 1st graders showed this performance, 60% of 3rd graders, and 75% of 5th graders. The results of this study illustrate the development of these abilities from preschool through elementary school.

Piekny and colleagues (2014) investigated preschool children's experimentation abilities longitudinally from four to six years of age using the Mouse House task. They found that approximately 30% of both four- and five-year-olds could distinguish between testing a hypothesis and producing an effect, succeeding on both the feed and find out questions. Between the age of five and six, there was an improvement in children's abilities, and with around half of six-year-olds succeeding on both tasks. These findings suggest that by six years of age, children have a beginning understanding of experimentation and the difference between testing a hypothesis and producing an effect.

The previous three studies show that preschool children show beginning abilities in distinguishing between testing a hypothesis and producing an effect, though the proportion of children showing this ability ranged from between 20-30% of four- to five-year-olds, to

30-50% of six-year-olds. Even once in elementary school, performance was quite varied, ranging from 30-55% of 1st graders and 60-86% of 2nd graders.

Köksal-Tuncer and Sodian (2018) investigated young children's hypothesis testing skills using the blicket detector paradigm. They first led children to believe that the weight of objects was what activated the detector. They then showed children evidence that was inconsistent with this hypothesis and subsequently revealed an alternative cause of the effect, a sticker hidden on the bottom of the object. With this procedure, they expected children would be more systematic in their hypothesis testing behavior than in a typical exploration task. Hypothesis testing behavior was measured by which types of objects children chose to place on the box after observing the disconfirming evidence. The choices consisted of four objects, one heavy object with a sticker and one without, and one light object with a sticker and one without. To test the sticker-cause hypothesis, children should place objects that control for the weight as a potential cause, i.e., the heavy object without a sticker or the light object with a sticker. They found that children did not prefer unconfounded objects over confounded objects in the exploration phase. When children were categorized by their overall hypothesis testing patterns, approximately half of the children showed a contrastive testing pattern.

After this hypothesis-testing phase, children's beliefs about the cause of the effect were checked by having them sort novel objects into two categories, objects that would make the box light up and objects that would not make the box light up. Finally, the experimenter would present children with a false hypothesis (the heavy objects make the box light up) and show children confounded evidence to support this false hypothesis. In this phase, Köksal-Tuncer and Sodian (2018) investigated if children would spontaneously argue against the false claim and use evidence to disprove it. In the counterargumentation task, 64% of children argued against the experimenter's false hypothesis, with 43% of

those children also providing disconfirming statements. Children interacted more with the unconfounded objects to generate disconfirming evidence than with the confounded evidence and 38% of children generated only disconfirming evidence. Thus, they showed that young children have a beginning ability to use unconfounded evidence to disconfirm claims they know to be false.

Testing a hypothesis becomes more difficult with an increase in variables, thus, many studies look at the selection of controlled tests by presenting children with options to choose from. Bullock (1991) asked children to help a character figure out if a lantern with a top on it will stay lit better in windy conditions than a lantern without a top. Children had to select two cards from a set of cards showing different lanterns designed with different features. There were three features with two levels each: the presence or absence of a top on the lantern, the size of the holes in the panes of the lantern, and the size of the candle inside the lantern. By 3rd grade, children showed an understanding of the need for performing a contrastive test of the focal variable. By 4th grade, children could select two cards, one showing a lantern with a top and a second card showing a lantern without, but both lanterns had the same other features, the same size holes and the same candle size, representing a controlled comparison. However, they were unable to generate a suggestion for a controlled test before being presented with the card options. Half of the children who selected a controlled test were also able to justify their choice in terms of controlling variables.

The study by Bullock and Ziegler (1999) already described in the previous section focused on the difference between selecting and producing controlled tests. They found that across all ages, selecting a controlled test was easier than producing one. By 4th grade, a majority of children could select a controlled test, replicating the results of Bullock (1991). Further, when asked to justify their choice of a controlled test, more than half of

4th graders, 80% of 5th graders, and almost all of the 6th graders justified this choice in terms of controlling variables, showing even better performance than Bullock (1991), and suggesting an explicit understanding of CVS. In addition to this, children recognized the importance of isolating and manipulating the focal variable, resulting in the production of a contrastive test. In 3rd grade, over 70% could produce a contrastive test. Where children struggled was in controlling the other variables at the same time. This finding has also been more recently replicated by Koerber, Sodian, Kropf, Mayer, and Schwippert (2011).

Siegler and Chen (1998) investigated four- and five-year-olds' abilities in experimentation using a scale and some objects. The weight of objects could be manipulated and the distance from the fulcrum could be manipulated. Objects were placed on the scale, but initially, the scale did not tip because there were blocks under each side, holding it stable. In 16 trials, children were first asked to predict which side of the scale would tip down when the blocks were removed and also asked to explain why they thought that would happen. The blocks were then removed, and children could observe which direction the scale tipped. When only the variable weight was under question, five-year-olds initially performed better than the four-year-olds in making correct predictions (80% vs. 56%). However, by the end of the 16 trials, both age groups were performing equally well, at around 85% accuracy. When distance was the variable under question, four-year-olds showed fairly stable prediction performance across all trials at around 40-45% accuracy, but five-year-olds started at around 40% accuracy and improved to about 70% accuracy at the end of 16 trials. When the variables were combined and children had to make predictions, for example, if a heavier weight close to the fulcrum or a lighter weight further from the fulcrum would tip the scale, accuracy was much lower (~6-28%). These findings show that five-year-olds already have a pretty stable understanding of expected outcomes of single-variable experiments and four-year-olds can adjust their

predictions quickly on the basis of observed evidence, at least in the case when the variable (weight) was somewhat intuitive. However, both age groups struggled to make outcome predictions for multivariable experiments.

Van Schijndel and colleagues (2015) investigated children's (four- to nine-year-olds) performance of unconfounded experiments during exploratory play in the domain of physics, specifically shadow size (van Schijndel, Visser, van Bers, Raijmakers, 2015). Shadow size is determined by the size of the objects creating the shadow as well as their distance from the light source: bigger objects make bigger shadows, and objects closer to the light source make bigger shadows, but a small object closer to the light source could have a bigger shadow than a big object farther away from the light source. Thus, both factors must be considered when predicting shadow size. Van Schijndel and colleagues took advantage of children's intuitive theories that only object size is important to shadow size (see S.-M. Chen, 2009) and manipulated whether children observed evidence consistent or inconsistent with that theory. They used puppets of two different sizes and created a set-up that allowed for the placement of two puppets at three different distances from the light source.

In one condition, children observed evidence that violated their naïve theory, that the small puppet created a bigger shadow than the big puppet. In a second condition, children observed evidence that was consistent with their theory, that the big puppet produced a bigger shadow than the small puppet. After observing either the consistent or inconsistent evidence, children were allowed to play freely with the puppets and the shadow machine while the experimenters observed children's experimentation behavior. Van Schijndel and colleagues (2015) were particularly interested if children conducted controlled experiments, varying either puppet size or distance and keeping the other variable constant.

Children who observed evidence that was inconsistent with their theory were more likely to conduct unconfounded experiments (100%) than children in the consistent evidence condition (50%). Interestingly, though, the majority of those unconfounded experiments were ones that varied size and kept distance constant, in other words, it appears that children were trying to confirm their theory rather than look for alternative hypotheses (e.g., that distance matters). This explanation is somewhat reflected in the outcome measure, that very few children (10%) revised their theory to include distance from the light source as an important factor in shadow size. All of the children who conducted at least one unconfounded experiment varying distance revised their theory to include distance as an important factor. This study provides evidence of spontaneous use of CVS in young children during free exploratory play. It also reflects, again, the effect prior belief can have on an individual's ability to perform controlled experiments. However, though children spontaneously used CVS, they did not appear to do so to explicitly test a hypothesis, nor did many children revise their theory on what factors were relevant for shadow size (van Schijndel et al., 2015).

Van der Graaf and colleagues (2015) investigated preschoolers' (four- to six-year-olds) use of CVS using the ramps task and a dynamic assessment paradigm. The ramps tasks consisted of four different variables: slope, ramp surface, the weight of balls, and starting location. The dynamic assessment consisted of four levels with four experiments each and two chances to design each experiment. This meant that children would start on Level 1 and design an experiment contrasting just one variable, for example, contrasting the type of ball. They would have two chances to correctly design this experiment. They would then proceed to design three more experiments for each of the other three variables. As long as children designed at least one experiment correctly, they would proceed to the next level in which a second, then third, then fourth variable was added. After each

attempt to design an experiment, children were asked why they had built the ramps in that way and were then told if their design was correct or not and why or why not. If the children were unable to design the experiment after two attempts, the experimenter set up the experiment correctly and explained why it should be designed that way.

All of the children were able to design at least one experiment at Level 1 correctly ($M = 3.4$ correct experiments). Almost 90% of children were able to design at least one experiment with two variables correctly ($M = 1.8$ correct experiments). Forty-seven percent of children were able to design at least one experiment with three variables correctly ($M = 0.9$ correct experiments). And finally, 31% of children were able to design at least one experiment with four variables correctly ($M = 0.7$ correct experiments). If one considers points for each variable correctly assigned, i.e., contrasted or controlled, children assigned slightly less than half of all variables correctly. The authors also found that there were differences between younger and older kindergarten classes, such that older kindergarteners (five- to six-year-olds) performed significantly better than younger kindergarteners (four- to five-year-olds) (van der Graaf et al., 2015). Thus, this study found that preschoolers show a beginning understanding of using CVS to design controlled experiments. Important to consider, however, is that children received very explicit feedback repeatedly after each attempt to design an experiment. This would mean that children could have received between 4 and 8 instances of feedback on their experimental design at each level. Additionally, van der Graaf and colleagues (2015) admit that their sample of preschoolers came from what they called a “talent hotbed” school, which gives extra attention to science and technology.

This section presented a number of studies which suggest that young children do indeed show some beginning understanding of experimentation and the use of the control of variables strategy. They distinguish between testing a hypothesis and producing an

effect; they show sensitivity to the conclusiveness of tests; they recognize whether evidence is confounded or unconfounded (and understand that unconfounded evidence should be used to support claims). Some of the studies even show these abilities were beginning to emerge at the preschool age, with four- and five-year-olds able to design unconfounded experiments during exploratory play and design unconfounded experiments with up to four variables when provided with extensive and repeated feedback. Many of the experiments described above ask children for explanations or justifications for their selection of controlled experiments or for their conclusions. The next section will briefly discuss the role of explanation and argumentation for scientific reasoning and the control of variables strategy.

1.3.4.3 Explanation and argumentation

Another component of scientific reasoning, which we have not yet discussed, is that of explanation and argumentation. The process of explaining requires individuals to not only verbalize their knowledge but also to organize it in a way that allows for clear communication (De Vries, Lund, & Baker, 2002). As we will later discuss, this process of coordinating knowledge can be beneficial to learning, by identifying places where the learner has gaps or discrepancies in her or his knowledge. Explanations are often thought of as a way to communicate what has happened and why, i.e., to describe causal relations (National Committee on Science Education Standards, 1996). Explanations can also be used in argumentation to support or attack a claim. In this way, explanation and argumentation often go hand in hand (Berland & Reiser, 2009)

Argumentation is an important process for gaining and sharing scientific knowledge (Budke & Meyer, 2015). Argumentation can be an internal process, in which an individual weighs evidence for or against a claim, but it can also take place in the form of writing or oral explanation (Kuhn, 2000). The purpose of arguments is to convince

someone that a claim is correct (Osborne, 2010). Mercier (2011) believes that reasoning developed specifically for the purpose of argumentation. He also makes a distinction between different levels of argumentation skills, ranging from simply providing an argument to specifically planning a complete argument that also takes into account potential counterarguments. Through argumentation, reasoners can make their implicit understanding of scientific reasoning, or the control of variables strategy, explicit (Edelsbrunner, 2017).

Children show argumentation abilities almost as early as they learn to talk, for example, claiming “Mine,” as a justification for taking a toy away from a sibling (Dunn & Munn, 1987; Kuczynski & Kochanska, 1990). Additionally, causal explanations are common around two to three years of age, increasing in frequency with age (Callanan & Oakes, 1992; Crowley et al., 2001; Wellman, Hickling, & Schult, 1997). Important factors to consider when evaluating the quality of argumentation is that most people do not come up with the best possible argument on the first try, people are more likely (and more motivated) to argue against a claim, arguments tend to get better when people are asked to elaborate (Mercier, 2011), and individuals are also motivated to provide explanations for evidence that is inconsistent with prior knowledge (Legare, Gelman, & Wellman, 2010).

Many of the studies described above use children’s verbal explanations (also referred to as arguments or justifications) as a measure of a more developed, explicit understanding of the control of variables strategy. They often found that these abilities are developing in the later elementary school years, though with much variability, and are affected by many of the same task factors as described earlier.

For example, Bullock and Ziegler (1999) found that over half of 4th grade children who chose controlled tests could also provide a justification referring to controlling variables and this ability increased to almost all of 6th graders being able to do this. Croker

and Buchanan (2011) asked children (four- to ten-year-olds) to provide explanations for their selected test of a hypothesis and categorized them as either evidence- or theory-based explanations. They found an effect of age both for whether or not children provided any explanation and whether those explanations were evidence-based.

Similarly, Amsel and Brock (1996) found that justifications based on prior beliefs accounted for the majority of explanations from 3rd and 6th graders, but that evidence-based justifications increased with age. Bullock (1991) found that though a majority of 2nd and 3rd graders could correctly judge the causal status of a variable based on covariation data presented to them, very few could justify their decision from the data. She concluded that it is this ability to justify (on the basis of data), rather than basic scientific reasoning abilities, that is developing around this age.

Chen and Klahr (1999) also investigated children's (2nd to 4th grade) ability to justify their design of controlled experiments through the use of probing questions asking children why they designed their experiment in the way they did or how they knew their experiment was conclusive. About 15% of children could provide a CVS-related justification for their controlled test design before training. Chen and Klahr did not report whether age was a factor in the ability to provide a justification.

Recall also that Köksal-Tuncer and Sodian (2018) presented children with a false hypothesis and showed children confounded evidence to support this false hypothesis to investigate if children would spontaneously argue against the false claim and use evidence to disprove it. Indeed, a majority of four- to five-year-old children were motivated to argue against the experimenter's false claim.

In addition to being used as a measure of a more robust, explicit understanding of the control of variables strategy, argumentation or explanation has also been shown to affect learning. Some studies have shown that the act of generating explanations can

improve causal reasoning and facilitate generalization in preschool-aged children (Legare, 2014; Legare & Lombrozo, 2014; Wellman, 2011; Williams & Lombrozo, 2010).

Explaining may increase attention to or engagement with a problem and thus encourage deeper thought about the underlying mechanisms or reasons (Chi, 2009; Walker, Lombrozo, Legare, & Gopnik, 2013). Legare and Lombrozo (2014) propose that the act of explaining allows individuals to practice reasoning scientifically and to form connections between new ideas and their prior knowledge. Explanation further encourages generating predictions or hypotheses. For example, Rittle-Johnson, Saylor, and Swygert (2008) have found that five-year-olds benefit from self-explanation, specifically after they were first given feedback regarding the correct solution and were then asked to explain. Thus, the process of explaining could serve to draw attention to the feedback they received, which children must, in turn, reformulate.

The ability to argue or explain is, of course, related to children's verbal knowledge. In a study investigating the relation between verbal and non-verbal knowledge and children's explanations, Edelsbrunner (2017) found that at young ages (1st to 3rd grade) children's level of non-verbal knowledge is separate from their level of verbal knowledge, while in higher grades (4th to 6th) the two types of knowledge covary perfectly and are not separable. This finding explains how younger children can recognize or design controlled tests without yet being able to verbalize their reasoning or their understanding of the control of variables strategy. Thus, both knowing and being able to explain the strategy is more advanced than knowing and being able to apply it. In this study, Edelsbrunner (2017) found that, although overall verbal abilities were low in the early grades and developing over time, there were three 1st graders who achieved the maximum verbal knowledge scores, suggesting that though these abilities show development with age, there are also clearly individual differences in children's abilities.

In summary, children's explanation and argumentation can be used as a measure of their explicit understanding of aspects of scientific reasoning as well as their ability to verbalize that understanding. Further, the process of explanation itself can be beneficial to scientific reasoning by helping to organize and coordinate knowledge prior to explaining or during the process of explaining. Finally, the ability to explain scientific processes is clearly developing later in childhood and is related to language abilities and thus, also affected by individual differences.

1.4 Chapter Summary

Research has shown, on the one hand, that children (and adults) struggle with many of the components of scientific reasoning. Their interpretation of evidence and their recognition and design of experiments is influenced by the prior beliefs (Amsel & Brock, 1996; Carey et al., 1989; Croker & Buchanan, 2011; Klahr et al., 1993; Kuhn et al., 1988; Schauble et al., 1991; Tschirgi, 1980). They selectively use valid experimentation strategies to determine the causes of negative outcomes but not positive outcomes (Carey et al., 1989; Croker & Buchanan, 2011; Kuhn & Phelps, 1982; Schauble, 1990; Tschirgi, 1980; Zimmerman & Glaser, 2001). Their performance can also be affected by the content of the tasks, in some cases performing better on more constrained, "scientific" tasks (Linn et al., 1983) and in other cases performing better on tasks about "everyday" content (Song & Black, 1992). They can also be influenced by the design aspects of a task, as well as differing levels of difficulty of the task requirements (Bullock & Ziegler, 1999; Saffran et al., 2016; Taylor & Dana, 2003). One conclusion from such studies is that children do not have the necessary ability to distinguish between theories and evidence that is so critical for scientific reasoning (Kuhn & Franklin, 2007). But a second possibility is that, because of some of the task factors mentioned above, these studies do not appropriately assess

young children's abilities, and instead highlight their struggles with tasks measuring scientific reasoning.

In some cases, we have evidence of children's nascent abilities in scientific reasoning. For example, four- and five-year-olds can interpret covariation evidence, particularly when it covaries perfectly (Amsel & Brock, 1996; Koerber et al., 2005; Piekny et al., 2014; Piekny & Maehler, 2013). They understand that evidence can be the basis for conclusions or hypotheses (Ruffman et al., 1993). First and 2nd graders have shown the ability to distinguish between testing a hypothesis and producing an effect and also recognizing whether a test is conclusive or not (Sodian et al., 1991). Four- and five-year-olds can generate and revise hypotheses and use unconfounded evidence to argue against a false hypothesis (Köksal-Tuncer & Sodian, 2018). Third graders can select contrastive tests and 4th graders can select controlled tests (Bullock & Ziegler, 1999). Four- and five-year-olds spontaneously generate controlled tests during exploratory play (van Schijndel et al., 2015), and they can design controlled tests with up to four variables when provided with support and feedback (van der Graaf et al., 2015).

Thus, these studies show that even without training, children already have beginning scientific reasoning abilities, specifically regarding the control of variables strategy, and even preschoolers show some ability to design multivariable experiments when supported (for a detailed review see e.g., Zimmerman, 2000, 2007; Zimmerman & Klahr, 2018). Novel tasks that do not allow for any prior beliefs and simpler tasks that require limited prior knowledge could be used to further investigate unadulterated scientific reasoning abilities in younger children. In addition to the development of scientific reasoning abilities, the ability to verbalize the reasoning process is developing in the preschool years and up through early elementary school when it becomes inseparable

from non-verbal knowledge (Edelsbrunner, 2017). Thus, it is possible that even younger children would succeed in tasks that assess scientific reasoning non-verbally.

The fact that some studies show scientific reasoning abilities in children as young as four or five years of age brings us into the same age range as causal reasoning studies showing that children are sensitive to the informativeness of evidence, for example, they recognize when evidence is confounded and that, as a result, there is the potential for information gain (Cook et al., 2011; Gweon & Schulz, 2008; Schulz & Bonawitz, 2007). In those studies, children played and explored more with toys that had generated confounded evidence and also showed novel information-seeking behaviors such as isolating variables.

It seems that a critical difference between children's causal reasoning abilities and their scientific reasoning abilities is the ability to coordinate theory and evidence, which requires the recognition that a theory may be wrong and can or should be revised, and that evidence may support or help falsify a theory (Kuhn, 1989, 2002; Kuhn & Franklin, 2007). Additionally, the processes of generating and testing hypotheses are likely implicated in the distinction. All of these require reflection and thus rely on metacognitive abilities, which are developing around five years of age (Bullock et al., 2009; Perner, 1991; Sodian & Bullock, 2008; Wimmer & Perner, 1983).

So, though children may naturally or spontaneously show behaviors indicative of some of the processes needed for scientific reasoning, such as isolation or even control of variables (Cook et al., 2011; van Schijndel et al., 2015), the question remains if they intentionally seek knowledge, by generating and testing particular hypotheses during exploration. Abilities in causal reasoning likely serve as building blocks for scientific reasoning, but the development of metacognitive abilities and the metaconceptual

understanding of theory and evidence is what is missing and necessary for scientific reasoning.

1.5 The Present Studies

This thesis aims to investigate preschool children's abilities to reason scientifically, as measured by understanding of the Control of Variables Strategy. To date, there has been extensive research on children's abilities in using CVS, and on improving those abilities, however, there is limited research on preschool children's CVS abilities. In one study described above, we saw that children as young as four spontaneously used CVS to produce controlled experiments during exploratory play (van Schijndel et al., 2015). However, there was no evidence that children intentionally generated and tested hypotheses to determine what factors were important for shadow size. Additionally, children had specific prior beliefs that only size, but not distance, was a relevant factor in determining shadow size, thus, their prior beliefs likely affected their ability to reason scientifically.

In a second study, we saw that a majority of four- to six-year-olds could design both contrastive experiments and controlled experiments with two variables, and that some could even generate controlled experiments with up to four variables (van der Graaf et al., 2015). However, in this study, the researchers assessed CVS ability "dynamically," providing explicit feedback after each experiment attempt about whether the experiment was designed correctly or not and why or why not. Thus, it is unclear how well preschool children would perform on such a task without such extensive feedback. Further, the ramps context is quite complex both in content and in having to manipulate the variables. Consequently, there is a need to further investigate preschooler's abilities in CVS with a task that eliminates prior beliefs, content knowledge, difficult task requirements, and

assesses “pure” reasoning abilities without providing added feedback or support from the experimenter.

The causal reasoning literature presents a potential solution to these requirements. First, the causal reasoning literature has shown us that preschool children and even infants are sensitive to the informativeness of evidence (Gweon & Schulz, 2008; Schulz & Bonawitz, 2007), they use covariation evidence to make accurate causal inferences (Gopnik et al., 2001; Schulz & Gopnik, 2004), they infer causal relations according to evidence even when those relations conflict with their prior beliefs (Kushnir & Gopnik, 2007), and they spontaneously perform informative exploratory behaviors (Cook et al., 2011); all abilities that are potential building blocks for scientific reasoning. Second, the causal reasoning literature has developed a paradigm that reduces the influence of prior beliefs and prior knowledge and has extensively used this paradigm with infants and preschool-aged children: the blicket detector paradigm (Gopnik & Sobel, 2000).

We developed a novel, knowledge-lean task using the blicket detector paradigm to investigate preschool children’s abilities in CVS. Importantly, the task included components critical to the distinction between causal reasoning and scientific reasoning: the intent to find out, the testing of hypotheses, and goal-directed experimentation, as well as an additional task assessing their metacognitive understanding of the informativeness of evidence. Specifically, the CVS task requires children to select a controlled test after being presented with a hypothesis about the light effect of the blicket detector. There are two versions of the task: with two or three variables. We additionally asked children to justify their selection and to, finally, interpret the results of their test. We also presented children with confounded evidence of a light effect and asked them if they could know the cause or not (Interpretation of Confounded Evidence task (ICE)).

We used this novel task to assess the stability of CVS abilities in a test-retest study with preschoolers (Study 1), to conduct a more robust assessment of preschooler's abilities in CVS (Studies 2a & 2b), and to investigate adults' abilities in CVS (Study 3). We also used this novel task to examine how scientific reasoning abilities in preschool relate to other developing cognitive abilities (Study 4). Finally, we focused on the iterative design and development of materials for promoting CVS abilities in preschoolers (Studies 5a, 5b, & 6) and again used the novel CVS task to assess the effectiveness of a video tutorial in improving children's abilities (Study 6).

2 Preschoolers' (and Adults') Scientific Reasoning

Understanding causality, the relation between cause and effect, is a critical skill for making sense of the world. This skill is so essential that it appears to develop in very early childhood with even infants showing precocious causal reasoning capacities. As reviewed in greater detail in Chapter 1, infants can reason about caused-motion interactions, agents' goal-directed actions, and covariation information (Muentener & Bonawitz, 2018). Preschool children can use causal information to categorize events and objects (Gopnik & Sobel, 2000; Nazzi & Gopnik, 2003; Schulz et al., 2008), reason about counterfactual events (Harris et al., 1996), register conditional independence among events and use covariation information to make accurate inferences (Gopnik et al., 2001; Schulz & Gopnik, 2004; Sobel & Kirkham, 2006), appreciate the ambiguity of confounded evidence (Cook et al., 2011, Schulz & Bonawitz, 2007; Sodian et al., 1991), and can effectively intervene on causal systems (Cook et al., 2011; Gopnik et al., 2001; Gweon & Schulz, 2008). Many of these skills already available to preschool children resemble the skills required for scientific reasoning, for example, a sensitivity to the informativeness of confounded and unconfounded evidence as well as the potential for gaining information through isolation of variables (Cook et al., 2011; Gweon & Schulz, 2008; Schulz & Bonawitz, 2007).

Scientific reasoning includes the process of evidence evaluation and determining cause-effect relations, similarly to causal reasoning, among other activities such as problem identification, questioning, hypothesis generation, construction of artifacts, evidence generation, evidence evaluation, drawing conclusions, and communicating and scrutinizing scientific reasoning and its results (Fischer et al., 2014). However, despite similarities to causal reasoning, research on scientific reasoning has traditionally found

that young children's abilities are limited, and that scientific reasoning only begins to develop in adolescence (e.g., Inhelder & Piaget, 1958; Kuhn et al., 1988; Kuhn et al., 1995). It has been argued that this disparity between precocious causal reasoning abilities and limited scientific reasoning abilities in preschool children is a result of their inability to distinguish between theory and evidence (Kuhn, 2002; Kuhn & Franklin, 2007).

Another distinction between causal reasoning and scientific reasoning is the metacognitive awareness of the motivation to explore confounded evidence. This intentional knowledge-seeking process is considered a requirement for engaging in scientific reasoning (Sodian, 2018).

On the one hand, children's scientific reasoning abilities have been shown to be limited. For example, children do not systematically test hypotheses and fail to control variables and ignore or distort evidence that does not support their prior beliefs (Kuhn et al., 1988). On the other hand, there is some evidence that children's abilities have been underestimated (Zimmerman, 2007), with studies showing, for example, that elementary school children can select a conclusive test of a hypothesis (Sodian et al., 1991) or even a controlled experiment (Bullock & Ziegler, 1999), and that preschoolers can generate controlled experiments when provided with repeated feedback on experimental design (van der Graaf et al., 2015). In addition, as reviewed in Chapter 1, there are many factors related to the tasks used to assess scientific reasoning that can influence its measurement, such as the content or difficulty of the task, as well as children's own prior beliefs or knowledge about the tasks.

Considering the similarity between causal reasoning and scientific reasoning and the findings of precocious causal reasoning abilities in preschool and beginning abilities in scientific reasoning in elementary school, we sought to investigate preschoolers' scientific reasoning abilities. Specifically, we were interested in preschoolers' understanding of the

Control of Variables Strategy (CVS). CVS is a key component of scientific reasoning, which requires reasoners to manipulate only the variable in question to test a hypothesis while keeping all other variables constant (Tschirgi, 1980). Keeping non-focal variables constant results in an unconfounded experiment from which valid causal inferences can be made (Chen & Klahr, 1999).

When children are asked to produce controlled experiments to learn a causal structure, they do not appear to use CVS until adolescence (Inhelder & Piaget, 1958; Kuhn et al., 1995; Schauble, 1996). Other findings, however, suggest that young children might have an understanding of CVS. For example, Chen and Klahr (1999) showed that elementary school-aged children can use the Control of Variables Strategy when they are explicitly taught how to use it. More recently, van der Graaf and colleagues (2015) showed that even preschoolers show some understanding of CVS when provided with support and feedback to produce unconfounded experiments. This evidence indicates that CVS can be taught to young children, but leaves open the question if children can use CVS spontaneously to test hypotheses.

Regarding the influence of the task itself on the assessment of scientific reasoning abilities, factors such as prior knowledge or beliefs about the task content, as well as the difficulty of producing or recognizing a controlled test, have been shown to impact children's abilities. For example, Bullock and Ziegler (1999) found that 4th graders could choose which of two experiments was unconfounded, while even adults struggled to produce a controlled experiment. When children do not have prior knowledge about a task, particularly in unfamiliar decontextualized systems, they seem to make more rational inferences than adults (Lucas et al., 2014). In light of these findings, we developed a novel knowledge-lean CVS task that required children to select a controlled test of a hypothesis with either two or three variables. We used the "blicket detector" paradigm (Gopnik &

Sobel, 2000) to control for prior knowledge in reasoning by creating a learning environment in which the researcher knows what prior knowledge the child brings to the task (Sobel & Munro, 2009). “Blickets” are objects with an invented causal power of making a box light up or play music. Children observe interactions with the blicket detector and can then discover causal structures or determine whether objects are blickets.

To investigate if preschool children can select an unconfounded experiment, we showed children evidence, a set of bricks makes the box light up, and provided a hypothesis, that one of the bricks was the cause of the light effect. We then presented children with a number of options that could be placed on the box to find out if the specified brick was indeed the cause of the effect. Children could only select one option to test and were asked to provide a reason for their choice of that option. We presented children with 2-variable and 3-variable versions of this task to vary the difficulty of selecting the correct, controlled, test. In another task, we showed children that a set of bricks made the box light up and asked if they could know which of the bricks were the cause of the light effect. With this task, we aimed to assess an additional metacognitive understanding of what one knows, which is also critical to developing mature scientific reasoning abilities.

We presented these tasks to children aged three to six. These children have not yet entered into the formal education system and, thus, should not have received any formal STEM education. In this way, we can obtain a clear picture of the development of scientific reasoning abilities in young children before they begin to learn about science in school. This chapter presents three studies with preschool children and one study with adults. In Study 1, we examined children's performance on these scientific reasoning tasks using a test-retest strategy. Children performed these three tasks in one session and then were tested again two weeks later, with the same tasks but different materials, to

determine whether their performance was stable. In this way, we tested both the causal reasoning capacities related to scientific inference, but also whether those capacities represented stable reasoning abilities within the child. Study 2a replicated and extended the first procedure, modifying certain aspects of the way the questions were asked. Children performed two trials of each of the three tasks in one session. In Study 2b, we further replicated and extended the procedure, additionally looking at how well children could learn from the experiments they designed. In Study 3, we conducted the same procedure as Study 2b with an adult sample. We specifically looked at adults' justifications of their experiment designs and evaluations of the evidence produced by their experiments.

2.1 Study 1: Stability of Preschoolers' Scientific Reasoning Abilities

2.1.1 Method

2.1.1.1 Participants

The final sample consists of 60 children ($M_{\text{age}} = 59.88$ months, $SD = 7.77$; median = 59.20 months; range: 44 - 78 months; 28 girls). Five additional children were tested but excluded due to experimenter error (3), or unwillingness to participate (2). All participants were typically developing children from a large German city. Parental informed consent and child assent were obtained for all children. Sample size was determined through power analysis based on a linear regression assuming a fixed model with $\alpha = .05$, $\beta = .20$ and a medium-to-large effect size ($f^2 = .25$) based on Cohen (1992).

2.1.1.2 Materials

The machine was a custom-built wooden box (30 x 20 x 15 cm) with an LED strip around the top that was controlled by the experimenter via a foot pedal. Eighty Lego Duplo bricks in 30 unique colors and patterns were used as “lighters” or “non-lighters.” Seven sets of bricks were used (see Figure 2.1, Panels A-G): four individual bricks were used for familiarization and training (A), two sticks with four bricks each were used to assess children's interpretation of confounded evidence (B & E), two sets of three sticks with two bricks each were used in the 2-variable task (C & F), and four sets of four sticks with three bricks each were used in the 3-variable task (D & G). The bricks in all sticks were glued together so that they could not be separated.

A cardboard tray measuring 16 x 21 cm was used to present children with test choices. The task materials, the location of the correct choices (Left, Middle, or Right), and the order of the CVS tasks were counterbalanced. We will use the first order to

illustrate the procedure (shown in Figure 2.1). Pictures of all materials can be found in Appendix A.

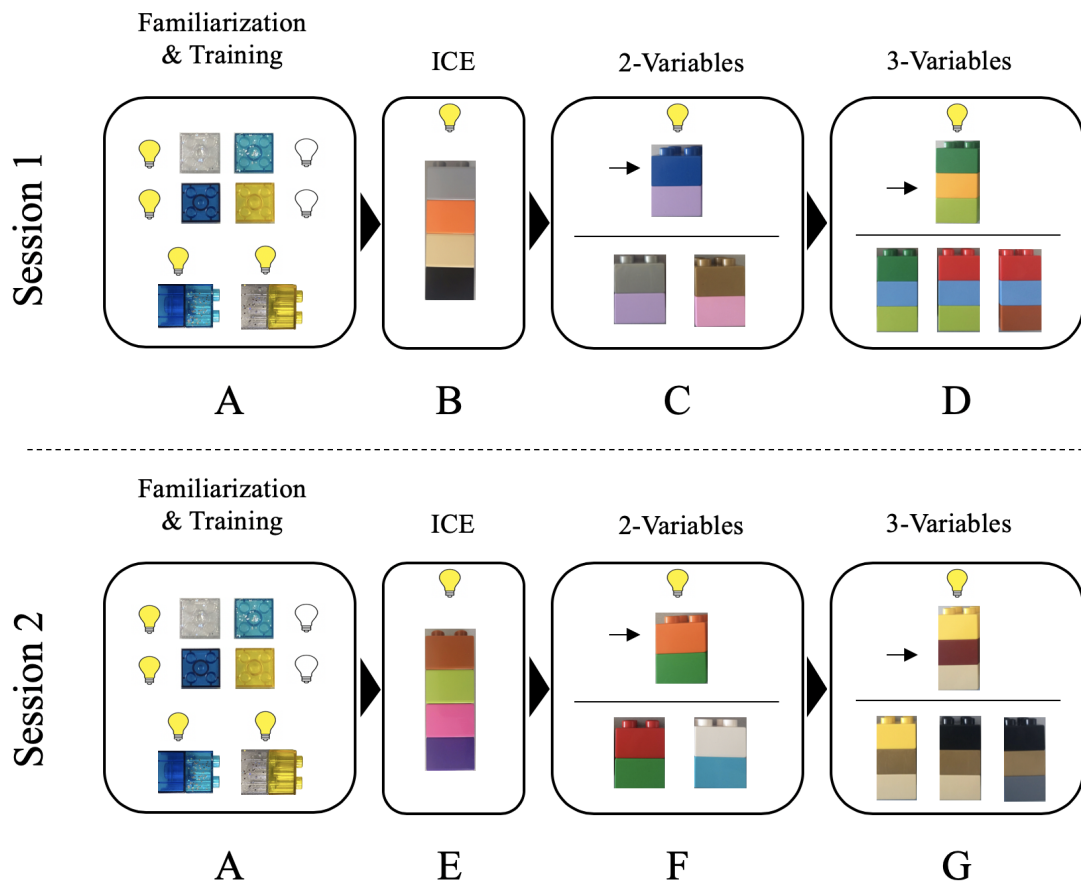


Figure 2.1. Materials and an example of the testing procedure for Study 1. ICE: Interpretation of Confounded Evidence task. Yellow light bulb indicates that the box lit up when the corresponding object was placed on it. Arrows indicate the brick in question (X).

2.1.1.3 Procedure

Data for this study were collected between July - August 2017 and November 2018 - February 2019. Each testing session took approximately 15 minutes. All individual sessions took place in local kindergartens¹ in separate, quiet rooms and were video

¹ Kindergarten (for children between 3–6 years) is not a part of the regular public school system in Germany. It is neither mandatory, nor free. Tuition is normally based on income.

recorded. Children sat at a table with a female experimenter. The experimenter first administered a matching puzzle game in order to familiarize the children with the testing environment as well as a color vision test to ensure that children could discern among the colors used in the procedure. Children were not included in the final sample if they failed the color vision test. In this sample, no child was excluded for this reason. The full protocol of the experiment can be found in Appendix B. At the outset of the experiment, the experimenter introduced the lightbox. Children were taught that some bricks make the box light up and some did not. Children were also taught that combinations of bricks made the box light up as long as there was a lighter present (Figure 2.1, Panel A).

Interpretation of Confounded Evidence Trial. Next, children observed a stick of four bricks. This stick was placed horizontally on the box, so all bricks touched the box, and the box activated (See Figure 2.1, Panel B). Children were asked if they knew which of the bricks in this stick make the box light up. They were also asked if they were certain or not, and to provide an explanation for their answer.

Control of Variables Trials. Children then received two types of CVS trials. In the 2-variable trial, children watched the experimenter place a stick of two bricks on the lightbox, which activated (Figure 2.1, Panel C). For the purposes of this demonstration, we will refer to the two brick colors as X and Y. The experimenter pointed to the top brick and said, “We want to find out if this brick (the X brick) makes the box light up.” The stick was placed in front of the child. Two additional sticks were then placed on the table. One (the correct choice) swapped the X brick with a novel color (Z) but kept the Y brick (so the stick was Z and Y). The other (the incorrect choice) swapped both bricks, resulting in a stick of two novel colors (P and Q). Thus, if the original stick was blue and purple and children were tasked to find out whether the blue brick made the box light up, the correct choice was a silver and purple stick and the incorrect choice was a gold and pink stick, see

Figure 2.1, Panel C. The experimenter presented the choices and explained, “You can pick one of these sticks to place on the box to find out if the X brick makes the box light up. Which stick do you want to pick?” The children then indicated their choice and the experimenter asked, “Why did you pick this stick?” Finally, the experimenter placed the chosen stick horizontally on the box. The box did not light up. Children were asked to interpret the evidence generated by their experiment. The experimenter asked, “Now do you know if the X brick made the box light up?”

The procedure for the 3-variable trial was the same as for the 2-variable trial, except that children were shown that a stick of three bricks made the box light up and were asked to find out if the middle brick made the box light up (Figure 2.1, Panel D). Children were given three sticks as choices. The correct choice varied the brick in question and kept the other two bricks the same; a second stick varied the brick in question as well as an additional brick; the third stick varied all three bricks, resulting in a stick with three novel colors.

Children received one ICE trial, one 2-variable trial, and one 3-variable trial in the first session (S1). In the second session (S2), two weeks later, children repeated the familiarization and completed one trial each of the ICE task and the 2- and 3-variable CVS tasks (Figure 2.1, Panels E-G), using different stimuli.

2.1.1.4 Coding

Verbal responses to the ICE task were coded in two steps. First, responses to the two questions - whether children knew which bricks make the box light up and whether they were sure of their response - were combined to create two categories of knowledge claims. Responses that indicated that children did not know which bricks made the box light up were coded as correct. This category includes children who claimed they did not know in the first question and children who first claimed they did know, but then indicated

they were not actually sure. The second category was an incorrect claim of knowledge and included responses that indicated children knew which bricks made the box light up and were sure of their knowledge claim.

Next, we coded children's explanations for how they knew or why they did not know. Any explanation for how they knew was coded as incorrect. Explanations for why they did not know were coded as correct or incorrect. A correct explanation indicated that children could not know which bricks made the box light up because the bricks were stuck together and could not be isolated and tested individually. Examples are shown in Table 2.1. We defined children as generating a robust ICE response if they generated a correct explanation for a correct knowledge claim.

Choices in the 2- and 3-variable CVS tasks were coded as correct or incorrect, based on whether children picked the appropriate stick. Justifications for choices were coded as relevant or irrelevant to CVS. Justifications for a controlled test were considered relevant if they referred to (1) the absence of the X brick in the choice, the presence of one (2) or two (3) of the control bricks in the choice or (4) if they referred to both the absence of the X brick and the presence of the control brick(s). For example, a child who answered "this brick is the same as this brick" while pointing to the control brick on the test stick and the control brick on the choice stick would be coded as providing a 'one control comparison' justification (refer to Table 2.2 for additional detailed examples of justifications). Other justifications, such as color preference, were not considered relevant. Additionally, we defined children as generating a robust response if they generated a valid justification for a correct choice.

The responses provided to the interpretation of evidence question were unable to be interpreted. Children generally responded 'Yes' or 'No,' but in coding it became clear that with 'Yes' children could mean 'Yes, I know it made the box light up' or 'Yes, I

know it didn't make the box light up.' Similarly, with 'No,' children could mean 'No, I don't know if it made the box light up' and 'No, it didn't make the box light up.' This issue is addressed in Study 2b.

All videos were coded by the author and an independent rater coded 20% of the data. Agreement for children's choice behavior was 94% ($Kappa = .87$). Agreement for verbal responses was 96% ($Kappa = .88$). In cases of disagreement, agreement was reached by a discussion of the two raters.

2.1.2 Results

In the ICE task, we analyzed whether children responded in a way that indicated an understanding of the inconclusiveness of evidence by answering that they did not know which bricks made the box light up. These data are shown in Figure 2.2. As a preliminary analysis, we built a Generalized Estimating Equation (GEE) with an independent working correlation matrix, a binomial distribution, and a cumulative logit link function (Zeger & Liang, 1986; Zeger, Liang, & Albert, 1988) looking at the role of gender and task materials on children's knowledge claim responses. Neither of these factors were significant (p -values $> .09$). As a result, these factors will not be considered further.

Next, we constructed a GEE to control for within-subject responses examining children's knowledge claim responses, looking at the role of age and session. Neither of these factors were significant, both p -values $> .19$. Fifty-three percent of responses correctly indicated a lack of knowledge of which bricks made the box light up ($S1 = 58\%$; $S2 = 49\%$). Taking into account children's explanations for their knowledge claims, we constructed a GEE to control for within-subject responses examining children's robust performance on the ICE task, looking at the role of age and session. Neither of these factors were significant, both p -values $> .17$. In the first session, 15% of children showed a robust understanding of confounded evidence by correctly responding they could not

know which bricks made the box light up because the bricks were stuck together. In the second session, 18% of children responded in this way. Seven percent of children responded consistently correctly and 75% responded consistently incorrectly, $Kappa = .31$, $p = .02$.

Table 2.1

Examples of Children's Correct and Incorrect Explanations for the Interpretations of Confounded Evidence Trial (Answers to "Can you know which bricks are lighters?").

Correct	Incorrect
I can't know because:	I don't know
The bricks are stuck together	I can't know because:
I can't try them out	These are different bricks
It could be any of the bricks	I haven't seen these bricks before
I haven't seen which ones light up	No one told me
I can't try them one at a time	I know because:
	It's yellow light the sun
	My mom told me
	Because they sparkle
	Because they made the box light up
	I have a book about them
	I'm a big kid
	I think so
	They are pretty

Note. Robust ICE Responses were indicated by children correctly saying that they did not know when block activated the machine and generating a correct justification.

Table 2.2.

Examples of Children's Justifications for their Test Selection (Answers to "Why is this the best stick to find out if the X brick makes the box light up?")

Relevant Justifications	
Example	Description
This brick is different from the X brick	Contrastive comparison
This stick doesn't have the X brick	Contrastive comparison
This stick also has this (control) color	One Control comparison
These two bricks are the same as those two bricks (controls)	Two Controls comparison
This brick is the same as that brick (control) and this brick is the same as that brick (control)	Two Controls comparison
These sticks are the same, but it doesn't have this (X) brick	CVS: Contrastive & Control comparison
Only this brick (X) is different	CVS: Contrastive & Control comparison
Irrelevant Justifications	
Example	Description
I don't know; It just is; My mom told me	Knowledge claim
I like this one; it's pretty; these look nice	Preference
I picked the other one last time	Strategy
Let's try it; We haven't tried it yet	Strategy
It is a lighter; maybe it lights up	Identity / Effect production
It is not a lighter	Identity
Because it is similar to the test stick	Comparison / Effect production

Note. Robust responses to the CVS trial indicated children chose the correct stick to test and generated a relevant justification.

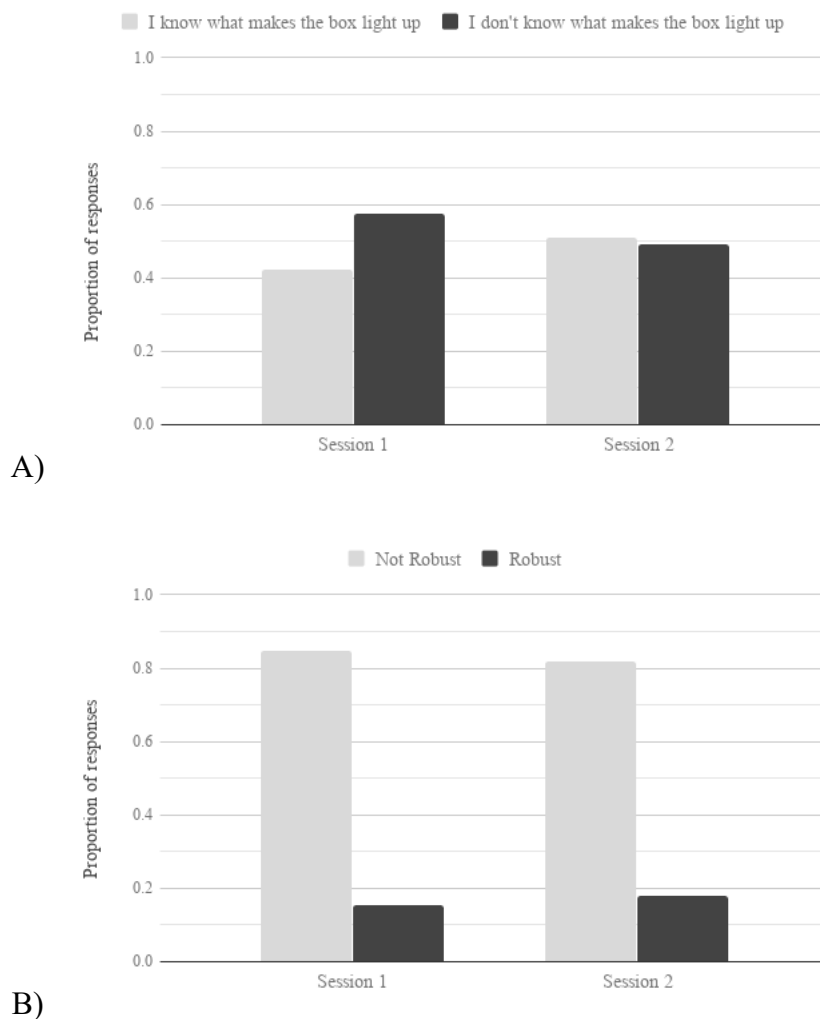


Figure 2.2. Children's performance on the interpretation of confounded evidence task. A) Children's knowledge claims about the effectiveness of the individual bricks. B) Children's robust performance: providing a relevant explanation for why they can't know which bricks make the box light up.

Next, we analyzed whether children chose the response that indicated a controlled experiment (CVS tasks). These data are shown in Figure 2.3. As a preliminary analysis, we built a GEE looking at the role of gender, the order in which children received the tasks, the task materials, and the location of the correct choice on children's responses. All of these factors were not significant (p -values $> .19$). As a result, these factors will not be considered further.

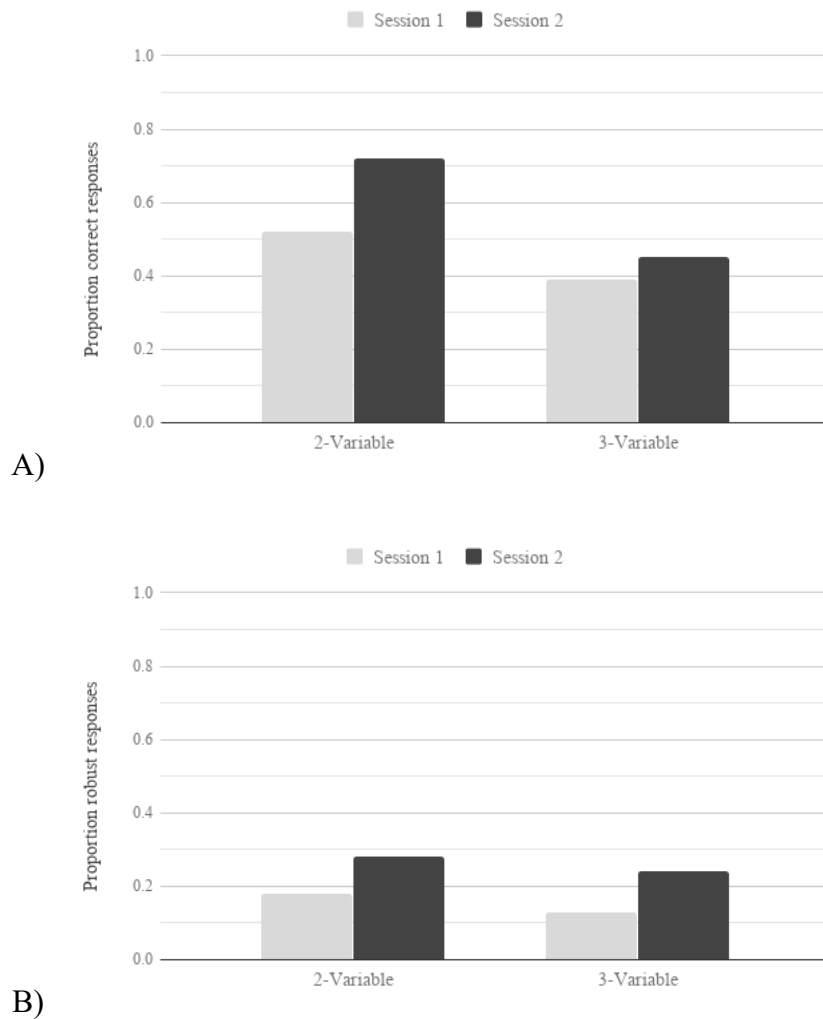


Figure 2.3. Performance on the CVS tasks. A) Choice Performance (2-var chance = 50%, 3-var chance = 33%), B) Robust performance.

For our main analysis, we constructed a GEE to control for within-subject responses examining whether children chose the response that indicated a controlled experiment, looking at the role of age, task (i.e., 2-variable vs. 3-variable), session, and performance on the ICE task during that session. This model revealed a main effect of session, $B = -0.88$, $SE = 0.26$, [95% CI = -1.39, -0.37], Wald $\chi^2(1) = 11.26$, $p = .001$, and a main effect of age, $B = 0.03$, $SE = 0.02$, [95% CI = 0.00, 0.07], Wald $\chi^2(1) = 3.81$, $p = .05$. Specifically, older children in our sample were more likely to select the correct choice than younger children. Children performed better in the second session than in the first session.

Performance in the 2-variable task was not stable, Kappa = .12, $p = .31$, with performance improving from session 1 to session 2, McNemar $\chi^2(1, N = 60) = 4.65$, $p = .03$. Children selected the controlled test at chance level during the first session (52%; $\chi^2(1) = 0.07$, $p = .80$, Cohen's $w = 0.03$), and improved to above-chance levels in the second session (72%; $\chi^2(1) = 11.95$, $p = .001$, Cohen's $w = 0.45$). The results of the 3-variable task between the two sessions was inconclusive. Performance between the two sessions was not stable, Kappa = .20, $p = .13$, but children did not significantly differ in their performance between the two sessions, McNemar Chi-Squared (1, $N = 60$) = 0.41, $p = .52$. Children selected the controlled test at chance level in the first session (39%; $\chi^2(1) = 0.71$, $p = .40$, Cohen's $w = 0.11$), and marginally above chance level in the second session (45%; $\chi^2(1) = 3.68$, $p = .06$, Cohen's $w = 0.25$).

We were also interested in children's justifications for their selections in the CVS trials. Twenty-three percent of justifications were considered relevant to CVS. We constructed a GEE to control for within-subject responses examining children's justifications, looking at the role of age, task, session, choice, and performance on the ICE task during that session. The model revealed a main effect of age, $B = 0.07$, $SE = 0.03$, [95% CI = 0.02, 0.12], Wald $\chi^2(1) = 6.37$, $p = .01$, and a main effect of choice, $B = -2.98$, $SE = 0.76$, [95% CI = -4.48, -1.48], Wald $\chi^2(1) = 15.23$, $p < .001$. Specifically, older children in our sample were more likely to provide relevant justifications than younger children. Children provided relevant justifications significantly more often for correct responses (36% of the time) than for incorrect responses (1% of the time). Finally, we contrasted the percentage of time children generated a correct response on the CVS trial and generated a relevant justification for that choice across the two sessions. On both the 2-variable and 3-variable CVS tasks, this performance was stable between the two

sessions (2-var: $Kappa = .39, p = .004$; 3-var: $Kappa = .36, p = .01$). A summary of the results can be found in Appendix C.

2.1.3 Discussion

Children were presented with confounded evidence and asked (1) to recognize that they cannot know something conclusively and (2) to explain that the reason they cannot know is because the evidence is confounded. This task proved difficult for children in this sample, with almost half of children claiming knowledge of what makes the box light up. These children did not recognize that, because all potential causes were on the box simultaneously and could not be isolated, they could not, in fact, know which bricks make the box light up. However, approximately half of children did respond in a way that suggests they recognized their inability to know what makes the box light up conclusively. About a third of these children were able to provide a reason for why they could not know, referring to the bricks being stuck together and unable to be tested individually. Taking into account children's explanations, performance across the two sessions was stable: children who were able to explain why they could not know in the first session tended to do so in the second session as well.

The low performance in providing an explanation for not knowing is surprising when one considers the procedure of this task. When introducing the stick of four bricks, the experimenter says, "These are stuck together. We can't take them apart," while emphasizing the stuckness. It is notable, then, that so few of the children who say they cannot know what makes the box light up are able to provide this explanation referring to the fact that the potential causes cannot be isolated. The difficulty of this task may lie in the metacognitive processes required for producing a correct response. The ability to recognize what one does not know has been shown to be difficult for children under the age of five (Rohwer et al., 2012). However, we found no age trends for this task in this

sample. The question was also phrased in a way that may have been biased toward positive responses: children were asked if they knew which bricks made the box light up. To address this issue, we changed the wording in Study 2a; instead of asking if they knew, children were asked if they could know for sure which bricks made the box light up, or if they could not know for sure.

Children's ability to select a 2-variable controlled test improved from the first to the second session, with performance increasing from chance to above-chance levels. This increase in performance occurred without feedback on any of the tasks, which is in line with studies showing that repeated interactions with systems for designing experiments improved children's performance in designing controlled tests (Kuhn, 2007b; Schauble, 1990). Children seem to have more difficulty recognizing a 3-variable controlled test: there was no improvement between sessions and children were only marginally above chance in the second session. We also found a developmental trend in children's selection of a test, with older children better able to recognize a controlled test.

The below chance performance in each task type in the first session could also be due to the wording of the choice question, while in the second session, their experience from the first session could have made the wording less of an issue. Children were asked which stick they would like to pick, which may have prompted them to choose based on preference. We address this possibility in Study 2a by placing more emphasis on the testing of a hypothesis; instead of asking which stick they would like to pick, they were asked which stick was the best to find out if the X brick made the box light up. Changing the emphasis of the test question could improve children's performance to S2 levels at the beginning.

Taking into account both children's choice and their justification for that choice, performance was stable across the two sessions, with children who show a robust

understanding in the first session tending to do so in the second session. However, overall, children generated few relevant justifications (23% of all responses). This low performance could be due to the wording of the justification question. Children were asked why they picked a particular stick, which may have elicited preference-based answers. We address this possibility in Study 2a by placing more emphasis on the reasoning behind why children chose a particular test; instead of asking why they picked a particular stick, they were asked why the stick they picked was the best to find out if the X brick made the box light up.

2.2 Study 2a: Preschoolers' Scientific Reasoning Abilities

2.2.1 Method

2.2.1.1 Participants

The final sample consists of 51 children ($M_{\text{age}} = 65.30$ months, $SD = 10.25$; median = 66.8 months; range: 45 - 82 months; 24 girls). Two additional children were tested but excluded due to experimenter error. All participants were typically developing children from a large German city. Parental informed consent and child assent were obtained for all children. Sample size was determined by a similar power analysis as in Study 1.

2.2.1.2 Materials

The lightbox was the same as in Study 1. Forty-four Lego Duplo bricks in 30 unique colors and patterns were used as Tomas (lighters) or not Tomas (non-lighters) (see Figure 2.4, Panels A-G). Seven sets of bricks were used: four individual bricks were used for familiarization and training (A), two sticks with four bricks each were used to assess children's interpretation of confounded evidence (B & G), two sets of three sticks with two bricks each were used in the 2-variable task (C & E), and two sets of four sticks with three bricks each were used in the 3-variable task (D & F). The bricks in all sticks were glued together so that they could not be separated. Bricks were never repeated outside their set.

A cardboard tray measuring 16 x 21 cm was used to present children with test choices. A clear plexiglass cover, measuring 17 x 24 x 8 cm, was placed over the choices to prevent children from grabbing for the options before hearing the critical questions. Eight different testing versions were created to counterbalance the task materials, the location of the correct choices (Left, Middle, or Right), and the order of the CVS tasks (as

shown in Figure 2.4 or starting with the 3-variable task and alternating). We will use the first order to illustrate the procedure (shown in Figure 2.4).

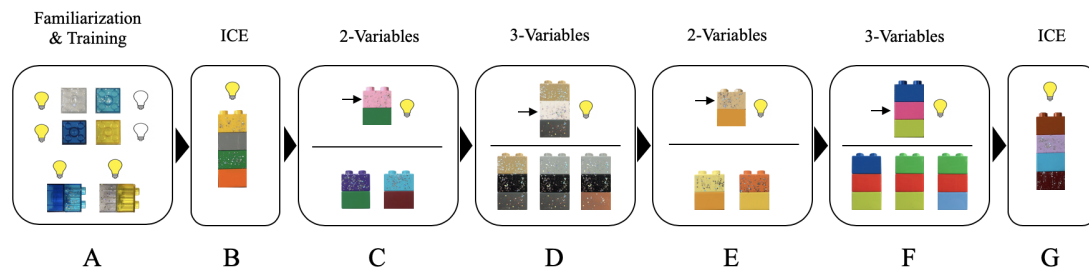


Figure 2.4. Materials and an example of the testing procedure for Study 2a. ICE: Interpretation of Confounded Evidence task. Yellow light bulb indicates that the box lit up when the corresponding object was placed on it. Arrows indicate the (X) brick in question.

2.2.1.3 Procedure

Data for this study were collected between February - April 2018. The procedure was the same as in Study 1 with the following changes (the changes to the protocol can be found in Appendix D, with the changes highlighted). Unlike Study 1, children performed two trials of each task type in one session. The session began with an ICE task, followed by two 2-variable and two 3-variable CVS tasks in alternating order and then finished with a second ICE task. We labeled bricks that made the box light up Tomas and labeled bricks that did not make the box light up not Tomas. This was done to make the wording of the critical questions shorter and less cumbersome. Additionally, when presenting children the CVS task choices, we placed a clear plexiglass cover over the choices, to prevent children from grabbing a choice before the experimenter had a chance to ask the critical questions.

In the training phase with the four introductory bricks, we assigned effects to each of the bricks rather than having the first and either third or fourth brick make the box light up. We made this change to show children that one glitter brick made the box light up and one did not and also that one blue brick, the dark blue one, made the box light up and the

other blue brick, the light blue one, did not. In this way, we could show children that neither glitter nor color could be relied upon to determine whether a brick made the box light up or not. In the familiarization with the combined bricks, we added a combination showing children twice that a lighter and a non-lighter would make the box light up.

In the ICE task, the procedure was the same, but the wording of the question was changed to place more emphasis on knowing for sure and to provide the alternative of not knowing for sure. Children were specifically asked, “Can you know for sure which of the bricks are Tomas or can you not know for sure?” In the 2- and 3-variable CVS tasks, the procedure again was the same, but the wording of the questions was changed to place more emphasis on the fact that the selection of a choice was for the purpose of finding something out. Children were specifically asked, “Which stick is the best to find out if the X brick is a Toma?” and “Why do you think this stick is the best to find out if the X brick is a Toma?”

2.2.1.4 Coding

Coding of the responses to the ICE and 2- and 3-variable CVS tasks was the same as in Study 1. All videos were coded by the author. We also counted the length of children's justifications in terms of number of words (MLUw). This provided a gross measure of children's language production. Additionally, in Study 2a, we also look at children's linguistic capacities by analyzing the length of their justifications, a gross measure of language ability, which might have influenced whether they can generate appropriate justifications.

The responses provided to the interpretation of evidence question were unable to be interpreted. Children generally responded ‘Yes’ or ‘No,’ but in coding it became clear that with ‘Yes’ children could mean ‘Yes, I know it made the box light up’ or ‘Yes, I know it didn’t make the box light up.’ Similarly, with ‘No,’ children could mean ‘No, I

don't know if it made the box light up' and 'No, it didn't make the box light up.' This issue is addressed in Study 2b.

2.2.2 Results

In the ICE task, we analyzed whether children responded in a way that indicated an understanding of the inconclusiveness of evidence by answering that they did not know which bricks were Tomas. These data are shown in Figure 2.5. As a preliminary analysis, we built a GEE with an independent working correlation matrix, a binomial distribution, and a cumulative logit link function (Zeger & Liang, 1986; Zeger et al., 1988) looking at the role of gender and task materials on children's knowledge claim responses. Both of these factors were not significant, all p -values $> .06$. As a result, these factors will not be considered further.

Next, we constructed a GEE to control for within-subject responses examining children's knowledge claim responses, looking at the role of age and trial on performance on the ICE task. This model revealed a main effect of trial, $B = .69$, $SE = 0.32$, [95% CI = 0.06, 1.31], Wald $\chi^2(1) = 4.64$, $p = .03$. Taking into account children's explanations for their knowledge claims, we constructed a GEE to control for within-subject responses examining children's robust performance on the ICE task, looking at the role of age and trial. This model revealed a main effect of trial, $B = 1.34$, $SE = 0.43$, [95% CI = 0.49, 2.18], Wald $\chi^2(1) = 9.56$, $p = .002$. In the first trial, 37% of children showed a robust understanding of confounded evidence by correctly responding they couldn't know which bricks made the box light up because they were all stuck together. In the second trial, 14% responded in this way. Seventy percent of children provided a correct knowledge claim response at least once out of two trials and 40% gave a robust ICE response at least once out of two trials.

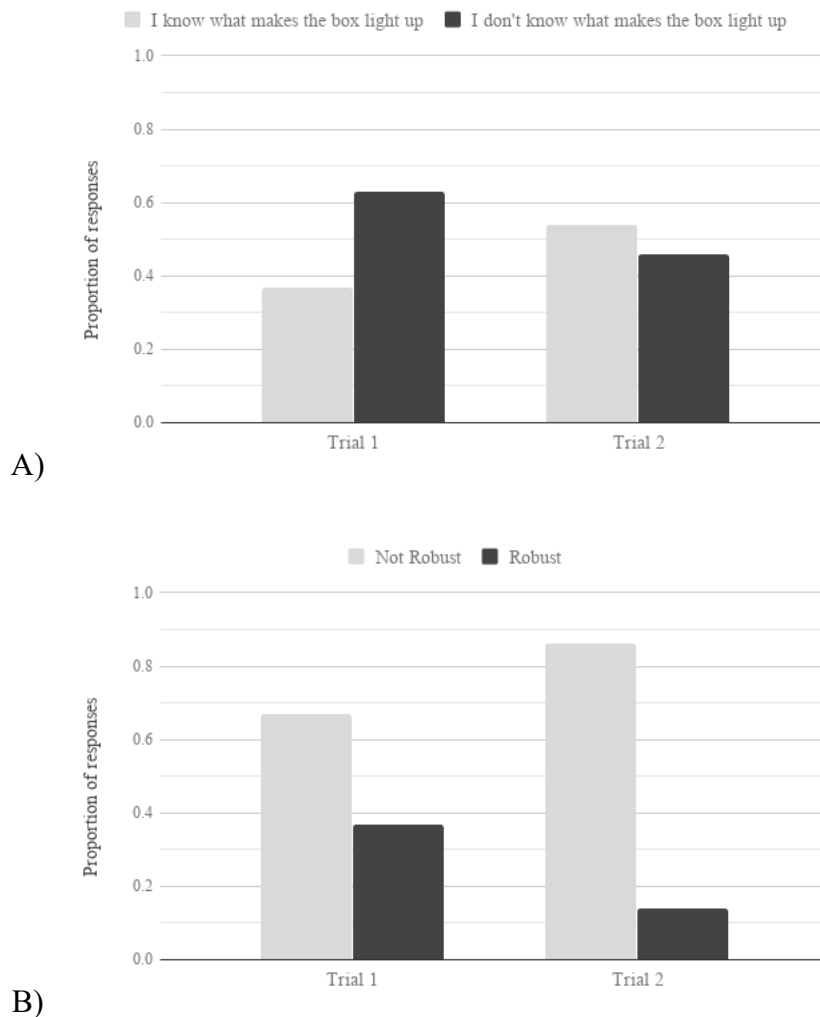


Figure 2.5. Children's performance on the interpretation of confounded evidence task. A) Children's knowledge claims about the effectiveness of the individual bricks. B) Children's robust performance: providing a relevant explanation for why they can't know which bricks make the box light up.

Next, we analyzed whether children chose the response that indicated a controlled experiment in the CVS tasks. These data are shown in Figure 2.6. As a preliminary analysis, we built a GEE looking at the role of gender, the order in which children received the tasks, the task materials, and the location of the correct choice on children's responses. All of these factors were not significant (p -values $> .25$). As a result, these factors will not be considered further.

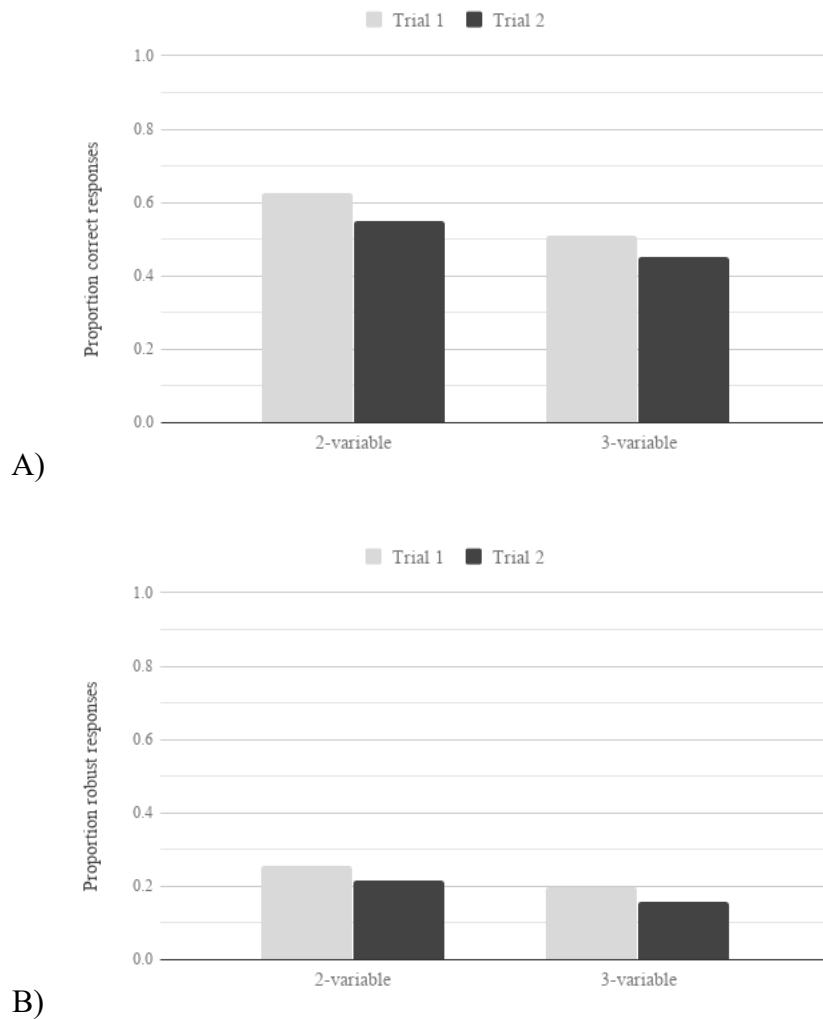


Figure 2.6. Performance on the CVS tasks A) Choice performance (2-var chance = 50%, 3-var chance = 33%). B) Robust performance.

For our main analysis, we constructed a GEE to control for within-subject responses examining whether children chose the response that indicated a controlled experiment on the CVS tasks, looking at the role of age, task (i.e., 2-variable vs. 3-variable), trial, and performance on the first trial of the ICE task. All of these factors were not significant (p -values $> .11$).

In the first trial of the 2-variable task, 63% of children chose the controlled test, no different than expected by chance, $\chi^2(1) = 3.31, p = .07$. Across two trials, 35% of the children selected the correct choice on both trials, 47% selected the correct test on one of the trials, and 17% of children selected the incorrect test on both trials. This distribution

was not different than expected by chance, $\chi^2(2) = 3.35, p = .19$, Cohen's $w = 0.26$.

Overall, 82% of children selected a controlled test with 2-variables in at least one trial.

In the first trial of the 3-variable task, 51% of children chose the controlled test, greater than expected by chance, $\chi^2(1) = 7.15, p = .01$. Across the two trials, 24% of the children selected the correct test twice, 49% selected the correct test once, and 27% of children selected the incorrect test twice. This pattern of performance was significantly different from chance, $\chi^2(2) = 10.63, p = .01$, Cohen's $w = 0.46$. Overall, 72% of children selected a controlled test with 3-variables in at least one trial.

We were also interested in children's justifications for their selections in the CVS trials. Twenty-six percent of their justifications were relevant to CVS. We constructed a GEE to control for within-subject responses examining children's justifications, looking at the role of age, task, trial, choice, performance on the first trial of the ICE task, and MLUw (to control for linguistic factors that might have contributed to children's ability to generate a justification). The model revealed a main effect of age, $B = 0.05, SE = 0.03$, [95% CI = 0.00, 0.10], Wald $\chi^2(1) = 3.89, p = .05$, a main effect of choice, $B = -1.83, SE = 0.47$, [95% CI = -2.75, -0.92], Wald $\chi^2(1) = 15.39, p < .001$, and a main effect of MLUw, $B = 0.36, SE = 0.11$, [95% CI = 0.14, 0.58], Wald $\chi^2(1) = 9.92, p = .002$. Specifically, older children in our sample were more likely to provide relevant justifications than younger children. Children provided relevant justifications more often for correct responses (39%) than for incorrect responses (12%). Finally, relevant justifications ($n = 53; M = 5.87 (1.32)$ words) were significantly longer than irrelevant justifications ($n = 151, M = 4.53 (2.29)$ words).

2.2.3 Discussion

Study 2a extended the procedure used in Study 1 while presenting children with slightly different test questions on both the ICE and CVS tasks designed to facilitate scientific reasoning.

In the first trial of the ICE task, more children could explain that they could not know which bricks made the box light up because the evidence is confounded (robust ICE) than in Study 1. By the second trial, though, children regressed to a level of performance similar to that of Study 1. The change in wording may have resulted in better robust performance in the first trial, but the longer duration of the experiment and the increased number of trials may have contributed to this improvement being lost by the end of the experiment. As in Study 1, there was no relation between children's age and their knowledge claims or their explanations.

In the 2-variable task, children performed marginally better than expected due to chance in the first trial, better than the first session of Study 1 but not as well as the second session of Study 1. Performance across two trials was not different from chance. In the 3-variable task, children performed better than expected due to chance in the first trials, better than the second session of Study 1. Their performance across two trials was also better than expected due to chance. This pattern is the reverse from what one would expect, that the 2-variable task should be easier than the 3-variable task. It is possible that the difficulty of the 3-variable task, both an increase in the number of variables and an increase in the number of choices, requires children to think more deeply about their choice, leading them to succeed, whereas with the 2-variable task, they may simply select a test without having thought too much about their choice.

Unlike in Study 1, children's age did not predict their performance on the selection of controlled tests in Study 2a. It is possible that in Study 1 older children were better able

to interpret the less-than-ideal wording of the questions to still select the correct choice, while in Study 2a, the improved wording resulted in all children understanding the task better. Then, differences in performance could come down to individual differences between children rather than age-related differences.

Children's ability to provide relevant justifications for their choice of test was similar to that in Study 1. Thus, it appears that the change in wording did not benefit children's justifications. We replicated the main effects of age and choice on children's justifications and additionally found a main effect of MLUw. Mean length of utterance in words was uniquely related to children's ability to provide relevant justifications, an effect which becomes clear when inspecting the difference between relevant and irrelevant justifications: relevant justifications were longer on average than irrelevant justifications. In Study 2b, we made one additional modification to the protocol to address the issue of the interpretation question in the CVS tasks.

2.3 Study 2b: Replication & Extension of Study 2a

In the previous two studies, the interpretation asked children, “Now do you know if the X brick makes the box light up?” Responses to this question were unclear, as children sometimes responded only yes, but then added an explanation that implied the X brick was not a lighter. Thus, we could not be sure if their response was indicative of their knowledge (Yes, I know; No, I don’t know) or of the brick’s category (Yes, it’s a lighter; No, it’s not a lighter). To address this issue, we reworded the question and provided children with three possible responses, “Is the X brick a lighter, not a lighter, or can you not know?”

2.3.1 Method

2.3.1.1 Participants

The final sample consists of 57 children ($M_{\text{age}} = 65.12$ months, $SD = 9.24$; median = 66.8 months; range: 41 - 81 months; 29 girls). Five additional children were tested but excluded due to color vision problems (1) or experimenter error (4). All participants were typically developing children from a large German city. Parental informed consent and child assent were obtained for all children. Sample size was determined by a similar power analysis as in Study 1.

2.3.1.2 Materials

The lightbox was the same as in Studies 1 and 2a. The same materials were used as in Study 2a.

2.3.1.3 Procedure

Data for this study were collected between April - May 2018. The procedure was the same as in Study 2a with the following changes (the changes to the protocol can be found in Appendix E, with the changes highlighted).

In the familiarization with the combined bricks, we added a combination, showing children that all four bricks stuck together made the box light up, before asking them if they could remember which of those bricks made the box light up as in the previous studies. In the CVS tasks, after children observed their choice placed on the box and the resulting effect (the machine never activated, regardless of their choice), they were asked to interpret the evidence generated by their experiment. In the previous experiments, the interpretation question and resulting responses were unclear and could not be interpreted. To address this issue, in Study 2b, the experimenter asked, “Now do you know if the X brick is a Toma, is it not a Toma, or do you not know?” Children were also asked if they were certain or not and to provide an explanation for their answer.

2.3.1.4 Coding

Coding of the responses to the ICE and 2- and 3-variable CVS tasks was the same as in Study 2a. At the end of the CVS tasks, children were asked to interpret the evidence generated by their choice. These questions were coded depending on the child's initial choice. If children chose the correct brick to test, then we considered whether they claimed that they were certain that the X brick was a Toma. If they chose an incorrect brick to test, then we considered whether they were not sure that the X brick was a Toma. Examples of children's responses are shown in Table 2.3.

All videos were coded by the author and by an independent rater. Agreement for children's choice behavior was 99% (Kappa = .97). Agreement for verbal responses was 94% (Kappa = .83). In cases of disagreement, agreement was reached by a discussion of the two raters.

Table 2.3.

Examples of Children's Justifications for their Interpretations of Evidence Generated by the Experiment (Answers to "How do you know it is/is not a Toma?" or "Why don't you know?").

Relevant Interpretation Explanations		
Type of test	Example	Description
Controlled	This stick doesn't light up	The stick used as a test does not make the box light up.
	That stick lit up and this one doesn't.	Comparison between the original stick and the stick used as a test.
	These bricks aren't lighters/ don't light up, so it has to be the X brick.	Reference to the inefficacy of the control bricks and the inference that can be made as a result.
Confounded	I don't know because the bricks are stuck together.	Recognizes that the experiment was confounded, cannot draw conclusion.

2.3.2 Results

In the ICE task, we analyzed whether children responded in a way that indicated an understanding of the inconclusiveness of evidence by answering that they did not know which bricks were Tomas. These data are shown in Figure 2.7. As a preliminary analysis, we built a GEE with an independent working correlation matrix, a binomial distribution, and a cumulative logit link function (Zeger & Liang, 1986; Zeger et al., 1988) looking at the role of gender and task materials on children's knowledge claim responses. Both of these factors were not significant, all p -values $> .34$. As a result, these factors will not be considered further.

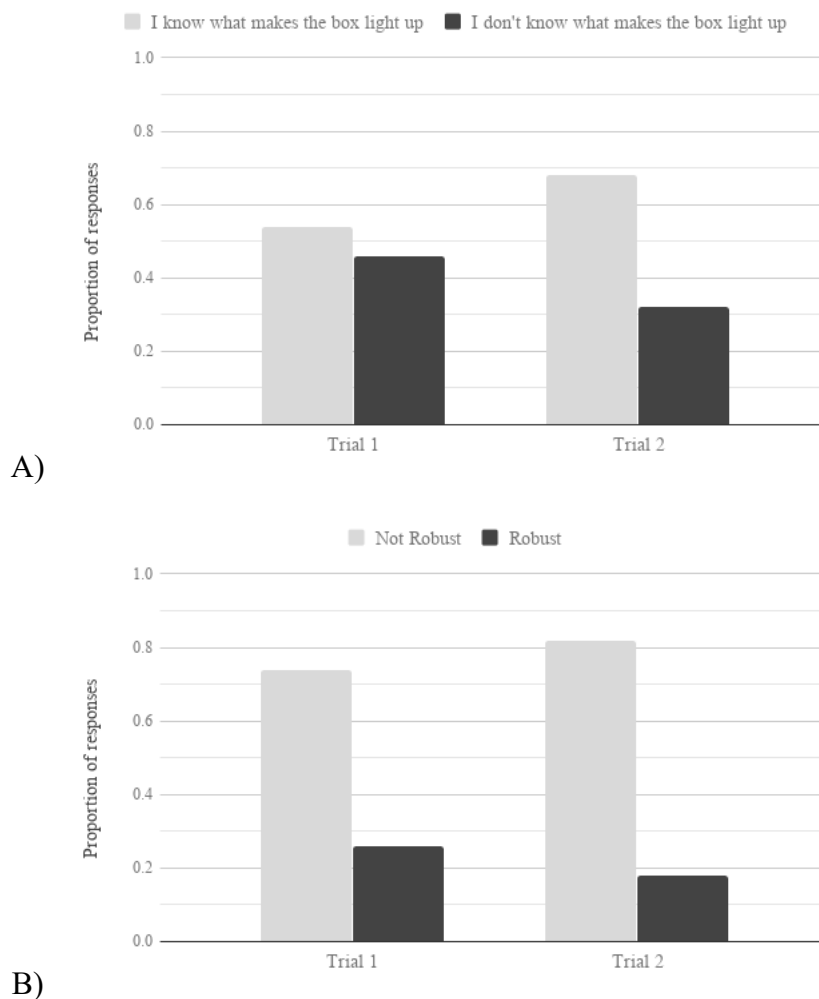


Figure 2.7. Children's performance on the interpretation of confounded evidence task. A) Children's knowledge claims about the effectiveness of the individual bricks. B) Children's robust performance: providing a relevant explanation for why they can't know which bricks make the box light up.

Next, we constructed a GEE to control for within-subject responses examining children's knowledge claim responses, looking at the role of age and trial on performance on the ICE task. This model revealed a main effect of age, $B = 0.08$, $SE = 0.03$, [95% CI = 0.03, 0.14], Wald $\chi^2(1) = 8.10$, $p = .004$. Taking into account children's explanations for their knowledge claims, we constructed a GEE to control for within-subject responses examining children's robust performance on the ICE task, looking at the role of age and trial. This model revealed a main effect of age, $B = 0.11$, $SE = 0.03$, [95% CI = 0.05, 0.16], Wald $\chi^2(1) = 12.83$, $p < .001$, and a main effect of trial, $B = 0.59$,

SE = 0.30, [95% CI = 0.002, 1.18], Wald $\chi^2(1) = 3.87$, $p = .05$. Performance on Knowledge Claim responses in the first trial was not significantly different from that in the first trial of Study 2a (63% - 47%, $p = .11$), but performance in the second trial was significantly worse than that in Study 2a (46% - 32%, $p = .01$). In the first trial, 26% of children showed a robust understanding of confounded evidence by correctly responding they couldn't know which bricks made the box light up because they were all stuck together, significantly fewer than that in Study 2a (37% - 26%, $p = .02$). In the second trial, 18% responded in this way, not significantly different from the proportion in the second trial of Study 2a (14% - 18%, $p = .32$). About half of children provided a correct knowledge claim response at least once out of two trials, significantly fewer than in Study 2a (70% - 54%, $p = .003$), and 28% gave a robust ICE response at least once out of two trials, again significantly fewer than in Study 2a (40% - 28%, $p = .02$). Thus, the majority of children did not generate robust responses on the ICE measure, but older children were more likely to do so, particularly on the first trial.

Next, we analyzed whether children chose the response that indicated a controlled experiment in the CVS tasks. These data are shown in Figure 2.8. As a preliminary analysis, we built a GEE looking at the role of gender, the order in which children received the tasks, the task materials, and the location of the correct choice on children's responses. All of these factors were not significant (p -values $> .27$). As a result, these factors will not be considered further.

For our main analysis, we constructed a GEE to control for within-subject responses examining whether children chose the response that indicated a controlled experiment on the CVS tasks, looking at the role of age, task (i.e., 2-variable vs. 3-variable), trial, and performance on the first trial of the ICE task. This model revealed a main effect of task, $B = 0.96$, SE = 0.27, [95% CI = 0.43, 1.48], Wald $\chi^2(1) = 12.60$,

$p < .001$. In the first trial of the 2-variable task, 70% of children chose the controlled test, greater than expected by chance, $\chi^2(1) = 9.28, p = .002$, but not significantly different than performance in Study 2a (63% - 70%, $p = .12$). Across two trials, 46% of the children selected the correct choice on both trials, 40% selected the correct test on one of the trials, and 14% of children selected the incorrect test on both trials. This distribution was different than expected by chance, $\chi^2(2) = 13.49, p = .001$, Cohen's $w = 0.49$, but not significantly different than performance in Study 2a ($p = .55$). Overall, 86% of children selected a controlled test with 2-variables in at least one trial.

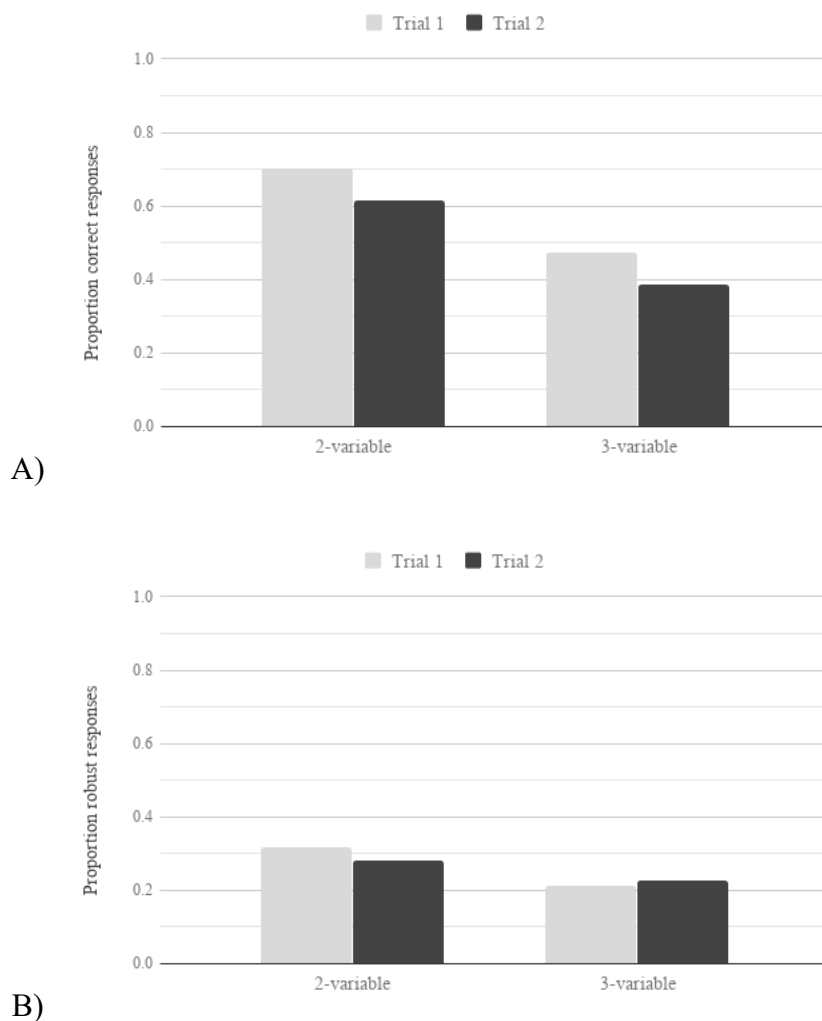


Figure 2.8. Performance on the CVS tasks. A) Choice performance (2-var chance = 50%, 3-var chance = 33%). B) Robust performance.

In the first trial of the 3-variable task, 47% of children chose the controlled test, greater than expected by chance, $\chi^2(1) = 9.74, p = .002$, and not significantly different than performance in Study 2a (51% - 47%, $p = .76$). Across the two trials, 17% of the children selected the correct test twice, 51% selected the correct test once, and 32% of children selected the incorrect test twice. This pattern of performance was not significantly different from chance, $\chi^2(2) = 4.78, p = .09$, Cohen's $w = 0.29$, and not significantly different than performance in Study 2a ($p = .86$). Overall, 68% of children selected a controlled test with 3-variables in at least one trial.

We were also interested in children's justifications for their selections in the CVS trials. Thirty-eight percent of their justifications were relevant to CVS, significantly more than in Study 2a (26% - 38%, $p < .001$). We constructed a GEE to control for within-subject responses examining children's justifications, looking at the role of age, task, trial, choice, performance on the first trial of the ICE task, and MLUw (to control for linguistic factors that might have contributed to children's ability to generate a justification). The model revealed a main effect of age, $B = 0.08, SE = 0.03, [95\% CI = 0.02, 0.14]$, Wald $\chi^2(1) = 6.10, p = .01$, a main effect of task, $B = -0.67, SE = 0.26, [95\% CI = -1.20, -0.17]$, Wald $\chi^2(1) = 6.87, p = .01$, a main effect of choice, $B = -1.55, SE = 0.34, [95\% CI = -2.22, -0.88]$, Wald $\chi^2(1) = 20.25, p < .001$, and a main effect of MLUw, $B = 0.28, SE = 0.09, [95\% CI = 0.10, 0.47]$, Wald $\chi^2(1) = 8.81, p = .003$. As in Study 2a, older children in our sample were more likely to provide relevant justifications than younger children, children provided relevant justifications more often for correct responses (50%) than for incorrect responses (23%), and relevant justifications ($n = 86; M = 5.99 (1.67)$ words) were significantly longer than irrelevant justifications ($n = 142, M = 4.62 (2.27)$ words). The new effect of task revealed that children provided relevant

justifications more often for the 3-variable CVS task (41%) than for the 2-variable CVS task (34%).

Our final analysis focused on children's interpretations of the evidence generated by their choice. Recall that after children chose which stick to place on the machine, that stick was placed on the machine and did not activate it. If children chose the correct stick, they should know that the X brick was a Toma; if children chose the incorrect stick, they are not able to conclude whether the X brick is a Toma or not and should instead indicate that they are unsure or cannot know. The distribution of children's responses is shown in Figure 2.9.

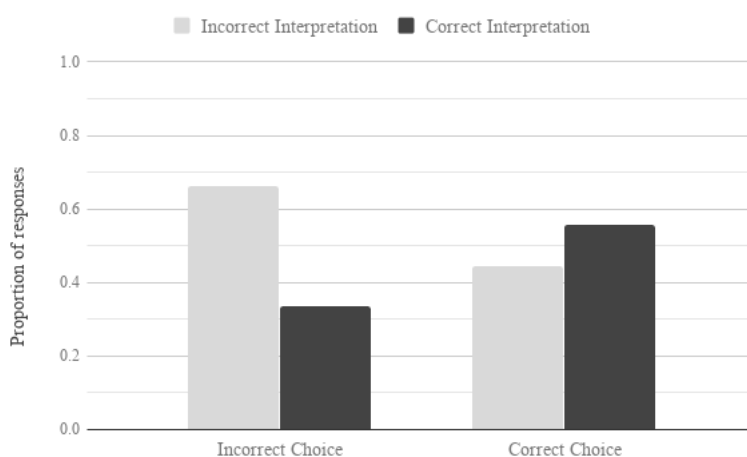


Figure 2.9. Interpretations of the outcome of the experiment (Is the X brick a Toma, not a Toma, or do you not know?)

We constructed a GEE to control for within-subject responses examining the correctness of children's interpretations, looking at the role of age, task, trial, performance on the first trial of the ICE task, and robust performance on the CVS tasks. The model revealed a main effect of robust performance on the CVS tasks, $B = -0.73$, $SE = 0.35$, [95% CI = -1.41, -0.04], Wald $\chi^2(1) = 4.32$, $p = .04$. Robust performance on a CVS task (i.e., making the correct choice and justifying it appropriately) uniquely predicted making a correct interpretation of the experiment rather than an incorrect interpretation.

2.3.3 Discussion

Study 2b extended the procedure used in Study 2a by presenting children with a slightly different interpretation question in the CVS tasks to address the ambiguity in the question and the resulting ambiguous responses.

In the first trial of the ICE task, fewer children could explain that they could not know which bricks made the box light up because the evidence is confounded (robust ICE) than in Study 2a. By the second trial, though, performance was similar to that of Study 2a. The lower performance in Study 2b as compared to Study 2a may be explained by an effect that was found for the first time in Study 2b, that children's performance on the ICE task was influenced by age. Older children were more likely to provide both correct Knowledge Claim responses and robust ICE explanations. Thus, the worse performance could be due to younger children in the present sample performing worse. There was one change to the protocol that may have influenced this result regarding the ICE task. In the present study, we added a brick combination during the training, showing children that all four bricks stuck together made the box light up, before asking them if they could remember which of those bricks made the box light up as in the previous studies. It is possible that this last combination and the question whether children could remember which of those bricks made the box light up affected younger and older children differently in the subsequent ICE task: younger children may have thought they were supposed to have learned something from the training and thus, should be able to identify which bricks in the ICE task made the box light up, while older children may instead have been able to recognize that they couldn't know for sure which bricks in the ICE stick made the box light up.

Performance on both the 2-variable and 3-variable CVS tasks in the present study did not differ significantly from performance in Study 2a. However, performance on the 2-

variable CVS task in the present study was significantly better than expected due to chance in both the first trial and across two trials, while that was not the case in Study 2a. For the 3-variable task, while performance across two tasks was significantly better than expected due to chance in Study 2a, it was not in the present study. It is interesting that the present study showed the opposite pattern of performance, that the 2-variable task was easier than the 3-variable task, though this is indeed what we would have expected in the previous study as well. This is consistent with the literature suggesting that the difficulty of control of variables tasks increases with an increasing number of variables (Tschirgi, 1980). It is further consistent with information processing models that suggest that the number of uncertain variables children are able to keep in mind is developing during this period (Beck, Robinson, Carroll, & Apperly, 2006; Erb & Sobel, 2014; Fernbach, Macris, & Sobel, 2012; Sobel, Erb, Tassin, & Weisberg, 2017). We again found no effect of children's age on their performance on the selection of controlled tests, as in Study 2a.

Looking at children's justifications for their choice of test, we saw a significant increase in relevant justifications from Study 2a to Study 2b. Though no changes to the protocol were made regarding children's justifications, it is possible that the change to the interpretation question that was asked after each trial had a beneficial effect of facilitating children's justifications. We replicated the main effects of age, choice, and MLUw on children's justifications and additionally found a main effect of task. Interestingly, though children selected 2-variable controlled tests more often than 3-variable controlled tests, children provided relevant justifications for 3-variable controlled tests more often than for 2-variable controlled tests. This could be because a controlled test in the 3-variable task contained two control bricks (as opposed to just one in the 2-variable controlled test), which may have increased attention to the similarities between the test stick and the choice stick, leading children to provide a relevant justification referencing the control bricks.

Finally, in the present study, children who provided relevant justifications for their controlled tests in the CVS tasks, by referring to contrastive and/or control variables, were also more likely to interpret the data they observed correctly to come to the appropriate causal inference.

2.4 Interim Discussion

The present studies measured preschoolers' use of the Control of Variables strategy in order to learn about a novel causal system. Three to six-year-olds, who have not yet entered formal schooling, showed a nascent understanding of controlled experiments with both 2- and 3-variables. Children recognized that ambiguous data was ambiguous and selected appropriately controlled tests to determine particular hypotheses about the causal structure. Older children could also provide relevant justifications for their choices. These findings support recent advances in the literature suggesting that preschoolers have some understanding of Control of Variables Strategy (van der Graaf et al., 2015) and that recognizing an unconfounded test is easier than producing one (Bullock & Ziegler, 1999).

This broad and systematic investigation of children's CVS abilities revealed early competence in recognizing controlled tests for diagnostic inference. In all three studies, children selected controlled tests with both 2- and 3-variables marginally or significantly more often than expected due to chance. Our findings are in line with previous findings showing that preschoolers can select a conclusive test of a hypothesis (Koerber & Osterhaus, 2019; Piekny & Maehler, 2013) and that they have an early capacity for using the Control of Variables Strategy (van der Graaf et al., 2015) even without any instruction or support.

These findings further suggest an important difference between children's causal and scientific reasoning. Scientific inference in the laboratory often involves reasoning

through many potential confounding variables. Rarely are experiments so clear cut as to be able to isolate a single causal factor. It seems possible that the building blocks of scientific inference are these causal reasoning capacities, and as the complexity of the inference increases via the number of possible variables one can vary, the more difficult the inference becomes.

The complexity of real-world scientific inference suggests an important role for play in learning. Preschoolers can learn causal structure from exploratory play (Schulz, Gopnik, & Glymour, 2007; Sobel & Sommerville, 2010). Preschoolers also engage in more exploratory play when faced with ambiguous evidence (Schulz & Bonawitz, 2007) or when they need to confirm explanations (Legare, 2012). Given the confounded nature of everyday situations, such play might be a necessary component in learning causal structure. Moreover, given that young children might not have an understanding of determinism (Bullock, Gelman, & Baillargeon, 1982), not all exploratory play would necessarily look rational. Repetitive behaviors and irrelevant actions - hallmarks of children's play - might help children determine what variables to control or ignore for subsequent learning.

Independent of age (and MLUw in Studies 2a & 2b), whether children select the correct test relates to their ability to provide a relevant justification: children are more likely to generate a relevant justification if they choose the unconfounded action to learn the causal structure. There are a few different ways to interpret this result. One possibility is that children choose correctly because they understand the underlying strategy of controlling variables (as measured by their generating a relevant justification). But another possibility is that seeing both the test stick and the choices simultaneously encourages comparison among the stimuli, resulting in the selection of the correct choice and consequently the relational inference and justification.

The results of the interpretation of confounded evidence task suggest that children are less competent in recognizing that they cannot know something conclusively because the evidence is confounded. In this task, children were asked an open-ended question that required them to reflect on their own knowledge and what conclusions they could make from observed evidence. This task relies on metacognitive processes and was difficult for children in this sample. Alternatively, the CVS tasks first presented children with evidence and then asked them to select a good test, using a forced-choice procedure. Performance in these tasks suggests that children can recognize a controlled test of a hypothesis on the basis of previously presented evidence.

Previous findings on young children's metacognitive awareness of the inconclusiveness of confounded evidence showed that three-quarters of five- to six-year-olds could correctly respond that they could not know something on the basis on confounded evidence and 40% of children could provide an explanation for their knowledge claim by referring to the confounded evidence in at least one of three trials (Köksal-Tuncer, Sodian, & Legare, 2019). In Study 2b, approximately half of children made a correct knowledge claim and 28% provided a valid explanation in at least one of two trials. Looking only at the five- to six-year-olds, two-thirds of the children made a correct knowledge claim and 43% provided a valid explanation in at least one of two trials. Taken together, these results seem to suggest that, at least by the age of five, children can recognize the uninformative nature of confounded evidence. This is especially meaningful considering that the ICE task in the present study was likely more difficult in a number of ways. The confounded evidence in the ICE task consisted of four candidate causes, whereas the confounded trials in Köksal-Tuncer et al. (2019) consisted of two candidate causes. As with the CVS tasks, it is likely that the difficulty of the ICE task increases with an increasing number of variables. Further, the bricks in the ICE task were

glued together, creating what could be interpreted as a new object. This could have made it even harder to recognize that we asked about the four individual bricks as potential causes, rather than the whole stick. Finally, the fact that the ICE task showed a developmental trend while the CVS tasks did not, as well as the lack of effect of ICE on performance on the CVS tasks, may suggest that these abilities are separate components of scientific reasoning.

The present findings, that young children show early competence in using the Control of Variables Strategy, raise the question of why CVS seems to be so difficult to master throughout development and into adulthood (Schwichow et al., 2016; Zimmerman, 2007). Bullock and Ziegler (1999) discussed that producing a controlled experiment is more difficult than recognizing a controlled experiment and that having selected a controlled test does not guarantee that an individual can also produce a controlled test. The increased cognitive effort involved in actively manipulating variables and designing two different tests seems to greatly limit individual's abilities to use CVS. Individuals' prior beliefs also often interferes with their ability to design controlled tests and interpret evidence: children have been shown to test hypotheses only when evidence is consistent with their beliefs about causal mechanisms (Croker & Buchanan, 2011), they ignore or distort evidence that conflicts with prior beliefs (Kuhn et al., 1988), and adults allow their beliefs to influence their interpretation of results (Kuhn, 2007a). Further, the majority of CVS research has been conducted in the domains of science or everyday-life contexts (e.g., ramps, pendulums, cake-baking, etc.; Chen & Klahr, 1999; Inhelder & Piaget, 1958; Tschirgi, 1980). Both prior beliefs and domain-specific content knowledge could get in the way of the basic abilities present already at the preschool age. Future research could further investigate CVS abilities throughout development using this novel, context-lean task.

To conclude, the present studies suggest that preschoolers possess nascent, but fragile, understanding of scientific inquiry, particularly in their ability to select controlled interventions to learn causal relations. Although children in the present sample showed a lot of coherence in their responses, performance on the present measures leave much room for improvement. This highlights that the complexity of the causal inferences involved in scientific reasoning is important for children's capabilities. Children in the present studies were more capable of reasoning about interventions involving a smaller number of variables, suggesting that information processing demands limit children's reasoning abilities. As children enter formal schooling, these information processing abilities improve, as does their metacognitive abilities, which might make it easier for them to relate their understanding of what they do not know to specific implemental designs for controlled experiments. This might allow them to understand the control of variables strategy explicitly, particularly when instructed (as suggested by Chen & Klahr, 1999), and apply it to their everyday thinking.

2.5 Study 3: Adults' Scientific Reasoning Abilities

Though they generally perform better on scientific reasoning tasks than children do, adults also show many of the same weaknesses on measures of scientific reasoning and CVS and typically do not show full competence (i.e., 100% correct performance) (e.g., Klahr & Dunbar, 1988; Klahr et al., 1993; Kuhn, 1989, 1991; Kuhn et al., 1988, 1995; Masnick & Klahr, 2003; Schauble, 1990, 1996; Schauble & Glaser, 1990). For example, Tschirgi (1980) found that adults used the VOTAT strategy to test hypotheses 55% of the time. In addition, their strategy usage was affected by the outcome: they were more likely to use a valid strategy when the outcome was bad (75%) than when the outcome was good (35%). Only 17% of adults fell into a category that represented consistent use of the VOTAT strategy.

The role of prior beliefs seems to play a critical role in adults' scientific reasoning abilities. Adults show biased interpretations on the basis of their prior beliefs, selecting evidence which confirms that belief (Kuhn et al., 1988). Their judgements of causality from covariation data are affected by prior beliefs (Amsel & Brock, 1996). In addition, adults are more likely to test hypotheses they believe to be plausible (88%) compared to hypotheses they believe are implausible (54%) (Klahr et al., 1993).

Kuhn and colleagues (1995) found that adults investigated about one-third of the full experiment space, and they allowed their prior belief to selectively direct their attention to variables they believed were causal (rather than noncausal). They initially provided valid inferences only 25% of the time, though this did increase to 70% over the course of ten weeks in a microgenetic study.

There are also differences in adults of different levels of education, for example, university students and college-educated adults tend to perform better than community college students or non-college educated adults (Klahr et al., 1993; Kuhn et al., 1988).

Schauble (1996) found that non-college-educated adults used CVS 50% of the time on a first task and improved to 63% on the second task. They made valid inferences approximately 70% of the time; however, their inferences were, as seen in studies above, also influenced by their prior beliefs about what variables would have an effect.

Amsel and Brock (1996) found that college students were better able to identify the causality of variables on the basis of covariation data than non-college-educated adults. They also provided more evidence-based justifications. However, all participants tended to judge causality based on prior belief, for example, that sunlight was causal, and a charm was noncausal for the health of plants, even when both variables were presented with the same evidence.

Kuhn (2007a) showed that adults, sampled from passengers waiting in a train station, were not able to make use of evidence presented to them regarding the effect of different entertainment features on ticket sales for a fundraiser. Specifically, adults were presented with data that supported a causal relation for only one of four features. However, 83% of adults judged that two or more of the features increased ticket sales, and most adults were “very certain” that their assessment was correct. Additionally, they made their judgement on the basis of their personal beliefs, for example, more often claiming that door prizes (83%) had an effect compared to costumes (33%) even though in reality, the data did not support either feature as causal. Kuhn also found that college-educated adults performed better than non-college-educated adults.

Adults are not consistent in their use of strategies, mixing valid and invalid strategies (Kuhn et al., 1995), they are also inconsistent in their predictions, making different causal attributions across consecutive predictions (Kuhn & Dean, 2004), and they do not always distinguish between theory and evidence in their justifications (Kuhn et al., 1995; Kuhn et al., 1992).

Further, Schauble (1996) found wide variability in performance, with some adults performing as poorly as most 5th to 6th graders. Kuhn and colleagues (1995) found similar results regarding variability, showing that one child even outperformed all the adults, and at the same time, some adults were at the same level as most children.

The results described above show that adults' performance in scientific reasoning and CVS is far from perfect. Their strategy use and judgements are affected by prior beliefs or the quality of the outcome. They are inconsistent in their use of strategies and in their justifications.

The finding that adults have knowledge of both valid and invalid strategies and mix their usage suggests that the development of mature scientific reasoning abilities is a gradual one and may never reach full maturity in every individual. The fact that individuals show inconsistent use of valid strategies also brings into question the value of single assessments, which may over- or underestimate abilities at any given time (Kuhn et al., 1995). This also makes it difficult to determine if variability is a result of different task contents or the individual's abilities. Finally, though there is clear progression in scientific reasoning abilities from the early years through to adulthood, the finding of wide variability within age groups suggests that individual differences likely play an important role.

The aim of Study 3 was to investigate adults' understanding of confounded evidence and the Control of Variables Strategy using the same knowledge-lean task originally developed for use with preschoolers. Though adults generally perform much better than young children in scientific reasoning and CVS assessments, the studies described above reveal that adults share many of the same struggles as children in regard to being influenced by prior knowledge and beliefs about the task content (e.g., Kuhn et al. 1988; Kuhn et al. 1995). Thus, we wanted to use the novel knowledge-lean task described

in Study 2a to investigate adults scientific reasoning abilities on tasks about which they should have no prior knowledge or beliefs.

2.5.1 Method

2.5.1.1 Participants

Participants were 40 students at a university in a large German city ($M_{\text{age}} = 24.95$ years, $SD = 3.95$; range: 18.95 - 34.64 years, 29 women). One participant was excluded due to experimenter error. Participants also completed a demographic survey. Forty percent were native German speakers, 23% were native English speakers, and the remaining participants spoke a native language other than English or German. Thirty-three percent of participants had completed a high school degree, 43% had completed a bachelor's degree, and 24% had completed a master's degree. Seventy percent of participants were studying psychology or education. The remaining participants studied physics, medicine, engineering, informatics, sports science, or political science. Participation was voluntary and participants received course credit and/or candy for participating.

2.5.1.2 Materials

The color vision test and Lego scientific reasoning tasks described in Study 2a were used.

2.5.1.3 Procedure

Data for this study were collected from June - November 2018. All sessions took place in a quiet room and were video recorded. The session lasted approximately 20 minutes. The procedure proceeded as described in Study 2a (Appendix D). The experiment was conducted in English.

2.5.1.4 Coding

Interpretation of confounded evidence task.

Knowledge Claims. Participants' knowledge claims (KCs) were coded as either correct or incorrect. A KC was correct if it indicated that it was not possible to know which bricks in the stick of four made the box light up. Any other response was considered incorrect.

Explanations. Participants' explanations for why they could not know which bricks were responsible for making the box light up were coded as described in Table 2.4. The percentages add up to more than 100% because participants provided between one and three explanations each. Participants could receive up to five points for explanations that referred to confounded or new evidence, the need for isolating or testing the bricks, or the knowledge that at least one lighter was present. Explanations that referred to the lack of a pattern or rule did not receive any points. There was one instance of a participant being unsure if the non-lighters had inhibitory effects when they outnumbered the lighters. This explanation was not included in the scoring. If participants provided an incorrect statement as an explanation, they received no points.

Control of Variables tasks. Participants choices were coded as correct or incorrect. The justifications for their choices were coded as described in Table 2.5. Participants could receive a maximum of six points for their justification of a 2-variable task and a maximum of seven points for a justification of a 3-variable task. However, if participants made an incorrect inference or suggested an incorrect experiment design, they received no points.

Table 2.4

Example explanations for the Interpretation of Confounded Evidence task, the percentage of participants providing each type of response, and the scoring system for responses.

ICE Explanations			
Example	Description	Percentage	Points
It could be any of them; at least one is a lighter	Presence of lighter	70%	1
These are different bricks	New evidence	7.5%	1
They're stuck together	Confounded	5%	1
I can't take them apart	Isolation of variables	25%	1
I can't test them individually	Testing	25%	1
There's no clear rule	Rules	7.5%	0
Color/glitter doesn't seem to matter	Rules	5%	0
Position/pattern of lighters doesn't seem to matter	Rules	15%	0
I don't remember the colors, so I don't know	Referring to previous evidence	2.5%	0
I don't know if one lighter is enough to overcome two non-lighters	Inhibitory non-lighters	2.5%	0
They light up when there are two Tomas All of them are not Tomas	Incorrect statement	7.5%	0

Table 2.5

Example justifications for the CVS Tasks.

CVS Justification Scoring			
Description	Percentage 2-variable	Percentage 3-variable	Points
Reference to control variable (1)	58%	18%	1
Reference to control variables (2)	-	70%	1
Reference to absence of focal variable	13%	45%	1
Reference to potential outcome: box lights up	15%	20%	1
Reference to potential outcome: box doesn't light up	30%	38%	1
Reference to potential inference based on outcome: box lights up	8%	10%	1
Reference to potential inference based on outcome: box doesn't light up	23%	35%	1
Incorrect inference/ testing strategy	30%	2.5%	0

Participants' interpretations of the outcome of the test were coded as described in Table 2.6. Participants could receive a maximum of six points for their interpretation if they chose a controlled test. If participants chose a confounded test, they could receive one point for making the correct inference that they can't know if the X brick is a lighter. However, if participants made an incorrect inference, they received no points.

Table 2.6

Example interpretations of the outcome of the CVS Tasks.

CVS Interpretation Scoring			
Description	Percentage 2-variable	Percentage 3-variable	Points
The X brick is a lighter.	83%	90%	1
I am sure/certain.	70%	68%	1
The stick I chose did not light up.	40%	58%	1
Therefore, none of those bricks are lighters.	60%	58%	1
The original stick did light up.	40%	53%	1
Therefore, the X brick must be the lighter.	68%	83%	1
Incorrect choice, correct inference	5%	5%	1
Incorrect inference	10%	5%	0

2.5.2 Results

Only the first trials of each task were included in the analysis because participants stopped providing detailed (or any) answers as the procedure continued. Descriptives for the variables of interest can be found in Table 2.7. Simple bivariate correlations were calculated for age, gender, native language, level of education, university subject, and ICE

Explanation score, 2-variable justification and interpretation, and 3-variable justification and interpretation (Table 2.8).

2.5.2.1 *Interpretation of Confounded Evidence*

In the interpretation of confounded evidence task, all participants provided correct Knowledge Claims, that they could not know for sure which bricks made the box light up. Participants provided between one and three explanations for their Knowledge Claims. Eighty-five percent of participants received at least one point for a correct explanation of their Knowledge Claim. Three participants did not provide any valid explanation and three provided an incorrect explanation. ICE Explanation score was significantly positively correlated with participants' age and both the 3-variable justifications and interpretations. A linear regression analysis was conducted to predict ICE Explanation score from age. The model was significant, $F(1,38) = 5.37, p = .03, R = .35, R^2 = .12, R^2_{Adjusted} = .10$.

Table 2.7

Descriptives for the scientific reasoning tasks.

Variables	Mean	SD	Min	Max
ICE	1.28	0.85	0	3 (5)
2-Var Justification	1.45	1.54	0	5 (6)
2-Var Interpretation	3.79	2.04	0	6
3-Var Justification	3.05	1.50	0	7
3-Var Interpretation	4.08	1.80	0	6

Note. () absolute maximums in brackets

Table 2.8

Correlations Among and Descriptive Statistics for Gender, Age, Native Language, Level of Education, University Subject, and ICE Explanation Score, CVS Justification Scores and CVS Interpretation Scores

Variables	1	2	3	4	5	6	7	8	9	10
1. Gender	–									
2. Age	-.13	–								
3. Native English	-.07	.03	–							
4. Education level	.01	.39*	.29	–						
5. Subject	.45**	.50**	-.04	.22	–					
6. ICE explanation	-.07	.35*	.11	-.01	.22	–				
7. 2-var justification	.15	-.37*	-.20	-.23	.01	-.14	–			
8. 2-var interpretation	-.18	-.14	.09	.12	-.16	.11	.16	–		
9. 3-var justification	-.02	.38*	.02	.003	.35*	.39*	.18	.17	–	
10. 3-var interpretation	-.26	.22	-.09	.12	-.16	.34*	-.15	.26	-.07	–

Note. * $p < .05$. ** $p < .01$. *** $p < .001$

2.5.2.2 *Control of Variables*

2-variable task. In the 2-variable CVS task, 93% of participants selected the correct stick, which represented a controlled test. Sixty-eight percent of participants who chose the correct test received at least one point for a correct justification of their test choice. One participant did not provide a valid justification and eleven participants provided an incorrect justification. Of the 11 participants who provided an incorrect justification, eight of them did so in the first trial of the task. The three participants who chose the incorrect confounded test received one or no points for their justifications. The 2-variable CVS Justification score was significantly negatively correlated with participants' age. A linear regression analysis was conducted to predict CVS Justification score from age. The model was significant, $F(1,38) = 5.95, p = .02, R = .37, R^2 = .14, R^2_{Adjusted} = .11$.

Of the participants who chose a controlled test, 92% made a correct inference that they could conclusively say the X brick was a lighter. Of the participants who chose a confounded test, 67% made a correct inference that they could not conclusively know if the X brick was a lighter. In total, 10% of participants made an incorrect inference as a result of their choice and the outcome of the test. There were no significant correlations with 2-variable Interpretation score.

3-variable task. In the 3-variable CVS task, 95% of participants selected the correct stick, which represented a controlled test. Ninety-seven percent of participants who chose the correct test received at least one point for a correct justification of their test choice. One participant provided an incorrect justification. The two participants who chose the incorrect confounded test received one point for their justifications. The 3-variable CVS Justification score was significantly positively correlated with participants' age, their study subject being something other than psychology or education, and the ICE

explanation. A multiple linear regression analysis was conducted to predict CVS Justification score from age, university subject, and the ICE explanation. The model including these effects on CVS Justification was significant, $F(3,36) = 4.02, p = .01, R = .50, R^2 = .25, R^2_{Adjusted} = .19$. Specifically, the coefficient for the ICE explanation was marginally significant, $B = 0.51, SE = 0.27, t = 1.86, p = .07$.

Of the participants who chose a controlled test, 95% made a correct inference that they could conclusively say the X brick was a lighter. Of the participants who chose a confounded test, 100% made a correct inference that they could not conclusively know if the X brick was a lighter. In total, 5% of participants made an incorrect inference as a result of their choice and the outcome of the test. The 3-variable CVS Interpretation score was significantly positively correlated with the ICE explanation. A linear regression analysis was conducted to predict CVS Interpretation score from the ICE explanation. The model including this effect on CVS Interpretation was significant, $F(1,38) = 4.93, p = .03, R = .34, R^2 = .12, R^2_{Adjusted} = .09$.

Overall performance. Looking across both CVS tasks, only one individual chose the incorrect test in both tasks. The other three individuals who chose incorrectly did so only once. In many cases, participants did not provide fully elaborated responses for both the justification of their choice and their interpretation of the outcome, as this could be repetitive. For this reason, we looked at the justification and interpretation scores together and saw that all participants received at least one point overall across both tasks. The average combined score for both responses was 6.19 (2.66) and ranged from 0 to 12 out of 13 possible points.

A few additional observations were made in coding participants' responses. Four individuals originally chose an incorrect test but, while providing their explanation for their choice, realized their mistake and switched to the correct test. Two individuals

mentioned a preference for testing the focal variable when choosing a test. Two individuals who chose an incorrect test, explained wanting to try all new colors. Five individuals forgot about the fact that the original stick had already been placed on the box and lit up. Ten individuals mentioned that they were assuming that bricks of the same color had the same effect on the box. Three individuals mentioned that they had to assume that the sticks of three performed in the same way as the sticks of two, because they had not seen a stick of three previously.

2.5.3 Discussion

The aim of Study 3 was to investigate adults' understanding of confounded evidence and the Control of Variables Strategy using the same knowledge-lean task originally developed for use with preschoolers. Because research has shown that the content of a task, as well as individuals' prior knowledge and beliefs (e.g., Kuhn et al., 1988; Kuhn et al., 1995), can influence adults' ability to reason scientifically, we wanted to further investigate their abilities with this simple novel task.

In the interpretation of confounded evidence task, all of the adults in this sample recognized that when presented with confounded evidence, a stick of bricks that makes the box light up, they could not know conclusively which of those bricks were effective. When asked to explain why this was the case, however, not all adults could: three adults could not provide a valid explanation referring to the confounded evidence, and three other adults provided an incorrect explanation. This is surprising considering the simplicity of the task and the fact that a correct response simply required participants to mention that the bricks were stuck together, could not be tested individually, or that they knew there was at least one lighter present but could not know which one. However, it is also in line with the research suggesting that adults do not show complete competence in scientific reasoning, as reviewed in the introduction to this study. Participants provided

fairly limited explanations, with an average of 1.4 out of 5 potential points. Further, age was predictive of ICE explanations, such that older participants provided higher scoring explanations.

When looking at performance on the CVS tasks, participants performed better on the 3-variable than on the 2-variable task in selecting a controlled test, as well as both the justification and interpretation measures. Fewer participants selected the incorrect test in the 3-variable task, more participants received at least one point for their justification; fewer participants provided an incorrect justification, and fewer participants made an incorrect inference in the 3-variable task than in the 2-variable task. This unexpected pattern also revealed itself in the relation between CVS justifications and age: justifications on the 2-variable task were negatively correlated with age, while justifications on the 3-variable task were positively correlated with age. It appears as if the 2-variable CVS task is somehow less intuitive and trips up the adult participants in recognizing the correct test, explaining the test, and interpreting the outcome.

Similarly to the ICE explanations, participants provided fairly limited justifications for their choices, as well as interpretations of their tests. Interestingly, adults performed much better in interpreting the outcome of their test than in explaining why they chose that test in the first place. Perhaps because when asked to interpret the outcome, participants were asked if they could now know for sure if the X brick was a lighter, and in order to show that they knew “for sure,” they provided more thorough explanations than when explaining why they picked a particular test. Also, interpreting the outcome should be easier because participants no longer have to reason in the hypothetical. When selecting a test, participants have to consider that there are two potential outcomes, the box lighting up or not lighting up, and what those outcomes mean for the hypothesis. It seems as though many participants did not fully verbalize the hypotheticals when selecting a test,

but once the outcome was known, they had less trouble verbalizing what could be interpreted from that outcome.

The additional observations also present some important points to consider. For example, the fact that some participants expressed a preference for testing the focal variable (even though or because it was not present in the choices) suggests that, even in adulthood, there is occasionally a desire to test the focal variable rather than to control the other non-focal variables. It was further interesting that two individuals suggested trying out completely new colors, which of course, cannot provide any information at all about the question. This could be due to misunderstanding the task, or because of a desire to find out more about how the box works, or due to a lack of understanding about CVS.

The fact that five individuals forgot that the original stick had been placed on the box and made it light up was especially unexpected. We were under the impression that this task was fairly simple (at least that was our intention in designing it), but if some participants forget a key piece of information for making a decision about which test to choose or how to interpret the test within 30 seconds of observing this evidence, then it could be that there is too much load on working memory. Another possibility, though, is that adults think the task is so simple that they do not pay full attention and thus forget parts of the task. Either way, this would be important to address, either by emphasizing the importance of paying attention or by helping participants to keep track of what has already been tested and what effect occurred.

Additionally, adults vocalized two assumptions that they felt they had to make in order to make decisions about their choice of test or interpretation of the outcome. First, that they had to assume that bricks of the same color have the same effect on the box. This assumption was correct and one that we assumed participants would make. But the fact that some participants needed to state this assumption may suggest that we should state it

first so that it is clear from the beginning. Second, some participants mentioned that they had to assume that the sticks of three perform in the same manner as the sticks of two (or four) as they had not previously seen a stick of three. The reasoning behind this was that if there were two non-lighters and one lighter in a stick of three, they were not sure if two non-lighters would overpower or cancel out the lighter. This confusion is fair considering we did not present sticks of three in the familiarization and can be easily addressed by including sticks of three in the familiarization phase.

The present study used a novel, knowledge-lean task to investigate CVS abilities in adults. Selecting a controlled test was not an issue for participants, with over 90% able to select a controlled test. This performance was slightly better than performance in Bullock and Ziegler's study (1999), which could be expected considering this task was designed to be simpler. Providing justifications for and interpreting the outcome of a test was more difficult, with most participants providing fairly limited correct responses and some participants providing incorrect justifications and inferences. These results support previous research suggesting that even adults are not completely competent in scientific reasoning and shows that even in a simple, knowledge-lean task without the influence of any prior knowledge or beliefs about the task, adults struggle to form complete justifications and interpretations for controlled experiments.

2.6 General Discussion

The four studies presented in this chapter investigated the understanding of confounded evidence and the Control of Variables Strategy in preschoolers and adults.

Studies 1, 2a, and 2b revealed that three- to six-year-olds show a nascent understanding of controlled experiments, recognizing that ambiguous data was ambiguous and selecting appropriately controlled tests to determine particular hypotheses about the causal structure. In addition, older children could also provide relevant justifications for

their choices. These findings support recent advances in the literature suggesting that preschoolers have some understanding of Control of Variables Strategy (van der Graaf et al., 2015) and that recognizing an unconfounded test is easier than producing one (Bullock & Ziegler, 1999).

Study 3 revealed that, with the same knowledge-lean task, almost all adults can recognize and select a controlled test, but they struggle to provide fully-formed justifications for a controlled test or interpretations of the outcome of their chosen test. Though these results support previous research showing that adults do not show complete competence in CVS, it is surprising considering the simple, knowledge-lean nature of this task. It suggests that adults' poor performance on CVS tasks in the past cannot be solely a result of their prior knowledge of or beliefs about the task content interfering with their reasoning. It suggests that there are more basic deficits in reasoning abilities or at least in the verbalization of the reasoning process, which perhaps could be addressed through training on such knowledge-lean tasks before introducing more complicated, usually scientific, task content.

Additionally, a number of observations from Study 3 hold relevance for future investigations using this task with children. First, the surprising pattern of better performance on the 3-variable CVS task than on the 2-variable task, as we also saw in preschoolers in Study 2a. If the 2-variable task gave adults more trouble than the 3-variable task, this should be kept in mind when using the task with children. It might make more sense to focus investigation on children's abilities on the 3-variable task to ensure that something else is not getting in the way of their reasoning process. Second, the difficult process of reasoning in the hypothetical for choosing a test and also justifying the test. Adults were better able to interpret the outcome of an experiment than to provide a justification for it, and this could also be the case for children. Future investigation could

demonstrate confounded and controlled tests to children and have them interpret the outcome, rather than having them select and justify a test themselves.

Third, the fact that some adult participants forgot that the test stick had been placed on the box and made it light up was unexpected. If this was indeed due to working memory demands, then this could also be a problem for children, but one that is not noticed because they do not speak up about having forgotten. This could be addressed in the same way by helping participants to keep track of what has already been tested and what effect occurred. Fourth, it could be beneficial to address particular assumptions early on to ensure that there is no confusion, for example about the fact that bricks of the same color behave in the same way, that sticks of three behave in the same manner as sticks of two or four, or that non-lighters do not have inhibitory effects when they outnumber the lighters.

An important realization from the observations from Study 3 is that, if adults have such issues, it is possible and even likely that children also have them, but they are unable or unwilling to verbalize these issues to the experimenter. Addressing these issues may reveal that young children perform even better on these tasks and, thus, may be more capable of scientific reasoning than as suggested by the current results.

To further investigate young children's scientific reasoning abilities, the next chapter describes the relation between those abilities and other cognitive factors, such as executive functioning and Theory of Mind.

3 The Structure and Correlates of Scientific Reasoning in Preschool

3.1 Introduction

The development of scientific reasoning does not occur independently of the development of other cognitive abilities; thus, it is important to consider and investigate the relations between scientific reasoning and other cognitive abilities to gain a more robust understanding of the development of scientific reasoning.

3.1.1 Metacognition and Theory of Mind

As discussed in the previous chapters, metacognition and Theory of Mind likely play an important role in children's developing scientific reasoning abilities (see Kuhn, 2010). Metacognition, the ability to reflect on one's own or others' thinking processes and to understand how one acquires knowledge (Kuhn, 1989; Kuhn et al., 2008; Sodian & Frith, 2008), consists of two components. The first component includes understanding of the mental world, such as knowing, believing, wanting (ToM), as well as an understanding the recursive nature of mental states, or in other words, the idea that a person can have a belief about a belief, which has been termed Advanced Theory of Mind (AToM; Koerber & Osterhaus, 2019; Osterhaus et al., 2017). The second component includes monitoring and self-regulating the process of knowledge acquisition (Schneider, 2008).

The recognition of the existence of different mental states and that different people can have different beliefs, as well as the appreciation of alternative possible outcomes, could help children to perform well on scientific reasoning tasks that require them to think about what would happen in one experiment versus another experiment and how the outcomes could be affected by different variables (and whether or not those variables are controlled or manipulated). The ability to self-monitor and reflect on what one knows or does not know (and how) should also be related to children's performance on scientific

reasoning tasks that require them to interpret evidence and recognize if they can or cannot reach conclusions on the basis of evidence. Finally, the ability to revise beliefs should be related to the ability to revise hypotheses as well (Kuhn & Pearsall, 2000). Indeed, the literature reviewed in Chapter 1 showed that metacognitive training improves performance on scientific reasoning tasks (Amsel et al., 2008; Kuhn et al., 2008).

3.1.2 Executive functioning

Executive Functioning is a set of conscious cognitive processes or skills that control and regulate attention, thoughts, and behaviors and consists of working memory, inhibition, and cognitive flexibility (Diamond, 2013; Miyake et al., 2000; Zelazo & Müller, 2010). Executive functioning has been shown to relate to both school readiness and academic achievement across domains such as language, math, and science (e.g., Bustamante, Greenfield, & Nayfeld, 2018; Clark et al., 2014; Nayfeld, Fuccillo, & Greenfield, 2013; Welsh, Nix, Blair, Bierman, & Nelson, 2010; see Diamond, 2013 for a review).

Working memory is responsible for holding information in short-term memory to keep it available for processing and has a limited capacity (Miller, 1956). It could be implicated in children's learning, for example, facilitating the learning of cause and effect relations by holding the information that a child observes as well as rules about causality in mind (Gropen, Clark-Chiarelli, Hoisington, & Ehrlich, 2011). Inhibition is the ability to suppress certain behaviors or thoughts in favor of other more necessary or appropriate behaviors or thoughts. In other words, to avoid impulse reactions and instead control responses (Diamond, 2013). It is likely important for helping to focus attention on certain aspects, for example, of cause and effect relations and suppressing prior beliefs that could affect the learning or interpretation of those relations. It could also support the revision of hypotheses (Gropen et al., 2011). Cognitive flexibility is implicated in the changing of

perspectives, adaptation of thought processes, adjusting to different demands, and shifting between different tasks (Diamond, 2013).

Thus, executive functioning likely supports the development of causal and scientific reasoning abilities by facilitating that children attend to important features and suppress noncritical features and prior beliefs, by holding important information in mind for processing and incorporation into existing knowledge schemas, and by allowing that children can flexibly revise and update hypotheses. Further, Gropen and colleagues (2011) argue that executive functioning and Theory of Mind are also likely related, such that the development of executive functioning facilitates the development of Theory of Mind abilities. As support for this claim, they point out that individual differences in executive functioning are correlated with false belief tasks and also predict later performance on such tasks (Zelazo, Carlson, & Kesek, 2008). Research on the structure of executive functioning has found, on the one hand, that executive functioning is best described as a unitary construct in early childhood (e.g., Wiebe, Espy, & Charak, 2008; Wiebe et al., 2011), and, on the other hand, it may rather be a set of distinct factors, such as inhibition and working memory (e.g., M. R. Miller, Giesbrecht, Müller, McInerney, & Kerns, 2012).

In addition to Theory of Mind and executive functioning, other cognitive factors such as intelligence and language abilities are also related to scientific reasoning skills. A number of studies have investigated the relation between scientific reasoning and other cognitive abilities. We describe a selection of those studies below. The first set of studies we present have investigated whether scientific reasoning is a unique skill, controlling for other cognitive factors such as intelligence or language abilities. The second set of studies investigates how individual differences in other cognitive skills explain individual differences in scientific reasoning.

3.1.3 The structure of Scientific Reasoning abilities

The studies described in this section have investigated the uniqueness of scientific reasoning as a cognitive ability and the structure of scientific reasoning and its subcomponents. Mayer and colleagues (2014) used a paper-and-pencil measure of scientific reasoning with ten-year-old children. Specifically, they investigated understanding the nature of science, understanding theories, designing experiments, and interpreting data. They found that two-dimensional models representing scientific reasoning as separate from both reading comprehension and intelligence fit the data best, suggesting that scientific reasoning is a separate construct (Mayer, Sodian, Koerber, & Schwippert, 2014). Koerber and colleagues (2015) developed a measure of eight- to ten-year-old children's scientific thinking with 66 story problems assessing five components: goals of science, theories and interpretive frameworks, experimentation strategies, experimental designs, and data interpretation. They used this scale to investigate the structure of scientific thinking and its relation to other cognitive factors. They, like Mayer and colleagues (2014), found that the scale could measure scientific thinking as a unitary construct, which was separate from reading comprehension and intelligence.

Koerber and Osterhaus (2019) developed a measure of kindergarten children's scientific thinking with 30 multiple-choice questions assessing three components: experimentation, data interpretation, and understanding the nature of science. They used this scale to investigate the structure of scientific thinking and its relation to other cognitive factors. They found that the scale could measure scientific thinking as a unitary construct, suggesting that there is a core ability underlying the development of scientific thinking.

Though the studies described above show that different components of scientific reasoning are related and represent subcomponents of a unitary structure of scientific

reasoning, other studies have shown that measures of components of scientific reasoning such as experimentation and evidence evaluation are not necessarily related (van der Graaf, Segers, & Verhoeven, 2016) and when they are related, they are not predictive of each other (van der Graaf, Segers, & Verhoeven, 2018; see also Bullock et al., 2009; Piekny et al., 2013). These contradictory findings could be a result of the different types of tasks used to assess scientific reasoning. For example, the studies described above used paper-and-pencil measures and multiple-choice responding, whereas the studies by van der Graaf and colleagues (2016, 2018) used a hands-on experimentation task (Chen & Klahr, 1999). In the next section, we will describe studies that have focused on the relation between individual differences in other cognitive skills and scientific reasoning abilities.

3.1.4 Correlates of Scientific Reasoning

This section describes studies that have found relations between Theory of Mind and false belief understanding, language abilities, intelligence, and executive functions and scientific reasoning abilities.

Klein (1998) found that children's (1st, 3rd, 5th graders) performance on a covariation evidence evaluation task (Ruffman et al., 1993) was related to their ability to design experiments that contrasted the focal variable. Klein also found that their understanding of conclusive and inconclusive evidence (Mouse House task; Sodian et al., 1991) was related to their ability to design controlled experiments in a pendulum task (Inhelder & Piaget, 1958). Klein concluded that Theory of Mind, in terms of distinguishing between beliefs and evidence, is fundamental for experimentation abilities and that this perspective can account for the different traditions of understanding scientific reasoning. Theory of Mind as the basis for scientific reasoning allows, on the one hand, that reasoning is based on knowledge about beliefs and causes which is domain-specific,

and on the other hand, allows for domain-generalty because this knowledge can be applied to reasoning about beliefs in other domains as well (Klein, 1998).

Astington et al. (2002) found a relation between second-order false belief ability (beliefs about beliefs, which may be false) and the ability to distinguish between causes of a situation and reasons for believing the situation in five- to seven-year-olds, after controlling for language and nonverbal reasoning abilities. They suggest that second-order understanding is fundamental to children's development and that it is this ability that likely facilitates the ability to understand evidence and reasoning.

Piekny and colleagues (2013) investigated children's developing scientific thinking abilities longitudinally at four- and five-years-old. They found that understanding of false belief at age four predicted experimentation abilities at age five, after controlling for intelligence, language, executive functioning, and working memory. But false belief understanding and experimentation skills were not related at the same point in time. Working memory was not related to experimentation skills at four or five. Intelligence was related to false belief understanding at age four and to experimentation skills at age five. They also found that different skills did not emerge all at the same time: evidence evaluation skills emerged first, with experimentation and hypothesis generation skills emerging later, suggesting that scientific reasoning consists of different subskills rather than depending solely on the development of the understanding of the theory-evidence distinction.

Sodian and colleagues (2016) found that both first- and second-order false belief understanding at five years predicted experimentation skills at eight years, independently of intelligence and executive functions (Sodian, Kristen-Antonow, & Koerber, 2016). They did not find an effect of metacognition (of own ignorance (Rohwer et al., 2012).

These findings suggest that the ability to represent beliefs as independent from reality precedes and is fundamental for the development of scientific reasoning abilities.

Osterhaus and colleagues (2017) found that, in children age eight to ten years, understanding of the nature of science (NoS) and experimentation skills were related to general information processing skills; specifically, that experimentation was related to inhibition and NoS was related to intelligence and language abilities. Advanced Theory of Mind (AToM) predicted NoS, which in turn predicted experimentation skills, after controlling for the general information processing skills listed above.

Mayer and colleagues (2014) found that intelligence, problem-solving, spatial skills, and reading skills were related to scientific reasoning skills in ten-year-old children. Inhibition, however, was not related to scientific reasoning abilities. They offer a possible explanation for this finding: that their paper-and-pencil measure of scientific reasoning may not have triggered prior beliefs that would then need to be inhibited.

Koerber and colleagues (2015) found a strong influence of intelligence on scientific thinking, such that children (2nd, 3rd, 4th graders) with higher intelligence performed better on the measure of scientific thinking. The level of parental education was also related to children's performance, such that children of parents with higher education performed better on the measure. There was also a positive effect of age and an effect of schooling, such that a nine-year-old in 4th grade would perform better than a nine-year-old in 3rd grade.

In addition, Koerber and Osterhaus (2019) found that in kindergarteners (6-year-olds) intelligence and language abilities were related to scientific thinking performance and that AToM predicted scientific thinking after controlling for intelligence and language abilities. This finding replicated their earlier findings with elementary school children (Osterhaus et al., 2017). The Munich longitudinal study (Bullock et al., 2009)

also revealed that intelligence and formal reasoning were correlated with measures of scientific reasoning in children from age nine to age 23. Haslbeck and colleagues (2018) have also found a relation intelligence and between the ability to plan experiments in both elementary and preschool children (Haslbeck, Lankes, Fritzsche, Kohlhauf, & Neuhaus, 2018).

Wagensveld and colleagues (2015) investigated how children's (4th & 6th grade) ability to acquire Control of Variables Strategy (CVS) skills was related to their other cognitive skills. Children who received no instruction but had to learn CVS through their experiences with the tasks were able to do so by relying on their existing knowledge (as measured in a pretest) and their existing reading, vocabulary, and verbal reasoning skills. The fact that linguistic factors were related to CVS (and nonverbal reasoning was not) suggests that language is important for scientific reasoning and science learning (Wagensveld, Segers, Kleemans, & Verhoeven, 2015).

Van der Graaf and colleagues (2016) investigated the role of cognitive factors in four-year-olds' abilities in experimentation and evidence evaluation. They found that executive functioning, specifically inhibition and verbal working memory, was indirectly related to scientific reasoning skills through grammatical ability. Vocabulary, visuospatial working memory, spatial visualization, and cognitive flexibility were not related to the measures of scientific reasoning. They also found that their two measures of scientific reasoning for experimentation and evidence evaluation did not correlate, suggesting that those may be separate components of scientific reasoning. Van der Graaf and colleagues propose that language can play a role in reasoning by providing structure to the reasoning process, by supporting mental representations of experiments and evidence, by generating analogies that help facilitate comparisons.

Van der Graaf and colleagues (2018) investigated kindergarteners' developing scientific thinking abilities, specifically experimentation, evidence evaluation, and domain-specific knowledge, longitudinally halfway through and at the end of senior kindergarten. As predictors, they assessed short-term memory, inhibition, and cognitive flexibility, as well as grammar and vocabulary, in junior kindergarten. They found overall improvement in scientific thinking from the first measurement point to the second and found that performance on scientific thinking at the first measurement point predicted performance at the second. Inhibition, verbal working memory, and grammatical abilities were predictive of all three measures of scientific thinking, replicating their previous results (van der Graaf et al., 2016). They additionally found that vocabulary predicted experimentation. Experimentation and evidence evaluation were related but not predictive of each other, suggesting that they are separate but related components of scientific reasoning and may develop independently.

Studies have also investigated the neural activity occurring during scientific reasoning. For example, Kwon and Lawson (2000) found that the activity in areas of the brain associated with inhibition (and also representation) were associated with scientific reasoning abilities. Further, in a review of the literature, Nenciovici and colleagues (2019) have shown that areas of the brain associated with executive functioning are active during scientific reasoning tasks involving hypothesis generation, causal reasoning, and overcoming misconceptions in scientific domains (Nenciovici, Allaire-Duquette, Masson, 2019).

Bauer and Booth (2019) investigated the relation between scientific literacy and executive functioning and causal reasoning abilities in three-year-old children. They found that executive functioning was correlated with scientific literacy, such that children with greater executive functioning scores performed better on both causal reasoning and

scientific literacy measures. They suggest that inhibition plays a role in helping children pay attention to important information and working memory can help them keep track of and process information to learn about scientific concepts.

In addition, causal inference ability was associated with scientific literacy. They surprisingly found that their measure of cause-effect relation (a blicket detector task) did not correlate with the other measures of causal reasoning or with scientific literacy. The authors hypothesize that such blicket detector tasks simply require children to isolate relations between objects and events, while other causal reasoning tasks require more reasoning about causal mechanisms, and that the detection of cause-effect relations may not be close enough to higher-level reasoning skills to be involved in scientific reasoning.

Finally, most studies did not find any relation between gender and scientific reasoning abilities (Astington et al., 2002; Bauer & Booth, 2019; Bullock et al., 2009; Koerber et al., 2015; Sodian et al., 2016; van der Graaf et al., 2018). Some studies have found gender differences in regard to science in school, showing that boys show higher science achievement in kindergarten and 3rd grade (Saçkes, Trundle, Bell, & O'Connell, 2011) and better performance in science and reasoning in grades seven to ten (J.-T. Kuhn & Holling, 2009).

3.1.5 Summary

To summarize this section, there is ongoing discussion as to the structure of scientific reasoning as a unitary construct or a number of separate but related subcomponents (Bullock et al., 2009; Koerber et al., 2015; Koerber & Osterhaus, 2019; Mayer et al., 2014; Piekny et al., 2013; van der Graaf et al., 2016, 2018).

A number of studies have shown that false belief understanding and (advanced) Theory of Mind abilities are related to several aspects of scientific reasoning, for example, experimentation, understanding evidence and justifications, as well as the understanding

of the nature of science (Astington et al., 2002; Klein, 1998; Osterhaus et al., 2017; Piekny et al., 2013; Sodian et al., 2016). Regarding the relation between executive functioning and scientific reasoning, most studies show that inhibition is related to scientific reasoning (Kwon & Lawson, 2000; Osterhaus et al., 2017; van der Graaf et al., 2016, 2018), but some have found that it was not (Mayer et al., 2014), perhaps because the tasks did not require inhibition. The relation between working memory and scientific reasoning is less clear, especially since there is a distinction between verbal and nonverbal working memory. For example, van der Graaf and colleagues found that verbal working memory was related to scientific reasoning (2016, 2018), but visuospatial working memory was not (2016). Piekny and colleagues (2013) assessed working memory with seven subtests and generated a composite score, which was not related to scientific reasoning, but since the composite score combined both verbal and nonverbal aspects of working memory, this could have diluted the effect of verbal working memory.

Intelligence has been shown to be related to scientific reasoning abilities across a number of different tasks and age groups (Bullock et al., 2009; Haslbeck et al., 2018; Koerber et al., 2015; Koerber & Osterhaus, 2019; Mayer et al., 2014; Piekny et al., 2013). Language abilities, including reading comprehension (Koerber et al., 2015; Mayer et al., 2014; Osterhaus et al., 2017; Wagensveld et al., 2015), language receptiveness (Koerber & Osterhaus, 2019), vocabulary (van der Graaf et al., 2018; Wagensveld et al., 2015), and grammatical abilities (van der Graaf et al., 2016, 2018) have been shown to be related to scientific reasoning abilities. Finally, most studies show that gender is not related to performance on scientific reasoning measures (Astington et al., 2002; Bauer & Booth, 2019; Bullock et al., 2009; Koerber et al., 2015; Sodian et al., 2016; van der Graaf et al., 2018).

As there has been limited research on preschooler's scientific reasoning in general, and no research with knowledge-lean tasks, it is important that we further investigate how scientific reasoning, as measured by our knowledge-lean tasks, relates to the development of other cognitive factors.

3.2 Study 4: The Structure and Correlates of Scientific Reasoning in Preschool

The aim of the present study is to further investigate scientific reasoning abilities in early childhood and how they relate to other cognitive factors that may be foundational for developing scientific reasoning. Specifically, we investigated the relation between the scientific reasoning measures from Chapter 2 (selection and justification of a controlled test and recognition of the inconclusiveness of confounded evidence) and intelligence (general knowledge), language abilities (grammar), executive functions (inhibition, working memory, planning, and cognitive flexibility), and Theory of Mind (knowledge access, content false belief, and explicit false belief) in four-year-old children. With this design, we gain a clearer picture of young children's developing scientific reasoning abilities.

3.2.1 Method

3.2.1.1 Participants

Data for this study were collected as part of a larger longitudinal study investigating the development of children's scientific reasoning skills. Our study sample included 187 children (91 girls) from a large German city. Thirty-five children were excluded due to problems with color vision (13), unwillingness to participate (10), experimenter error (6), technical problems with the materials (4), or not completing the CVS task (2). Participating children were four years of age at the first session ($M_{\text{age}} = 48$ months, $SD = 1.59$; range = 39-51 months). Children did not have any diagnosed

developmental delays or disorders, and they understood German “well” or “very well” as reported by a parent.

Sociodemographic information was only available for two-thirds of the sample at the time of publication. Eighty-nine percent of children were native German speakers. The remaining 11% spoke Bosnian, Bulgarian, Chinese, French, Greek, Polish, Portuguese, Russian, Spanish, Turkish, and Ukrainian as native language with German as their second language. Overall, 26% of the sample was multilingual, with English as the most common second language (44%). All sessions were carried out in German. The sample was racially, ethnically, and socioeconomically representative of our recruitment area.

With respect to maternal education, less than 1% of mothers reported having no degree, 8% held a Hauptschule, Realschule, or Berufsschule degree, 6% held a high school degree, and 48% had a university degree, 2% reported another form of education. With respect to paternal education, 9% held a Realschule or Berufsschule degree, 10% held a high school degree, 45% had a university degree, and 2% reported another form of education.

3.2.1.2 Measures & Coding

We used the tasks from Chapter 2 to measure scientific reasoning abilities, three Theory of Mind tasks (Knowledge Access, Content False Belief, Explicit False Belief), four executive functioning tasks (Day-Night, inhibitory control; Backwards Digit Span, working memory; Truck Loading, planning; and Dimensional Change Card Sorting, cognitive flexibility), the General Knowledge subscale for intelligence, and the Formation of Morphological Rules subscale to measure language abilities.

Measuring Scientific Reasoning. We used a version of the scientific reasoning tasks in between the protocol of Study 1 and that of 2a. Specifically, we did improve a number of the questions as in Study 2a, but not all of them, and we also did not refer to the

bricks as Tomas. To shorten the task and prevent children from becoming frustrated during a long multi-task session (~80 minutes), the interpretation question after each CVS task was not asked. For the full protocol used in the present study, please refer to Appendix F. Additionally, a fixed order was used, such that children always received the tasks in the following order: ICE - CVS2 - CVS3 - CVS2 - CVS3 - ICE, and the materials were not counterbalanced (refer to Figure 2.4). The tasks were coded as described in Study 1. Thirty percent of the data was double coded. Agreement for ICE Knowledge Claims was 94% (Kappa = .88), and 99% for Robust ICE (Kappa = .96). Agreement for CVS choices was 98% (Kappa = .96) and 97% for justifications (Kappa = .84).

For the analyses investigating the relation between scientific reasoning and other cognitive abilities, combined scores for each of the three tasks were generated as follows. For the ICE task, children could receive 1 point for a correct knowledge claim and 2 points for a correct knowledge claim plus a valid explanation (robust ICE) per trial. Thus, across two trials of the ICE task, children could receive between 0 and 4 points. For the CVS tasks, per trial, children could receive 1 point for a correct choice and 2 points for a correct choice plus a valid justification (robust CVS). Thus, across two trials of each CVS task, children could receive between 0 and 4 points. We generated combined scores for ICE and CVS and also generated a general scientific reasoning score, combining all three tasks, for a score out of 12 possible points.

Measuring Theory of Mind. Children's Theory of Mind (ToM) was assessed using the German-language version of the Theory of Mind Scale (Wellman & Liu, 2004; see Hofer, Hauf, & Aschersleben, 2004; Kristen, Thoermer, Hofer, Aschersleben, & Sodian, 2006 for the full German-language version). We used three subscales: knowledge access, content false belief, and explicit false belief.

In the Knowledge Access (KA) task, children were shown a box with a drawer and asked, “What do you think is inside the drawer?” (Belief question). The experimenter then opened the drawer to reveal a toy dog inside. The drawer was closed again, and children were asked, “What was in the drawer?” (Memory question). Next, the experimenter introduced a toy figure, Anna, and said, “Anna has never looked in this drawer before.” Children were then asked the critical KA question, “Does Anna know what is in the drawer?” This was followed up with a Control question, “Has Anna looked in the drawer?” Children must provide a relevant answer to the Belief question and also correctly answer both the Memory and the Control questions in order for their response to the KA question to be considered. The KA question is coded as correct if children respond that Anna does not know what is in the drawer. Thirty percent of the data was double coded: agreement for the test question was 100%.

In the Content False Belief (FB) task, children were shown a Smarties candy container and asked, “What do you think is inside this container?” (Belief question). The experimenter then opened the container revealing a toy pig inside. The pig was then placed back inside the container, and children were asked, “Can you remember what’s really inside this container?” (Memory question). Next, the experimenter introduced a toy figure, Lukas, and said, “Lukas has never seen what is in this container.” Children were then asked the critical FB question, “What does Lukas think is in this container: Smarties or a pig?” This was followed up with a Control question, “Has Lukas looked in this container before?” Children must provide a relevant answer to the Belief question and also correctly answer both the Memory and the Control questions in order for their response to the Content FB question to be considered. The Content FB question is coded as correct if children respond that Lukas thinks there are Smarties in the container. Thirty percent of the data was double coded: agreement for the test question was 97% ($Kappa = .94$).

In the Explicit False Belief (FB) task, children were introduced to a toy figure, Paul, and told that Paul is looking for his gloves. The gloves could be in his backpack or in the closet, which are depicted as images and presented to children. Children are told that, in reality, Paul's gloves are in his backpack, but Paul thinks they are in the closet. Children are then asked the critical False Belief question, "Where will Paul look for his gloves: in his backpack or in the closet?" This was followed up with a Control question, "Where are Paul's gloves really: in the closet or in his backpack?" Children must correctly answer the Control question in order for their response to the Explicit FB question to be considered. The Explicit FB question is coded as correct if children respond that Paul will look for his gloves in the closet (where he thinks they are). Thirty percent of the data was double coded: agreement for the test question was 100%. These three tasks can be considered individually or combined and averaged into a general Theory of Mind score.

Measuring Intelligence. Children performed the General Knowledge subscale of the Wechsler Preschool and Primary Scale of Intelligence - Third Edition (WPPSI-III; Petermann & Lipsius, 2009; Wechsler, 2012). Raw scores were used for analysis because norms are not available for children under four years of age. The General Knowledge task measures a child's general cultural knowledge, long-term memory, and acquired facts with questions such as "How many eyes do you have?" or "What do you use to cut something?" The task consists of 28 trials worth 1 point each. The task ends after the child receives zero points on five trials in a row. Twenty percent of the data was double coded. A high degree of reliability was found between the two raters: two-way mixed, absolute agreement, single measures ICC was found to be .97.

Measuring Executive Functioning. To assess children's executive function abilities, we included four tasks measuring different components of executive function, specifically, inhibitory control, working memory, planning, and cognitive flexibility.

Day-Night Stroop (inhibitory control). To measure inhibitory control, we used the Day-Night-Stroop task (Gerstadt, Hong, & Diamond, 1994). Children are shown pictures of a sun or a moon and asked to say “Day” when the sun is presented and “Night” when the moon is presented. Then the desired response is switched, and children must say “Night” when the sun is presented and “Day” when the moon is presented. This requires children to inhibit the original matching response for the non-matching response. Children receive a point for each correct trial. Thirty percent of the data was double coded. A high degree of reliability was found between the two raters: two-way mixed, absolute agreement, single measures ICC was found to be .98.

Backward Digit Span (BDS; working memory). To measure working memory, we used the Backward Digit Span task (Davis & Pratt, 1995). In this task, the experimenter speaks a series of numbers, and children are instructed to repeat those numbers in the reverse order. Children first received two training trials with two numbers each and then two test trials per each level. Children pass a trial by repeating all numbers in the correct reverse order. If children pass at least one trial per level, they proceed to the next level. Two scores were generated from performance, the longest series of numbers successfully repeated, and also the total score resulting from all trials completed. Thirty percent of the data was double coded. A high degree of reliability was found between the two raters: two-way mixed, absolute agreement, single measures ICC was found to be .99 for both scores. For analyses, we used the total score from all trials for more variance.

Truck Loading (organization and planning). To measure organization and planning, we used the Truck Loading task (Carlson, Moses, & Claxton, 2004). In this task, children must help the postman deliver letters; however, there is a one-way road, and letters can only be delivered from the top of the pile to the bottom out of the truck. Therefore, children must load the letters into the truck in the opposite order of delivery,

i.e., the first letter to be delivered needs to be loaded into the truck last. Children complete two trials of each level (2 - 5 houses). They receive a point for the level if they load the truck correctly in at least one trial per level. If they fail in both trials of one level, they do not proceed to the next level. Children can receive a maximum of four points. Thirty percent of the data was double coded. A high degree of reliability was found between the two raters: two-way mixed, absolute agreement, single measures ICC was found to be .98.

Dimensional Change Card Sorting (cognitive flexibility, working memory and inhibition). To measure all three core components of the executive functions, i.e., working memory, inhibition, and cognitive flexibility, we used the Dimensional Change Card Sorting task (DCCS; van der Ven, Kroesbergen, Boom, & Leseman, 2013; Zelazo, 2006). Children were presented with two boxes with slots in the top, one with a blue circle and one with a red triangle. In the first phase, children were instructed that in the Color Game, they should place all blue cards in the box with the blue circle and place all red cards in the box with the red triangle. Children completed six trials of this game. Next, children were introduced to the Shape Game in which they should place all circles in the box with the blue circle and all triangles in the box with the red triangle. Children completed six trials of this game. Finally, children were introduced to the Border Game in which, when the card has a black border, they should play according to the rules of the Color Game, and when the card does not have a border, they should play according to the rules of the Shape Game. Children completed 12 trials of this game and received a point for each correct trial.

Following Zelazo (2006), children were classified as failing or passing the Shape Game of the DCCS; children passed by correctly sorting at least five out of six cards. Children were classified as failing or passing the Border Game; children passed by correctly sorting at least 9 out of 12 cards. For a total score, children could receive two

points if they passed both the Shape and Border Game, one point if they passed the Shape Game, and zero points if they failed the Shape Game. Thirty percent of the data was double coded: agreement for the Color Game was 100%; agreement for the Shape Game was 98% (Kappa = .99), agreement for the Border Game was 97% (Kappa = .94).

Measuring Language. To assess children's language abilities, we used the formation of morphological rules subscale from the Language Development Test for Children Aged 3-5 Years (Sprachentwicklungstest für drei- bis fünfjährige Kinder, SETK 3-5; Grimm 2015).

Formation of morphological rules (Morphologisches Regelverstehen). The subscale Formation of Morphological Rules measures the ability to generate the plural form of a word. The experimenter first speaks words in the singular form and asks the child to provide the plural form. Then the experimenter speaks a made-up word in the singular form and again asks the child to provide the plural form. For example, "Look, here is one apple... Here, there are even more. So, here are three... [apples]" or "Look, here is one kland... Here, there are even more. So, here are three... [klants]." The task consists of 18 trials. Children receive 2 points for each correct plural form they generate or 1 point for specific attempts to pluralize. Twenty percent of the data was double coded. A high degree of reliability was found between the two raters: two-way mixed, absolute agreement, single measures ICC was found to be .95.

3.2.1.3 Procedure

Data for this study were collected between October 2017 - August 2018 over two sessions lasting approximately 80 minutes each. Both sessions took place in the laboratory of the university, with the second session occurring approximately 12 days after the first ($M = 11.84$ days; $SD = 11.71$ days, range = 2-98 days). In the first session, parent's written consent and children's verbal consent was obtained. Children completed the Truck

Loading, Backwards Digit Span, Lego scientific reasoning, Day-Night Stroop, and Formation of Morphological Rules tasks, among others not included in the present study. Parents also completed a questionnaire that included questions concerning parental education, race and ethnicity, as well as household composition, income, and literacy environment. In the second session, children completed the Dimensional Change Card Sorting, General Knowledge, and the Theory of Mind tasks, among others not included in the present study. Children were tested in a colorfully decorated room. Sessions were video recorded for later coding of participant responses.

3.2.2 Results

We will first report the results of the scientific reasoning tasks as analyzed in Studies 1, 2a, and 2b (a summary of all results can be found in Appendix C). We will then report the results of the investigation of the relations between the scientific reasoning tasks and the other measures assessed in this study.

3.2.2.1 *Scientific Reasoning tasks*

In the ICE task, we analyzed whether children responded in a way that indicated an understanding of the inconclusiveness of evidence by answering that they did not know which bricks were lighters. These data are shown in Figure 3.1. We constructed a GEE with an independent working correlation matrix, a binomial distribution, and a cumulative logit link function (Zeger & Liang, 1986; Zeger et al., 1988) looking at the role of gender, age, and trial on children's knowledge claim responses. None of these factors were significant, all p -values $> .07$. Across two trials of the ICE task, 42% of the children responded correctly on both trials, 26% responded correctly on one of the trials, and 31% of children responded incorrectly on both trials.

Taking into account children's explanations for their knowledge claims, we constructed a GEE to control for within-subject responses examining children's robust

performance on the ICE task, looking at the role of age and trial. This model revealed a main effect of trial, $B = 1.02$, $SE = 0.25$, $[95\% \text{ CI} = 0.53, 1.51]$, $\text{Wald } \chi^2(1) = 16.70$, $p < .001$. In the first trial, 19% of children showed a robust understanding of confounded evidence by correctly responding they could not know which bricks made the box light up because they were all stuck together. In the second trial, 8% responded in this way. Across two trials of the ICE task, 6% of the children provided a robust response on both trials, 14% provided a robust response on one of the trials, and 80% of children did not provide a robust response on either trial.

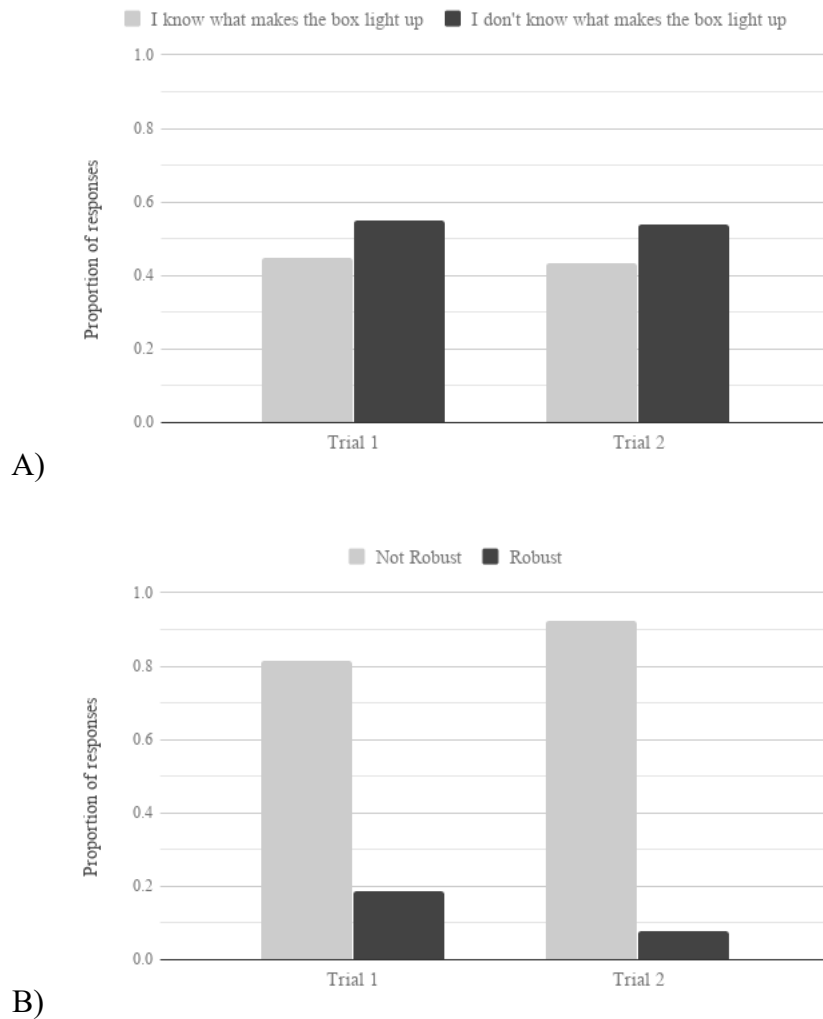


Figure 3.1. Children's performance on the interpretation of confounded evidence task. A) Children's knowledge claims about the effectiveness of the individual bricks. B)

Children’s robust performance: providing a relevant explanation for why they cannot know which bricks make the box light up.

Next, we analyzed whether children chose the response that indicated a controlled experiment in the CVS tasks. These data are shown in Figure 3.2.

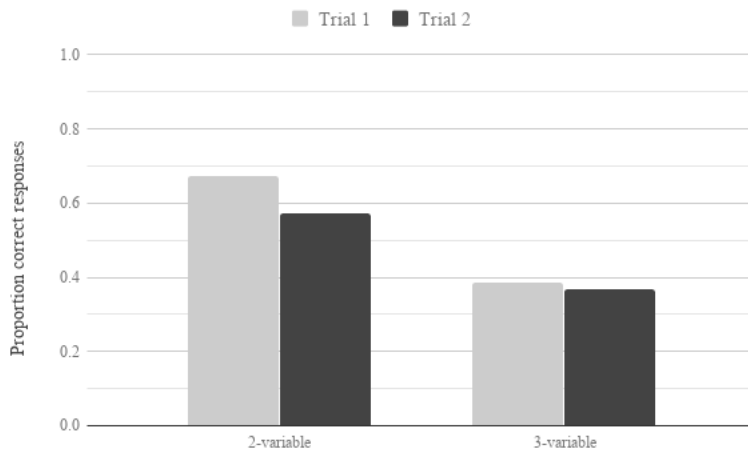


Figure 3.2. Choice performance on the CVS tasks (2-variable chance = 50%, 3-variable chance = 33%).

As a preliminary analysis, we built a GEE looking at the role of gender and experimenter on children’s responses. The model revealed a main effect of gender, $B = 0.37$, $SE = 0.15$, [95% CI = 0.09, 0.65], Wald $\chi^2(1) = 6.47$, $p = .01$, such that girls were more likely to select the correct response. Figure 3.3 examines this result in more detail by looking at the performance of boys and girls separately on each trial. Since we did not have any theoretical reason to expect an effect of gender and as there were no effects of gender in the studies in Chapter 2, we chose to complete the same analyses as previously, without including gender in the equations.

For our main analysis, we constructed a GEE to control for within-subject responses examining whether children chose the response that indicated a controlled experiment on the CVS tasks, looking at the role of age, task (i.e., 2-variable vs.

3-variable), trial, and performance on the first trial of the ICE task. This model revealed a main effect of task, $B = 1.02$, $SE = 0.15$, [95% CI = 0.72, 1.32], Wald $\chi^2(1) = 43.97$, $p < .001$. In the first trial of the 2-variable task, 67% of children chose the controlled test, greater than expected by chance, $\chi^2(1) = 21.81$, $p < .001$. Across two trials of the 2-variable task, 36% of the children selected the correct choice on both trials, 52% selected the correct test on one of the trials, and 12% of children selected the incorrect test on both trials. This distribution was different than expected by chance, $\chi^2(2) = 22.57$, $p < .001$, Cohen's $w = 0.35$. In the first trial of the 3-variable task, 39% of children chose the controlled test, no different than expected by chance, $\chi^2(1) = 1.88$, $p = .14$. Across the two trials of the 3-variable task, 17% of the children selected the correct test twice, 41% selected the correct test once, and 42% of children selected the incorrect test twice. This pattern of performance was significantly different from chance, $\chi^2(2) = 6.52$, $p = .04$, Cohen's $w = 0.19$.

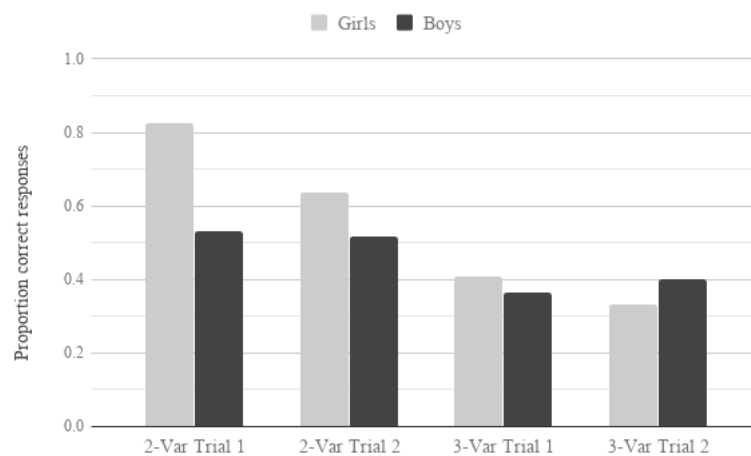


Figure 3.3. Choice performance on the CVS tasks split by trial and gender

We were also interested in children's justifications for their selections in the CVS trials. Eleven percent of their justifications were relevant to CVS. We constructed a GEE to control for within-subject responses examining children's justifications, looking at the

role of age, task, trial, choice, and performance on the first trial of the ICE task. The model revealed a main effect of task, $B = -0.80$, $SE = 0.24$, [95% CI = -1.28, -0.33], Wald $\chi^2(1) = 11.00$, $p = .001$, a main effect of choice, $B = -1.35$, $SE = 0.34$, [95% CI = -2.02, -0.68], Wald $\chi^2(1) = 15.49$, $p < .001$, and a main effect of robust ICE, $B = -0.86$, $SE = 0.39$, [95% CI = -1.63, -0.09], Wald $\chi^2(1) = 4.75$, $p = .03$. Children provided relevant justifications more often for the 3-variable CVS task (14%) than for the 2-variable CVS task (9%). Children provided relevant justifications more often for correct responses (16%) than for incorrect responses (6%). Finally, children who provided a robust ICE response in the first trial were less likely to provide relevant justifications.

3.2.2.2 Relation between Scientific Reasoning and other cognitive abilities

A correlation table including each trial of the three scientific reasoning tasks, age, gender, intelligence, language, the four executive function tasks, and the three Theory of Mind tasks can be found in Appendix G. The three Theory of Mind tasks were correlated with each other, so we combined them into a general Theory of Mind score. Both trials of the ICE task were highly correlated and both trials of the 3-variable task were correlated. The two trials of the 2-variable task were not correlated, but the second trial of the 2-variable task correlated with the 3-variable tasks. As a first step, we combined the two trials of the ICE task into one ICE score and the four trials of the CVS task into one CVS score. In a second step, we chose to combine the scores from all trials of the three scientific reasoning tasks to create a general scientific reasoning score. The means, standard deviations, ranges, and sub-sample sizes for age, intelligence, language, executive functions, Theory of Mind, and scientific reasoning are displayed in Table 3.1. We will first describe the relations between the other variables before we report the results of the ICE and CVS scores, followed by the results of the scientific reasoning score.

Table 3.1

Descriptives for variables of interest.

Variables	Mean	SD	Min	Max	<i>n</i>
Age	48.53	1.55	39.89	51.26	185
Intelligence	16.34	3.70	4 (0)	24 (28)	177
Language	22.57	7.59	0	35 (36)	170
Working Memory	1.65	2.07	0	8	148
Inhibition	0.73 ^a	.25	0	1	179
Planning	1.61	1.18	0	4	165
Cognitive Flexibility	0.79	0.55	0	2	130
Theory of Mind	0.59 ^a	.37	0	1	171
Scientific Reasoning	3.65	1.91	0	11 (12)	180
ICE	1.35	1.18	0	4	187
CVS tasks	2.27	1.37	0	8	187
2-variable	1.36	0.78	0	4	187
3-variable	0.91	0.99	0	4	187

Notes. Age in months; () absolute minimums and maximums in brackets; ^a represents proportion correct

Theory of Mind & Executive Functioning. *Theory of Mind* was positively correlated with age, intelligence, language, working memory, inhibition, and planning. *Theory of Mind* was not correlated with gender or cognitive flexibility. *Working memory* was positively correlated with age, intelligence, language, planning, cognitive flexibility, and *Theory of Mind*, and marginally related to inhibition. *Working memory* was not correlated with gender. *Inhibition* was positively correlated with language, planning, and *Theory of Mind*, and marginally to working memory. *Inhibition* was not correlated with

age, gender, or intelligence. *Planning* was positively correlated with intelligence, language, working memory, inhibition, and Theory of Mind. Planning was not correlated with age, gender, or cognitive flexibility. *Cognitive flexibility* was correlated with age, intelligence, language, and working memory. Cognitive flexibility was not correlated with gender, inhibition, planning, or Theory of Mind.

ICE & CVS Scores. Simple bivariate correlations between the variables of interest are presented in Table 3.2. The ICE score was significantly correlated with intelligence, planning, and Theory of Mind. ICE was not correlated with age, gender, language, working memory, inhibition, or cognitive flexibility. A multiple linear regression analysis was conducted to predict ICE from intelligence, planning, and Theory of Mind. The model including these effects on scientific reasoning was significant, $F(3,147) = 5.76, p = .001, R = .32, R^2 = .11, R^2_{Adjusted} = .09$, accounting for approximately 11% of the variance in ICE score. The regression coefficients are presented in Table 3.3. Neither intelligence nor planning contributed to the multiple regression model. Theory of Mind had a significant positive regression weight, indicating that children with higher Theory of Mind scores were expected to have higher ICE scores.

The CVS score was significantly correlated with gender (girls performed better) and significantly positively correlated with inhibition and Theory of Mind. CVS was not correlated with age, intelligence, language, working memory, planning, or cognitive flexibility. A multiple linear regression analysis was conducted to predict CVS from gender, inhibition, and Theory of Mind. The model including these effects on CVS was significant, $F(3,160) = 5.72, p = .001, R = .31, R^2 = .10, R^2_{Adjusted} = .08$, accounting for approximately 10% of the variance in CVS score. The regression coefficients are presented in Table 3.4.

Table 3.2
Correlations between variables of interest.

Variables	1	2	3	4	5	6	7	8	9	10	11
1. Age	–										
2. Gender	-.01	–									
3. Intelligence	.12	-.07	–								
4. Language	.10	.03	.62***	–							
5. EF - Working Memory	.20*	.07	.35***	.38***	–						
6. EF - Inhibition	-.07	-.06	.11	.16*	.15 [†]	–					
7. EF - Planning	.03	.03	.26**	.21*	.41***	.20*	–				
8. EF – Cognitive Flexibility	.18*	.03	.32***	.33***	.40***	.09	.11	–			
9. Theory of Mind	.16*	-.10	.26**	.38***	.18*	.22**	.28***	.08	–		
10. ICE Score	.10	-.11	.15*	.10	-.04	.11	.17*	.09	.30***	–	
11. CVS Score	-.04	-.16*	.09	.03	.11	.25**	.04	.07	.17***	.13 [†]	–
12. Scientific Reasoning	.02	-.15*	.11	.05	.03	.22**	.13	.06	.29***	–	–

Neither gender nor Theory of Mind contributed to the multiple regression model. Inhibition had a significant positive regression weight, indicating that children with higher inhibition scores were expected to have higher CVS scores.

Scientific Reasoning Score. Simple bivariate correlations between the variables of interest are presented in Table 3.2. The scientific reasoning score was significantly correlated with gender (girls performed better), and significantly positively correlated with inhibition and Theory of Mind. Scientific reasoning was not related to age, intelligence, language, working memory, planning, or cognitive flexibility. A multiple linear regression analysis was conducted to predict scientific reasoning from gender, inhibition, and Theory of Mind. The model including these effects on scientific reasoning was significant, $F(3,155) = 8.25, p < .001, R = .37, R^2 = .14, R^2_{Adjusted} = .12$, accounting for approximately 14% of the variance in scientific reasoning score. The regression coefficients are presented in Table 3.5. Gender did not contribute to the multiple regression model. Both inhibition and Theory of Mind had significant positive regression weights, indicating that children with higher scores on inhibition and Theory of Mind were expected to have higher scientific reasoning scores.

To further examine the relation between Theory of Mind and scientific reasoning, we created dichotomous mastery categories for both scientific reasoning and Theory of Mind (mastery \geq 50% correct) and conducted a McNemar's test. This significant McNemar's test ($p < .001$) revealed that Theory of Mind is an important precursor for scientific reasoning: While 56% of children ($n = 95$) showed mastery of Theory of Mind but not a mastery of scientific reasoning, less than 2% of children ($n = 2$) displayed the opposite pattern of no mastery of Theory of Mind but mastery of scientific reasoning (see Table 3.6; see Appendix H for tables for each of the three scientific reasoning tasks).

Table 3.3

Regression analysis predicting ICE.

Variables	<i>B</i>	<i>SE B</i>	β	<i>t</i>	<i>p</i>
(Constant)	.37	.41		0.89	.38
Intelligence	.02	.03	.06	0.75	.46
Planning	.11	.08	.12	1.30	.20
ToM	.83	.27	.25	3.06	.003

Table 3.4

Regression analysis predicting CVS.

Variables	<i>B</i>	<i>SE B</i>	β	<i>t</i>	<i>p</i>
(Constant)	1.72	.51		3.38	.001
Gender	-.38	.21	-.14	-1.78	.07
Inhibition	1.22	.45	.21	2.74	.007
ToM	.47	.30	.11	1.48	.14

Table 3.5

Regression analysis predicting scientific reasoning.

Variables	<i>B</i>	<i>SE B</i>	β	<i>t</i>	<i>p</i>
(Constant)	2.80	.69		4.07	<.001
Gender	-.55	.29	-.14	-1.89	.06
Inhibition	1.33	.60	.17	2.21	.03
ToM	1.33	.41	.25	3.27	.001

Table 3.6

Crosstabulation of mastery of Theory of Mind and mastery of scientific reasoning.

Theory of Mind	Scientific Reasoning		Total
	No mastery	Mastery	
No mastery	51 (30.0%)	2 (1.2%)	53 (31.2%)
Mastery	95 (55.9%)	22 (12.9%)	117 (68.8%)
Total	146 (85.9%)	24 (14.1%)	170 (100%)

This finding suggests that Theory of Mind is needed to be successful in scientific reasoning, but mastery of Theory of Mind does not guarantee success in scientific reasoning.

3.2.3 Discussion

In this study, we investigated four-year-olds' abilities in scientific reasoning and the relation between scientific reasoning and other cognitive abilities. We will first discuss children's performance on the scientific reasoning tasks in relation to the findings of Study 2b before discussing the relation between scientific reasoning and inhibition and Theory of Mind.

3.2.3.1 *Scientific Reasoning tasks*

Children's performance on the ICE task in the present study was similar to children's performance in Study 2b. In both cases, children were better able to correctly claim that they could not know which bricks made the box light up than to provide a reason for why they could not know that. Additionally, when providing a reason, children performed better on the first trial of the task than on the second trial in both studies. As discussed in Chapter 2, it is possible that this decrease in performance is a result of fatigue, and children were no longer willing to provide explanations by the end of the study, the sixth time they were asked to explain something. Overall performance was worse in the present study than in Study 2b, but this is to be expected considering there was an effect of age on both knowledge claims and explanations in Study 2b. The present study was conducted with children on the younger end of the age range in Study 2b, and the metacognitive abilities likely implicated in responding correctly to this task are developing around five years of age (Perner, 1991; Wellman, 1985).

Regarding performance on the CVS tasks, the present study found an effect of gender, specifically on the first trial of the 2-variable task, in which girls were more likely

to select the correct choice than boys. This finding was surprising considering we found no effects of gender in Studies 1, 2a, or 2b. Further, most studies show that gender is not related to performance on scientific reasoning measures (Astington et al., 2002; Bauer & Booth, 2019; Bullock et al., 2009; Koerber et al., 2015; Sodian et al., 2016; van der Graaf et al., 2018), and studies that have found gender differences generally show that boys perform better on measures of science achievement and reasoning throughout school (J.-T. Kuhn & Holling, 2009; Saçkes et al., 2011). This effect may be unrelated to children's scientific reasoning abilities. Taking a closer look at Trial 1 (refer to the procedure in Figure 2.4), it is possible that children's (specifically girls') color preferences affected their responses: the correct choice included a purple block. That being said, we attempted to design the materials in a way that would not influence the choice of a correct or incorrect response, and in Studies 2a and 2b in which these task materials were counterbalanced, there was no influence of specific task materials on children's performance. Interestingly, the effect of gender on the ICE tasks was marginally significant ($p = .07$) in the same direction, such that girls performed better. This would suggest that perhaps the effect of gender is something more than just an effect of task materials.

In general, the performance on the CVS tasks in the present study was again similar to performance in Study 2b. In both cases, there was an effect of task type, such that the 2-variable task was easier than the 3-variable task. And in both cases, the pattern of performance across both trials was significantly different from chance, with 17% correct in both trials of the 3-variable task in Study 2b and 16% in the present study, and with 46% correct in both trials of the 2-variable task in Study 2b and 36% in the present study. However, performance was better in the 2-variable task in Study 2b than in the present study. Looking at children's justifications, fewer justifications were considered

relevant in the present study (11%) than in Study 2b (38%). This is to be expected, again, considering there was an effect of age on children's justifications in Study 2b, and the children in the present study were on the younger end of the age range in Study 2b.

Finally, in the present study, there was an effect of robust performance in the first ICE trial on children's justifications. This effect was not found in Study 2b. However, this effect was in the opposite direction than one would expect. Our reason for including the first ICE trial in the analysis was to investigate if being able to explain why one cannot know something would be related to being able to provide a justification for an experimental design choice. However, the effect revealed that children who provided a robust ICE response were less likely to be able to provide a relevant justification for their choice. Theoretically, we do not have an explanation for this effect. Methodologically, it is possible that children who provided a robust response in the first trial of the experiment felt less inclined to provide additional explanations later on in the experiment, possibly resulting, again, from fatigue.

3.2.3.2 Relation between Scientific Reasoning and other cognitive abilities

We now move to the discussion of the results on the structure and correlates of scientific reasoning. When looking at the ICE and CVS tasks separately, we saw that ICE was significantly positively correlated with intelligence, planning, and Theory of Mind, while CVS was significantly correlated with gender and significantly positively correlated with inhibition and Theory of Mind. The fact that the two tasks show different relations with other cognitive measures could suggest that they indeed measure different aspects of scientific reasoning. Further, ICE was predicted by Theory of Mind, while CVS was predicted by inhibition.

When looking at a general scientific reasoning score, we saw that scientific reasoning was significantly correlated with gender and significantly positively correlated

with inhibition and Theory of Mind. The effect of gender was discussed in detail in the previous section. The positive correlation with inhibition makes sense both in terms of the ICE task and the CVS tasks. In the ICE task, children have to inhibit a natural bias to answer in the affirmative in order to correctly claim that they cannot know something. In the CVS tasks, children have to inhibit the incorrect choices, e.g., the novel colors, in order to select the correct choice representing a controlled test of a hypothesis.

The positive relation with Theory of Mind is also to be expected based on the literature in which a number of studies have shown that false belief understanding and (advanced) Theory of Mind abilities are related to several aspects of scientific reasoning, for example, experimentation, understanding evidence and justifications, as well as the understanding of the nature of science (Astington et al., 2002; Klein, 1998; Koerber & Osterhaus, 2019; Osterhaus et al., 2017; Piekny et al., 2013; Sodian et al., 2016). In the ICE task, children have to have a metacognitive understanding of their own ignorance and distinguish between their beliefs and the evidence they observed. In the CVS tasks, children have to have an understanding of alternative possibilities, i.e., the box could light up or not light up when their choice is placed on it. They also have to distinguish between the hypothesis (that one particular brick makes the box light up) and the evidence that different experimental design choices would produce. The pattern of performance showing that mastery of Theory of Mind appears to be necessary for mastery of scientific reasoning and not the reverse, suggests a direction to this relation.

Scientific reasoning was not related to age, though the CVS tasks were also not related to age in Study 2b, which had a much larger age range. It seems that the development of scientific reasoning abilities is more related to other individual cognitive abilities than to a general development with age. Scientific reasoning was not related to intelligence or language. The lack of a relation to intelligence may be explained by the

intelligence measure used. In the present study, we used a measure of general knowledge, which was also highly related to language. The scientific reasoning tasks in the present study are knowledge-lean, thus, it may be expected that they do not relate to a measure of intelligence that is knowledge-rich.

The lack of a relation to language is somewhat surprising considering the literature showing that a number of different language measures have been shown to be related to scientific reasoning. In particular grammatical abilities appear to be related to scientific reasoning (van der Graaf et al., 2016, 2018) and our language measure assessed children's grammar, specifically their understanding of morphological rules. Further, language was related to all other measures (except the Content False Belief task). This finding may suggest that the questions in the present study were not so difficult that they required extensive language processing. However, this finding may also suggest that children did not need to rely on their language abilities or the understanding of the questions to correctly solve at least the CVS tasks. Instead, children may have been able to solve the CVS tasks using a lower level perceptual similarity matching strategy.

Finally, scientific reasoning was not related to working memory, planning, or cognitive flexibility. Previous research has also found no relation between scientific reasoning and cognitive flexibility (van der Graaf et al., 2016, 2018). The relation between working memory and scientific reasoning is inconclusive in the literature, with some studies finding a relation and some not, and particularly when there is a distinction between verbal and non-verbal working memory (Pieknny et al., 2013; van der Graaf et al., 2016; 2018). However, van der Graaf and colleagues found that verbal working memory was related to scientific reasoning. The working memory task in the present study, the Backwards Digit Span, is also a measure of verbal working memory. Considering that we did not find a relation between language and our tasks, this may explain the lack of a

relation between verbal working memory and our tasks as well. Additionally, we had no expectations based on the literature regarding a relation between planning and scientific reasoning.

In summary, the present study reveals that even four-year-olds can consistently select a controlled test of a hypothesis more often than expected due to chance. However, they struggle to provide a reason for their experimental design choice, with only 11% of justifications being relevant to CVS. Over two-thirds of four-year-olds showed some recognition of the inconclusiveness of confounded evidence by correctly claiming a lack of knowledge about what makes the box light up at least once. Again, very few could provide an explanation for that lack of knowledge. In the present sample, gender seemed to play a role in these scientific reasoning abilities, with girls performing better than boys. The relation between scientific reasoning and inhibition is consistent with needing to inhibit a positive claim of knowledge or the incorrect choice in the CVS tasks and with the literature showing that inhibition is related to scientific reasoning and experimentation (Bauer & Booth, 2019; Kwon & Lawson, 2000; Osterhaus et al., 2017; van der Graaf et al., 2016, 2018). The relation between scientific reasoning and Theory of Mind is also consistent with the literature on distinguishing between hypotheses and evidence and the ability to represent alternative hypotheses or outcomes. The present study cannot definitively rule out the possibility that children solve the CVS tasks on the basis of perceptual matching; however, the relation to Theory of Mind would suggest that children are not solely relying on perceptual similarity to solve the CVS tasks. Future studies should investigate this further.

4 The Development of Educational Tools for Assessing and Promoting Control of Variables Strategy Abilities in Preschool

4.1 Introduction

In the previous chapters, we described in detail the use of the knowledge-lean Lego Control of Variables Strategy (CVS) task for assessing young children's abilities in control of variables strategy. We developed the knowledge-lean task for assessment purposes to avoid the influence of prior science content knowledge and prior beliefs on the assessment of children's abilities. In this chapter, we will discuss the potential for promoting CVS abilities. To this end, we chose to develop tools for training CVS using video and tablet applications and to introduce CVS in child-friendly and hopefully engaging contexts.

The results from the studies in Chapters 2 and 3, as well as the literature reviewed in Chapter 1, have revealed that young children have nascent abilities in scientific reasoning and specifically the Control of Variables Strategy. However, preschoolers are still far from fully competent, which presents the possibility of improving their abilities further. Indeed, a large portion of the CVS literature has focused on promoting these abilities through training and intervention studies (see Schwichow et al., 2016 for a review and meta-analysis). CVS trainings have been shown to be effective for children of all ages, starting in middle elementary school and also across different task domains (Schwichow et al., 2016). However, there is little research on training CVS abilities in preschoolers, though this is unsurprising considering there is also little research on preschoolers' abilities in CVS in general.

As discussed in Chapter 1, the context and content of tasks can affect children's ability to perform the task. When children have prior beliefs about task content, they

ignore or distort data that does not fit their beliefs (Kuhn et al., 1988). Further, if children already have knowledge about the task content, they could end up with correct performance based on their content knowledge rather than on their scientific reasoning abilities. However, when trying to promote CVS abilities, it is important that children are actively engaged in the task and are motivated to solve a problem or figure something out. For this reason, we chose to develop training tools within a “fun,” engaging context. In one case, we adapted the plane context from Bullock and Ziegler (1999), and in the second case, we developed a farm animal context inspired by Saffran and colleagues (2015).

Children up to eight years of age in the U.S. spend, on average, 2 hours and 20 minutes per day with “screen media” either watching television, watching videos on a phone or tablet, or interacting with applications (Rideout, 2017). Children, now, are digitally fluent from a very young age (Palaiologou, 2016). They are considered to be “digital natives” because they are born into an interactive world full of technologies, and they grow up learning how to use them, as opposed to adults who have had to discover and understand the digital world in adulthood (Prensky, 2001). They grow up with their parents’ phones or tablets in their hands and learn to interact with them from a very early age. In the U.S., 78% of families with children eight or younger own a tablet, and 42% of children under nine have their own personal tablet (Rideout, 2017). In the U.K., 65% of children three to four years old used a tablet at home (Ofcom, 2017).

Videos purporting to be educational are growing in numbers. A search for “educational videos preschool” on one of the most popular platforms for such videos, YouTube, produces over 2 million results (“YouTube,” 2019) and a recent survey revealed that, in the U.S., 81% of parents of children 11 years or younger allow their child to watch videos on YouTube (A. Smith, Toor, & van Kessel, 2018). Looking at digital applications, the Apple App Store boasts over 75,000 educational apps across all grades, subjects, and

learning styles (“Apple,” 2019). Educational videos and applications can keep young children entertained in the home, allowing the caregiver to accomplish other tasks or can provide educational pre-school content to children of parents who cannot afford to send their children to daycare or preschools. When content is designed appropriately, educational videos and applications can entertain and engage children and offer opportunities for learning.

Additionally, one of the reasons educational videos and applications are being produced at such a rate and are gaining popularity in use among parents and educators is the ease with which they can be used and integrated into daily life or school. This ease of use can be beneficial in research contexts as well. If everything needed to assess or train CVS abilities, for example, is located in one tablet, this simplifies testing procedures and reduces the need for lots of materials as well as the potential for experimenter or technical errors.

The goal of this Chapter and the studies herein are two-fold. First, we aimed to investigate and determine important factors for designing and evaluating CVS training tools for preschoolers. Second, we aimed to investigate if said tools could be used both to assess and promote preschool children’s CVS abilities. To this end, we begin with a brief review of the literature on teaching the Control of Variables Strategy before continuing with a broader review on children’s use of digital media, important instructional design theories, as well as design and evaluation criteria particularly important for our target group, preschoolers. The first two studies in this Chapter discuss the iterative design and evaluation process of a tablet application for assessing and training CVS abilities. The third study presents a video tutorial for teaching CVS. All three studies also investigate children’s abilities in CVS using the respective digital media.

Our more far-reaching goal with this work is to spark a discussion regarding the promotion of scientific reasoning abilities in early childhood education. Throughout this dissertation, we have emphasized the importance of scientific reasoning abilities for children and adults, for scientists and lay-people, and have referenced findings showing that the general population is not fully competent in scientific reasoning and could stand to benefit from more intensive instruction and training. The easiest way to accomplish this would be to place more focus on the promotion of these abilities throughout education. Indeed, there are already measures in place (e.g., American Association for the Advancement of Science [AAAS], 2009; European Commission, 2015; Next Generation Science Standards [NGSS], 2013; United Nations Educational, Scientific, and Cultural Organisation [UNESCO], 2005/2014). The work presented in the earlier chapters of this dissertation suggests that preschoolers have beginning abilities in CVS. We believe that, because of these findings, it could be appropriate to start introducing the concepts of scientific reasoning and to actively begin to promote these abilities much earlier than is already the case. We do not intend to argue that preschoolers should (or can) become and remain fully competent in scientific reasoning but that beginning to promote these abilities earlier rather than later may lay the foundation for higher-level abilities to develop earlier. To place the present studies in this broader perspective, they represent initial steps, first, in investigating how to develop training tools to be appropriate for this age group and second, in investigating whether such tools can successfully be used to assess and promote CVS abilities.

4.1.1 Promoting Control of Variables Strategy abilities

Because the Control of Variables Strategy has been shown to be important to broader scientific reasoning abilities and for learning about science in general (e.g., Bryant, Nunes, Hillier, Gilroy, & Barros, 2013), much research has been conducted to

determine how best to promote CVS abilities. A number of factors have been considered in this research, for example, the method of instruction, the content and type of tasks, the level of difficulty, the duration of training, whether the training takes place in the lab or in the classroom, as well as what type of assessment is used to measure learning and how long after the training the assessment takes place. In addition, individual factors such as age, achievement level, or cognitive abilities (as we saw in Chapter 3) could play a role in the effectiveness of training.

Perhaps one of the most investigated factors for CVS instruction is the method through which CVS is taught. Research in this area rises out of the discussion of whether direct, explicit instruction or discovery/inquiry learning is more effective. Direct instruction tends to be teacher-centered in that an instructor explicitly states what is to be learned. In the case of CVS, direct instruction could take the form of an instructor explaining that to design and perform a controlled experiment, one must only vary one variable at a time while keeping all other variables constant (e.g., Chen & Klahr, 1999).

Alternatively, discovery learning arose from the constructivist camp (Piaget, 1970) and the idea that children need to construct knowledge on their own to truly understand concepts in a deep and meaningful way (Dean & Kuhn, 2007; Hmelo-Silver, Duncan, & Chinn, 2007). In addition, having acquired the knowledge on their own facilitates their ability to extend and transfer that knowledge to other problems (Schauble, 1996). In the case of CVS, discovery learning would allow children to design experiments and observe outcomes and through this process, recognize that when they design confounded experiments, they cannot be certain of the effects of particular variables.

But, the task of having to discover knowledge oneself could result in misunderstandings and a lack of appropriate feedback from the system (Chen & Klahr, 1999; Klahr & Nigam, 2004). It also likely puts a heavy cognitive load on children who

have to devote their efforts to figuring something out rather than learning the knowledge which could instead be presented clearly to them (Kirschner, Sweller, & Clark, 2006; R.E. Mayer, 2009; Sweller, 1988). Another alternative, which provides more support than discovery learning, but does not explicitly teach concepts as in direct instruction, is the scaffolded instruction approach, which has been shown to be as effective as direct instruction (Lazonder & Egberink, 2013; Sao Pedro, Gobert, Heffernan, & Beck, 2009; Sao Pedro, Gobert, & Raziuddin, 2010).

The majority of intervention studies investigating this question tend to find that CVS is best promoted through direct instruction (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004; Lorch et al., 2010; Matlen & Klahr, 2013; Sao Pedro et al., 2009; Strand-Cary & Klahr, 2008; Toth et al., 2000; Triona & Klahr, 2003; Wagensveld et al., 2015; Zohar & Aharon-Kravetsky, 2005; Zohar & David, 2008; Zohar & Peled, 2008). However, a recent meta-analysis (Schwichow et al., 2016) found that the effectiveness of teaching CVS did not depend on whether the instruction included an explicit CVS rule or not, suggesting that training does not necessarily have to be direct to be effective. Below, we describe in detail the study by Chen and Klahr (1999) to illustrate an intervention study with direct and indirect instruction and various transfer tasks.

Chen and Klahr (1999) investigated the possibility of training CVS abilities in early elementary school-age children, approximately seven to ten years old (2nd, 3rd, 4th grade). Specifically, they were interested in the effects of direct and indirect instruction on the acquisition of CVS abilities. Their intervention consisted of five phases occurring over three different time points. They used slopes, springs, and sinking as the content domains. On the first day of the intervention, children were introduced to the materials and given the opportunity to explore the materials, name the variables, and were given a conceptual knowledge assessment about the effects of two of the variables. They were then allowed to

produce two tests for each of two target variables and answered probe questions (Exploration phase).

Following the Exploration phase, children were assigned to one of three groups for the Intervention phase: Training-Probe, No Training-Probe, and No Training-No Probe (Control). Training consisted of explicit training of CVS by providing examples of confounded and unconfounded experiments. Children were asked if the experiments were good or bad comparisons and why. The experimenter then followed up with correct judgements and explanations. Probe questions were asked during the Assessment phases after the children designed their own experiments and consisted of questions such as why they designed their experiment in that way and what could they conclude from their experiment. The Assessment phase required children to produce two comparisons for two target variables.

One week later, children performed two transfer tasks in which they had to identify variables, explain how they could affect the outcome, produce comparisons for two target variables, and provide their reasoning for the comparisons, as well as what they could conclude from them. Seven months later, children were given a paper-and-pencil post-test in which they had to evaluate comparisons as being a “good test” or a “bad test.” All the comparisons were made with three variables, each with two levels.

Children’s use of CVS in designing unconfounded comparisons was scored (*CVS score*), and a more stringent measure of their verbal justifications in combination with their CVS score was scored (*Robust use of CVS*). Robust use of CVS required children to provide explanations that included mentions of CVS, i.e., controlling all other variables. Chen and Klahr (1999) found that direct instruction of the Control of Variables Strategy improved children’s ability to design and understand unconfounded experiments, while the children in the control condition did not show any improvement. Use of CVS

increased from 34% of trials before training to 65% of trials after training, and this improvement was still apparent one week later in the two transfer trials.

They did, however, find effects of age such that only the 3rd and 4th graders were able to transfer their learning to other tasks a week later, and only the 4th graders were able to do so in the far transfer task seven months later. Almost 50% of the children in the training-probe group were considered “good experimenters,” by designing unconfounded experiments in seven out of eight comparisons in the transfer phases, compared to 22% in the no training-probe and 13% in the control group. These results show that explicit training through direct instruction, combined with probe questions as indirect instruction, was the most effective way to teach CVS and that it was more effective than indirect probe questions alone. Toth, and colleagues (2000) found similar results in a classroom setting, suggesting that such instruction can also be more broadly applied in a classroom environment.

Though in the studies described above, 2nd graders did not seem able to learn and retain the ability to apply the CVS skills in delayed transfer tasks, Chen and colleagues (2011) found that even six- to eight-year-old children can be taught to generate a valid test of a hypothesis in tasks similar to Sodian and colleagues’ (1991) Mouse House task and that even a year later they maintain around 60% correct performance (Chen et al., 2011 in Chen, 2012). Chen and colleagues, however, also found an effect of age, such that 2nd graders performed best (~80%), 1st graders followed (~65%), and kindergarteners were least successful (~25%) at the last assessment. Earlier research has also shown, similarly to Chen and colleagues (2011), that seven- and eight-year-old children benefit from CVS training (Case, 1974).

Another factor to consider in teaching CVS is whether the materials used for instruction are physical or virtual. Though most interventions use physical materials, such

as the ramps task (e.g., Chen & Klahr, 1999), some studies have also investigated if digital tasks can be used for training and have generally found that virtual tasks or simulations are effective and when comparing them with physical tasks, there is no difference in performance (e.g., Kittredge, Klahr, & Willows, 2015; Sao Pedro et al., 2009; Triona & Klahr, 2003; Van de Keere, Mestdagh, Dejonckheere, Vervaeke, & Tallir, 2014).

Two meta-analyses have been conducted to investigate the effectiveness of interventions for teaching the Control of Variables Strategy (Ross, 1988; Schwichow et al., 2016). In Ross's 1988 meta-analysis of 65 intervention studies, he found that CVS could indeed be taught (mean effect size $d = 0.73$). He also found that a number of factors moderated how effective CVS instruction was. For example, interventions that only focused on teaching CVS were more effective than studies also teaching additional skills. Studies that provided practice tasks with both in-school and out-of-school contexts were more effective than studies which only provided one context. Studies, where students received feedback, had larger effect sizes than those without feedback. Effect sizes were also larger when students were assessed on tasks that were similar to what they had encountered in the training as opposed to novel tasks. Finally, when children had to identify the relevant variables themselves, effect sizes were larger than when the variables were identified for them.

Schwichow and colleagues (2016) also found that CVS could be taught through intervention studies (mean effect size $d = 0.61$). Though they found a smaller overall effect size, this was comparable to Ross' meta-analysis when outliers were removed ($d = 0.61$). However, Schwichow and colleagues did not find many significant moderators of effectiveness of instruction. The use of demonstration and cognitive conflict in instruction made training more effective and larger effect sizes were found when the assessment task was a real hands-on task as opposed to written open-response, multiple-

choice assessments, or virtual assessments. Demonstration involved teachers showing children examples of correct experimental procedures, and cognitive conflict was used to support children recognizing that experimental strategies were not appropriate (without explicitly referencing CVS). This could mean, for example, pointing out that they cannot really know something for sure from a confounded experiment. Interestingly, they did not find effects of age, direct vs. discovery instruction, type of training tasks, use of feedback, contexts, number of variables, identification of variables by the instructor, duration of intervention, or delay between intervention and assessment.

Based on these findings, we believe it should be appropriate to use virtual training tasks to teach the Control of Variables Strategy. It is unclear, however, if children as young as preschool age will benefit from training. The few findings from previous studies suggest that six-year-olds showed limited improvement as the result of interventions teaching CVS (Case, 1974; Chen, 2012). Specifically, we use a direct instruction method in the studies presented in this Chapter. Further, in Study 6, we attempt to induce cognitive conflict in the tutorial by presenting examples of confounded experiments and both asking and pointing out that such an experiment is not a “fair” or “good” experiment because one cannot know to which variable the outcome is due.

4.1.2 Children, media, & technology

The use of media and technology in early childhood is a topic of frequent, sometimes heated, discussion among researchers and in the public. On the one hand, there is support for the value technology can bring to education and educational settings, and on the other hand, there is concern about the impact on children’s cognitive, emotional, and social development (Plowman & Stephen, 2003). What is important to consider, however, is the quality of the content that children are consuming during screen time and how it is being consumed. For example, children can use digital media and toys with adults, with

other children, or by themselves, and the content can be entertaining, or educational, or both (Anand & Krosnick, 2005). In terms of parents' beliefs about technology and media use, 67% of American parents think that screen media helps their child's learning (Rideout, 2017).

This belief seems to be supported by empirical research showing that digital media can have positive effects on learning. In a recent study, researchers showed that by playing the popular application *Angry Birds*, which involves sling-shooting birds at a structure to break it down and save their eggs from enemy pigs (Rovio Entertainment Oyj), five-year-olds showed an improvement in learning about how force affects projectile motion and predicting the parabolic pathway of an object. The authors concluded that interest in and motivation to learn science could be fostered through engagement with mobile games (Herodotou, 2018).

Previously, research has shown that digital media can introduce children to concepts such as mathematics or dynamic systems (Elliott & Hall, 1997; Resnick, 1998), can engage children in reasoning and problem-solving (Crawley, Anderson, Wilder, Williams, & Santomero, 1999; Lieberman & Linn, 1991; Yelland, 2005), and can improve vocabulary, spelling, and reading (Din & Calao, 2001). However, other factors must also be taken into account when looking at learning with technology. For example, a systematic review of the literature by Ching-Ting Hsin and colleagues (Ching-Ting Hsin, Ming-Chaun Li, & Chin-Chung Tsai, 2014) revealed that for children up to eight years old, older children showed greater improvement and performed better overall in interventions with technology, children with more prior content knowledge were more successful in learning with technology, and children who had more access to technology showed better performance in learning with technology.

When looking at the types of devices that young children are using and the devices that are popular in educational settings, tablets seem to be preferred. Use of tablets in school has been associated with positive learning outcomes (Haßler, Major, & Hennessy, 2016). There are many reasons why tablets are the device of choice. Tablets are mobile and have large screens which are larger than those of mobile phones. Further, the touchscreen functionality is intuitive and easy for children to learn how to use. For example, studies have shown that preschool children quickly learn to use tablets and can use them independently and confidently (Chiong & Shuler, 2010; Couse & Chen, 2010).

Although the functionality of the device itself is important, the design of the interface and the applications is also critical to ensuring that the interaction is intuitive and appropriate for children. In the next sections, we will discuss the importance of designing for learning and some specific design guidelines for digital media for young children.

4.1.3 Instructional design

When designing materials for learning, it is important to consider how people learn and what can help or hinder learning. The goal of instructional design is to systematically translate principles of learning and instruction into instructional materials that enhance learning (P. L. Smith & Ragan, 2004).

An important theory of cognitive architecture and the capacities available for learning is the Cognitive Load Theory (CLT). Cognitive load is defined as the mental effort required during learning and problem-solving activities (Sweller, 1988). It is based on the idea that working memory capacity is limited (Miller, 1956) and that consideration must be given to how it is allocated. Cognitive load is further split into three types: intrinsic, germane, and extraneous. Intrinsic cognitive load is a result of the element interactivity of the material being learned. In other words, how difficult the material is for an individual. More complex tasks will generate higher cognitive load (Wouters, Paas, &

van Merriënboer, 2008). This type of load cannot be reduced because it depends on the content of the task itself.

Germane and extraneous cognitive load are a result of the design or materials and can be manipulated. Germane cognitive load is cognitive load that is related to or relevant to the task and is considered effective because it increases cognitive load in a way that is beneficial to learning (Paas, Renkl, & Sweller, 2003). Extraneous cognitive load, however, is unnecessary and can result from poorly-designed materials. Extraneous load should be reduced to allow cognitive resources to be allocated elsewhere. The three types of cognitive load are additive, and their combination cannot exceed the capacity of working memory (Miller, 1956). Cognitive Load Theory has implications for instructional design, such that intrinsic cognitive load should be at an appropriate level based on how complex a task is for a particular individual; germane cognitive load should be allocated to help individuals learn, for example by acquiring schemas; and extraneous cognitive load should be avoided, by ensuring that instructional materials are well designed and do not focus on irrelevant information or distract the learner.

Taking into account cognitive architecture and the implications of Cognitive Load Theory, R.E. Mayer and colleagues (2014) investigated instructional design in terms of multimedia learning and design. Multimedia is defined as presenting words (printed or spoken) and pictures (illustrations, photos, animation, or video) simultaneously. Multimedia learning is defined as the process of building mental representations from the words and pictures presented (R.E. Mayer, 2005). R.E. Mayer termed the robust finding that people learn more from information presented in both words and pictures than from information presented in words alone the *multimedia principle*. The basis for this finding comes, again, from cognitive architecture with the explanation that we have two information processing systems, one for verbal and one for visual material (Paivio, 1990).

So, when information is presented in both modalities, we have more capacity to process that information than if it were presented in only one modality.

Critically, though, the quality and type of visual material, as well as how it is added to words, can be more or less effective. Thus, R.E. Mayer (2002) outlined eight additional principles to guide the creation or selection of visual materials for multimedia learning. The *spatial and temporal contiguity principles* claim that verbal and visual information should be presented close to each other in both space (next to each other) and time (simultaneously). The *coherence principle* claims that information should be coherent and exclude extraneous material. The *modality principle* claims that verbal information presented as narration (audio) in addition to visual material is more effective than verbal information presented as text (visual) in addition to visual material. The *redundancy principle* claims that information should not be presented twice in different formats. The *pre-training principle* claims that individuals learn better when they are already familiar with the terms and main concepts of the instructed material. The *signaling principle* claims that cues that highlight the organization of the material can help individuals learn better. The *personalization principle* claims that individuals learn better when the verbal material is informal and conversational in style.

A number of additional principles are also relevant for multimedia theory. For example, the *split-attention principle* (Ayres & Sweller, 2005) claims that people learn better when they do not have to split their attention between the information presented, which goes hand-in-hand with the *spatial and temporal contiguity principles* described above. The *segmenting principle* claims that individuals learn better when information is presented in shorter, learner-paced chunks, rather than all at once (Mayer & Pilegard, 2005). The *voice principle* claims that individuals learn better when the voice presenting information is unaccented (relative to the listener) and human rather than computer-

generated (R.E. Mayer, 2014). Finally, the *transient information principle* (Sweller, Ayres, & Kalyuga, 2011) claims that because information presented as spoken words is not permanent but transient, spoken information can generate extraneous cognitive load. Learning is affected because the spoken information is gone before the learner had the chance to process it.

A model which can incorporate both the theories of human cognitive architecture and the multimedia principles described above is van Merriënboer and Kester's (2005) four-component instructional design (4C-ID) model for multimedia learning. This model consists of four components claimed to be necessary for complex learning: learning tasks, supportive information, procedural information, and part-task practice. In detail, learning tasks refer to authentic and meaningful real-life tasks; supportive information refers to information that, for example, describes the organization of the task domain or provides suggestions regarding how to approach the problem; procedural information refers to specific routine or algorithmic steps that can be used to solve a learning task; finally, part-task practice refers to practicing particular aspects of a task to achieve a level of automaticity, which can be applied to future learning tasks. Many of the principles described above also apply to the different components of the 4C-ID model.

Finally, another concept important for instructional design is the idea of scaffolding. The metaphor of scaffolding suggests that individuals can benefit from temporary supporting structures to help develop new understandings, concepts, and abilities (Hammond & Gibbons, 2005). The concept of scaffolding is based on Vygotsky's zone of proximal development (ZPD; Vygotsky, 1980). The ZPD is the metaphorical space between where an individual is in terms of their level of development of problem-solving ability and where they could be with the support of adults or more-able peers. Scaffolding is then used within the ZPD to support children past the point of their abilities

on their own. As children learn more, the scaffolding is removed, or rather moved higher, to support further learning.

4.1.4 Educational products

There are a number of factors to consider when designing and evaluating the “educational” property of products for children. McManis and Gunnewig (2012) developed a form for educators and researchers to complete when evaluating a product for its educational value, which can also be used as guidelines for designing such products. For the purpose of being educational, the product should emphasize learning rather than a focus on winning. The content should be based on research or learning standards and should follow the developmental course as well as effective teaching strategies. Further, the product should include informative feedback that supports learning. A product should be appropriate for the target audience. This includes the cognitive skills required, the subject matter, and the functionality of the product or device. The context should also be interesting and appealing.

The product should be child-friendly, which for McManis and Gunnewig (2012) meant that there should be clear and simple choices and that after initial support from an adult, children should be able to use the product independently as a result of clear, understandable instructions and integrated supports and prompts. Further, there should be multiple opportunities for success, such that children should be able to think again about a situation and try to apply a more effective strategy if the first attempt was unsuccessful. Products should be engaging and enjoyable to use. There should be a variety of activities available, and they should match well to the target audience’s attention span. Rewards should be used appropriately to encourage engagement. Features should be individually customizable for each child’s needs. Finally, the last criterion is specific to the researchers

or teachers: the product should have the ability to monitor progress and potentially assess performance and present this information in a way that is easy to interpret.

For educational products to be effective for learning, Sharples and colleagues (2007) determined that the content should be learner-centered, building on children's skills and knowledge and enabling them to reason from their own experience. Products should be knowledge-centered, such that the content is a result of validated knowledge and is taught efficiently. A product should use a variety of concepts and methods for teaching and use them inventively. The assessment of performance or progress should be matched to learners' abilities and should diagnose and guide successful learning. Lastly, educational products should be community-centered, allowing students to form knowledge-sharing communities and support less-able students (Sharples, Taylor, & Vavoula, 2007).

4.1.5 Designing for preschoolers

In the early 2000s, researchers began to discuss the roles children could play in the design of technology for children (Druin, 1999, 2002). Traditionally, adults designed products for children, keeping children in mind, but not involving them in the design process. In this case, children take on the role of "user" in that they use a technology after it has already been developed. Adults then observe and learn from children's interactions with the product after it has already been designed and released. A second possibility is to include children as "testers." In this case, children use or test the technology at earlier stages of development, perhaps even as a prototype. Observations of and feedback from the children can still be implemented in the final product. Children can also contribute to the design of products as "informants," as they are more involved as sources of information throughout the process, prior to the design, and also regularly during testing and evaluation. Finally, children can be involved as "design partners," the highest level of

involvement in the design of a product. In this case, they are viewed as equal members of a design team, and their ideas and opinions are integrated from the very beginning of the design process. Depending on the age of the child, as well as individual characteristics, different roles may be appropriate.

Ensuring that products are developmentally appropriate is an important responsibility in designing for children (Bredekamp & Copple, 1997). One definition suggests that for something to be developmentally appropriate, it must be “challenging but attainable for most children of a given age range, flexible enough to respond to inevitable individual variation, and, most importantly, consistent with children’s ways of thinking and learning” (Clements, 2002, p. 161). Further, products should be designed to contribute positively to both the health and development of the children using them (Wartella, O’Keefe, & Scantlin, 2000).

4.1.6 Design and evaluation criteria

A number of criteria can be considered when designing, as well as evaluating digital media products. The first criterion, usability, is defined as the “extent to which a system, product, or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” (International Organization for Standardization (ISO), 2018). The three sub-criteria of usability, effectiveness, efficiency, and satisfaction will be described in more detail below.

Effectiveness is the accuracy and completeness with which a user achieves his or her goals using the product. For example, when interacting with a tablet device, a child must perform a gesture, and the gesture must also be recognized by the device in order for the gesture to be effective (Dubé & McEwen, 2017). To evaluate effectiveness, evaluators could look at the number of correct gestures over the total number of gestures performed.

Efficiency refers to the amount of resources used in relation to the results achieved (ISO, 2018). A common resource evaluated for efficiency is time, i.e., the amount of time it takes an individual to complete his or her goals with the product. Satisfaction is defined as the “extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user’s needs and expectations” (ISO, 2018).

In addition to usability, Markopoulos and colleagues (2008) presented a number of additional design and evaluation criteria for products for children (Markopoulos, Read, MacFarlane, Hoysniemi, 2008), though some are also included in the International Organization for Standardization (ISO, 2018):

Usefulness: Does the product provide benefits to users or help them address their needs and goals?

Learnability: Is the product easy to learn how to use? Can competence with the product be reached in a reasonable amount of time?

Accessibility: Is the product able to be used by the full range of potential users, including their needs, characteristics, and capabilities?

Safety: Is the device or product safe to use? For example, will children be exposed to unwanted advertisements or contact?

Additionally, designers should design for engagement, capturing and keeping attention, and avoiding boredom. Engagement is the idea that something can attract and hold our attention (Chapman, 1997). More specifically, it is characterized by “attributes of challenge, positive affect, endurance, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control” (O’Brien & Toms, 2008, p. 7). Designers can look to entertainment for ideas about how to create engaging software (Dickey, 2005). Indeed, it has been suggested that there is a shift of focus from usability

toward engaging experiences (Hassenzahl & Tractinsky, 2006). One reason for this may be because of the competition with other products or media. However, this presents a difficult question in terms of the distinction between a product that is fun and entertaining versus a product that aims to educate. We return to this discussion below.

Engagement is not only important in terms of a user's experience with a product but is also very important to learning. Researchers agree that learning takes place when learners are engaged and that the goal of educational software should be to provide an engaging learning environment (Sim, MacFarlane, & Horton, 2005; Webster & Ho, 1997). Young children have also been found to be more motivated to learn when they use engaging digital media (Lieberman, Bates, & So, 2009). In terms of evaluating for engagement, evaluators can further distinguish between behavioral, emotional, and cognitive engagement (Fredricks, Blumenfeld, & Paris, 2004).

A final criterion for design and evaluation is fun. It can be difficult to distinguish between engagement and fun. A definition of fun, for example, has many similarities with how engagement has been described above: "Things are fun when they attract, capture, and hold our attention by provoking new or unusual emotions in context that typically arouse none, or by arousing emotions not typically aroused in a given context" (Carroll, 2004, p. 38). Researchers argue that fun contributes to motivation to pursue an activity, which can, in turn, contribute to effective learning (Malone & Lepper, 1987; Prensky, 2001). However, other researchers have shown that though there was a relation between fun and usability, there was no relation between fun and learning (Sim et al., 2005). Though they found no relation between fun and learning, in this particular study, children did prefer to use the product that they rated as more fun (Sim et al., 2005). There seems to be a discrepancy between these findings and those of Liebermann and colleagues (2009) that, on the one hand, children are more motivated to learn as a result of engagement with

the material, and on the other hand, though they prefer to interact with “fun” products, this is not associated with learning. This brings us to the possible distinction between a product or task being engaging versus being fun: perhaps engagement implies that a child is connecting with the task in a way that facilitates learning while fun implies that, though a child may enjoy performing a task, he or she may not necessarily be attending to the task in a way that highlights the knowledge to be learned and the connections to be formed between the task and existing knowledge. This can be an issue when considering products that are meant to be entertaining and products that are meant to be educational, as well as products that attempt to be both.

Consequently, there is a fine and sometimes blurry line between entertainment and educational products. Thus, it is important to highlight qualities for the design and evaluation of products that focus on the educational aspect. Hirsh-Pasek and colleagues (2015) discuss five qualities to look out for when evaluating “educational” products, but these guidelines could also be taken into account earlier in the design process. First, do the activities or enhancements in an application actually add value and increase engagement, or are they distracting? Second, does the application have too many choices, increasing distraction and decreasing engagement? Third, is the application really educational with specific learning goals, or rather just rote memorization? Fourth, is the application entertaining, but lacking educational content? Finally, is there too much going on, too often switching between screens or tasks, rather than focusing on repeating and learning content? (Hirsh-Pasek, Zosh, Golinkoff, Gray, Robb, & Kaufman, 2015).

During the design process, it is important to evaluate the product often with users and then to redesign based on the evaluation and then reevaluate the new design, and to continue this process until a final design is reached that meets all the requirements or has

solved all the problems. This process of design-evaluation-redesign-reevaluation is called iterative design (Gould & Lewis, 1983).

4.1.7 Evaluating products for children with children

Designers of educational products, as well as the parents and teachers who provide educational products to children, should want and be able to evaluate said products in terms of usability and educational value. To evaluate products based on the design factors described above, there are two possible approaches. First, evaluators can observe individuals interacting with the product and look for behaviors suggestive of engagement, fun, usability, etc. A second possibility is to ask individuals for their own assessments of the interaction.

One common method for evaluations is the Think Aloud method. With this approach, individuals must verbalize their thoughts throughout the interaction with the product. Think Aloud has been considered particularly valuable and is often the default practice for usability testing (Dumas & Redish, 1999; Nielsen, 1994). Some researchers have suggested Think Aloud is not effective until children are in their teens (Hanna, Ridsen, & Alexander, 1997), while others have shown that Think Aloud can be used with children as young as seven years old (Donker & Markopoulos, 2002).

However, because preschool children can have a short attention span, be prone to social desirability bias, or have difficulty verbalizing their thoughts, the Think Aloud method may not be the best approach for evaluation (Hanna et al., 1997; Oerke & Bogner, 2013; Read & Fine, 2005). For example, preschool children's language skills are not fully developed, so they may have trouble keeping up a constant flow of verbalization. The extra cognitive load placed on them to verbalize may cause them to remain quiet during the most difficult moments, which is when verbalization would be the most useful to an evaluator (Fransen & Markopoulos, 2010; Markopoulos et al., 2008).

The Think Aloud process may present additional difficulties for children because of the unusual social dynamics of the situation. Children can be shy when interacting with unfamiliar adults, and the Think Aloud process essentially requires them to talk constantly with no (or very limited) feedback or response from the adult. Further, children may not understand that it is the product that is being “tested” and not them. They may feel bad or embarrassed when they make mistakes (Fransen & Markopoulos, 2010). An easy adaptation to the Think Aloud was proposed by Donker and Reitsma (2004). They called this method the voluntary Think Aloud and looked only at any spontaneous utterances children made during evaluation sessions.

There are a couple of alternatives that have been developed as a response to these issues with the Think Aloud method. Active Intervention has the experimenter engage more actively with the child, asking more questions, prompting, helping children to keep talking (Monk, Davenport, Haber, & Wright, 1993). This creates a more natural setting. However, there is the danger of leading the child in a particular direction, focusing on issues that the evaluator thinks are a problem, rather than discovering new problems through the child’s interaction.

A second alternative is to use a social robot as a stand-in for the evaluator. In Robotic Intervention, the evaluator typically controls the social robot from a different room, allowing the child to interact freely with the robot. This could create an environment in which the child is more comfortable, has fun talking to the robot, and is less inhibited in his verbalizations. Researchers have found that, in terms of evaluation, there is no difference in the quality of verbalizations children make with an adult vs. a robot, but that children considered the robot more fun (Fransen & Markopoulos, 2010). However, the workload for the researcher controlling the robot is very high.

In addition to the verbalization methods described above, survey methods can also be used to capture children's opinions about products in usability testing. The Fun Toolkit is an assortment of different methods for evaluating children's assessment of how fun a product was to use (Read, MacFarlane, & Casey, 2002). The first is the Funometer, which has children "fill up" a thermometer to the level of fun for a specific product (Risden, Hanna, & Kanerva, 1997). The amount of fun, the line children draw, is measured with a ruler.

The Smileyometer is a variation of the Funometer with five discrete fun ratings, including awful, not very good, good, really good, and brilliant. An issue with the Smileyometer is that children mostly provide positive responses and very rarely select either of the two negative smileys. For example, across two trials, 84% of children aged six to nine picked "Brilliant" (Read et al., 2002). A solution to this problem was developed by Hall, Hume, and Tazzyman (2016), which presents children with five different levels of happy smileys. In this case, children select across all five levels instead of just three, as in the original version.

The Fun Sorter can be used when children are evaluating two or more different products. It requires them to rank products on different criteria, for example, "worked the best," "most fun," and "easiest to do." Finally, the Again-Again Table asks children if they would like to use the product again with response options, yes, maybe, and no. The Again-Again Tables can be used with one or more products. The similarity of results of the Funometer and the Smileyometer led Read and colleagues (2002) to conclude that either the Funometer or the Smileyometer should be included in a toolkit, but that both are not necessary. The results of the Fun Sorter and the Again-Again tables were also similar, so the authors suggested again to only include one of the measures in a toolkit. Using two different methods to evaluate fun with children provides a broader understanding and

accounts for either positive bias or preference with methods such as the Funometer or Smileyometer, and at the same time, accounts for the greater difficulty of using the Fun Sorter or the Again-Again table for children. Read and colleagues (2002) concluded that either the Funometer or the Smileyometer should be included, and either the Fun Sorter or the Again-Again Tables should be included in evaluations of fun for children. However, some of the methods, such as the Funometer, were more successful with older children. It is likely that preschool children, younger than the children who tested out the Fun Toolkit, could struggle with understanding how to use the measures. Perhaps the Smileyometer would be the exception, though they still might not understand that the Smileyometer should be used to evaluate the product and may simply choose the smileys they like best.

Finally, an effective alternative to the methods described above is to simply observe children as they interact with a product. With structured observation, evaluators can determine a focus and develop observation guides to support their observations and help them know what to look for. Some researchers have found that most problems were identified by observation and that the Think Aloud provided supplementary information about the importance of the problems to children (Donker & Reitsma, 2004). For engagement, evaluators can look for positive and negative instantiations of engagement by observing children's interactions (Read et al., 2002). Researchers can observe facial expressions or comments, as well as usability issues that occur, and these behavioral signs can serve as a more reliable measure than children's direct responses (Hanna et al., 1997).

4.1.8 Summary

In summary, this section has outlined, first, that CVS abilities can be promoted in children in early elementary school and raises the question if this is also the case for young children before they enter into formal education. Second, we presented the literature on children's use of digital technology, as well as the importance of designing both for

children and for learning outcomes. Third, we outlined the importance of evaluating products for children and with children to ensure that children have an optimal experience that allows for both accurate assessment of abilities and effective promotion of those abilities. It is critical that usability of products, whether as research tools or educational tools, whether digital or paper-and-pencil, is optimized to ensure that the data obtained from them is both meaningful and accurate. In the case of the current studies, we wanted to create engaging research tools but at the same time, avoid that children become too overwhelmed or distracted that they do not take the tasks seriously.

4.2 The Present Studies

In the next sections, we will present three studies regarding the design and evaluation of digital media for assessing and teaching the Control of Variables Strategy with preschool children. With these studies, we investigated and determined important factors for designing and evaluating CVS training tools for preschoolers. In addition, we investigated if the developed tools could be used both to assess and promote preschool children's CVS abilities. Specifically, Studies 5a and 5b present the iterative design and evaluation process of a tablet application for assessing and training CVS abilities. Study 5a compared a paper-based tool and a tablet application. Study 5b built on the findings of Study 5a and continued the development of the tablet application. In addition, the evaluation was changed to place the focus on whether or not children could interact with the product in a way that could allow its use for assessments. Study 6 presented a video tutorial for teaching CVS and investigated how animation affected children's experience with the tutorial, as well as whether or not the tutorial affected children's performance on an unrelated CVS assessment.

4.3 Study 5a: The Development of a Tablet Application for Training CVS

4.3.1 Statement of Collaboration

In this section we describe the study design, the method, and the results of a study investigating both the potential of a digital task for use in scientific research and the effect of training through this task on children's abilities in Control of Variables Strategy. In terms of this study, I collaborated with Daniela Becker, a master student in Media Informatics at LMU. I developed the concept of the task for investigating CVS and we co-designed the material of the study i.e., the story and the script. The illustrations were created by an external collaborator, Alexander Schenker. Ms. Becker was responsible for the implementation of the task both as a tablet application and as a paper-based task. She collected the data under my supervision at a number of kindergartens in Munich. The work was relevant to her in terms of her master thesis on the design and evaluation of an educational application for preschoolers (Becker, 2018). Thereby, she conducted a preliminary evaluation of the user experience with the tasks based on observation (video recordings of testing sessions). Additionally, she carried out a preliminary analysis of the effect of the task training on children's CVS abilities. Ms. Becker's notes and coding-scheme for the evaluation of the tasks formed the basis of the final coding-scheme. In terms of this thesis, I report some results of Ms. Becker's analyses, which have been analyzed again by me. In addition, for the evaluation of the tasks, I have analyzed children's need for help, their engagement with the tasks, the frequency of issues over the course of the experiment, children's interactions with the application, and their abilities in a warm-up game which required them to perform a drag-and-drop action. For the evaluation of children's CVS abilities, I took a more fine-grained approach of evaluating whether children vary focal variables and control variables, I have investigated the effect

of age as a continuous variable, and I have described children's verbal explanations of their experiment design.

4.3.2 Introduction

The first goal of the present study was to design a paper-based storybook task and a tablet application to be used in scientific research on children's abilities in Control of Variables Strategy. To this end we iteratively designed and evaluated the tasks. In addition, we wanted to compare any advantages or disadvantages of the medium of the task for use with children. Further, we wanted to use the tasks as a training for CVS abilities and did so with a pre-test, training, post-test design. In designing the tasks, we considered principles of cognitive architecture, instructional design, and multimedia learning.

4.3.3 Method

4.3.3.1 Participants

A total of 23 preschool children participated in the study ($M_{\text{age}} = 67.38$ months, $SD = 7.38$; range: 51.53 - 79.87 months, 11 girls). All participants were typically developing children of lower- to upper-middle class background from a large German city. Parental informed consent and child assent was obtained for all children before the study. Three children participated in testing the first iteration of the application ($M_{\text{age}} = 72.38$ months, $SD = 3.15$; range: 69.20 - 75.50 months, 1 girl), two children participated in testing the first iteration of the paper storybook ($M_{\text{age}} = 74.38$ months, $SD = 0.35$; range: 73.77 - 75.00 months, 2 boys), nine children participated in testing the second iteration of the application ($M_{\text{age}} = 67.31$ months, $SD = 6.07$; range: 59.00 - 79.87 months, 3 girls), and nine children participated in testing the second iteration of the paper storybook ($M_{\text{age}} = 64.23$ months, $SD = 8.82$; range: 51.53 - 78.47 months, 7 girls).

4.3.3.2 Materials

A matching puzzle game was used to familiarize the children with the testing environment (Figure 4.1A). An iPad application, Fiete Puzzle (Ahoiii Entertainment, 2017), was used as an additional warm-up for the children who would interact with the tablet (9.7in iPad 2). This application required children to drag-and-drop puzzle pieces to their appropriate location. Children could choose which theme they wanted to play, either farm animals or fire department (Figure 4.1B).



Figure 4.1. A) Warm-up puzzle. B) Warm-up puzzle on tablet.

To investigate children's Control of Variables Strategy abilities, the following story was used as a context for designing experiments. The experimenter/narrator introduces Farmer Meyer (FM) and his many animals, especially his chickens. FM can decide what to feed his chickens (Variable 1): herbs or grain; he can decide their sleeping location (Variable 2): inside or outside; and he can decide the type of nest (Variable 3): straw nest or stick perch. One day, FM notices that some of his chickens are laying spotted eggs and others are laying plain eggs. He wonders why this happened and supposes it could be due to one of the previously mentioned variables, the type of food, the sleeping location, or the type of nest. He wants to find out but needs some help designing a test. The script of the story can be found in Appendix I.

An interactive storybook and an iPad application were developed using the story described above to investigate children's CVS abilities (Figure 4.2 A&B). The storybook was created with laminated DIN A4 illustrations and magnetic strips glued to the back. Small icons of the variables (3.8 x 3.8 cm) were also laminated and glued first to a foam sheet backing and then to a magnet strip. In this way, the variable icons could be placed on the storybook and also moved or removed. In addition to the variable icons, two blue arrow icons were created and used to mark the current variable under investigation. There were three "chapters" of the story/task which was indicated by three different background colors.

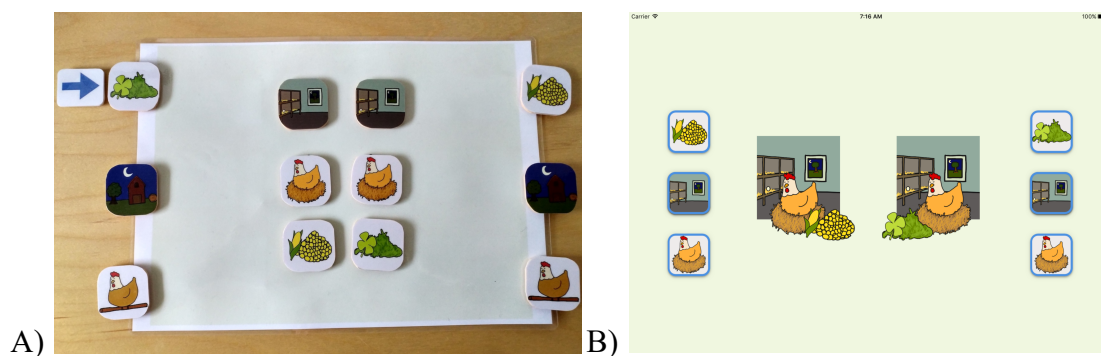


Figure 4.2. A) Paper-based storybook task. B) Tablet application.

In the iPad application, the user proceeds through the story by swiping or tapping on the right side of the screen. The variables are first introduced and then highlighted to indicate that they are now active and can be selected by tapping on them. Then, only the selected variable(s) remain highlighted. For a detailed description of the design and development of the tasks, please refer to Becker (2018).

4.3.3.3 Procedure

Data for this study were collected between February - March 2018. All sessions took place in local kindergartens in a separate, quiet room and were video recorded.

Children were tested individually in sessions lasting approximately 20 minutes. The experimenters were introduced to the children before the individual testings. When children entered the testing room, they were asked to sit at a low table and the experimenter sat across from them.

The testing session began with a warm-up game in which children were asked to complete a simple matching puzzle. The purpose of the warm-up was to make the child comfortable with the experimenter and the testing environment prior to the main tasks. The children who participated in testing the iPad application additionally completed a warm-up puzzle on the tablet. This required them to drag and drop puzzle pieces to complete a puzzle.

The session continued with the Farmer Meyer task. The procedure for the storybook interaction will be described below. The procedure for the tablet application was the same, except that children tapped on the icon buttons to make selections. The selected icons were highlighted and also appeared on or around the chickens in the middle. The tasks consisted of six phases: the story phase, the training phase, the question phase, the pre-test, the instruction and feedback phase, and the post-test (Figure 4.3).



Figure 4.3. Procedure of the six phases of the task.

The experimenter told the children the story of Farmer Meyer as described in the materials (story phase). When the three variables were first introduced, they were placed one by one next to a chicken and then the children were instructed to select one of the

variable levels to apply to the chicken. The child then moved the selected magnet icon and placed it on or near the chicken (training phase; Figure 4.4).



Figure 4.4. Training phase for interaction with task.

Next, the experimenter introduced the critical occurrence of Farmer Meyer's chickens producing two different types of eggs, spotted and plain. FM states that this difference could be due to the type of food, the sleeping location, or the type of nest. FM decides that he thinks that it is the type of food that makes a difference if his chickens lay spotted or plain eggs. Children are asked how they think FM could figure out if that is the case (question phase).

Children are then shown two chickens and instructed to select a food, a sleeping location, and a nest for each chicken in order to find out if the type of food makes a difference in whether the chickens lay spotted or plain eggs. The two chickens are located in the middle of the page and the magnet icons are located next to the chickens, one set for each chicken. The blue arrows or highlighting indicate the current variable of interest, in this case food (Figure 4.5).

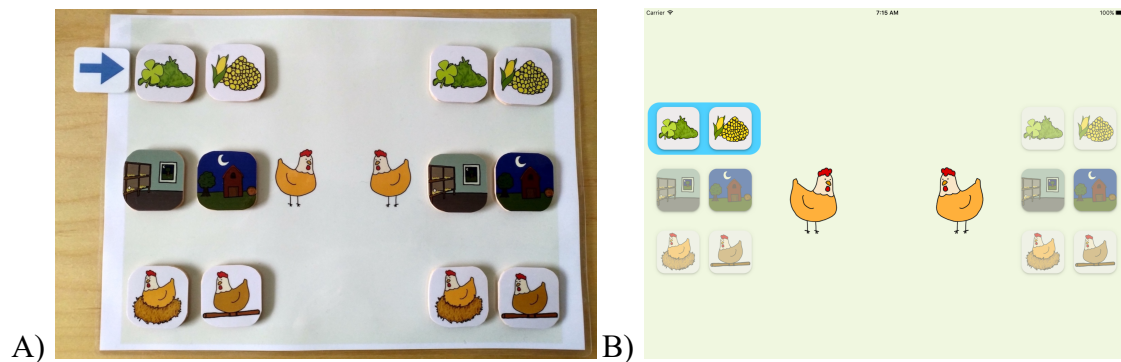


Figure 4.5. Initial state of the CVS task in the A) storybook version and B) tablet application.

Children are instructed to first select a type of food for Chicken 1 and then for Chicken 2, then a sleeping place for Chicken 1 and then Chicken 2, and finally a type of nest for Chicken 1 and then Chicken 2. They did this by moving one of the magnet icons from the side of the page and placing it on or near the associated chicken. Once children made all their selections, they were asked if they were finished or if they wanted to change any of their selections. If they wanted to change something, they were allowed to do so and then the same question was repeated. The remaining magnet icons were removed from the page. Then, children were asked if they thought this set-up was a good test to find out if the type of food makes a difference in whether the chickens lay spotted or plain eggs and why or why not. In this phase, children were not given any feedback or support for their selections or answers (pre-test).

Next, the task was repeated with a different variable of interest: the sleeping location. In this phase, children were given feedback after they had assigned each variable to the chickens (instruction/feedback phase). Once children had selected a sleeping place for each chicken, they were informed if their selection was correct or not and why. For example, if children assigned the chicken two different sleeping places, the experimenter told them: “Very good! We want to find out if the sleeping place makes a difference, so it’s correct that you picked two different sleeping places so that we can compare them.” If

they selected the same sleeping place for both chickens the experimenter told them: “You picked the same sleeping place for both chickens. We want to find out if the sleeping place makes a difference, so we actually need to pick two different sleeping places so that we can compare them. Let’s change one of the sleeping places so that the chickens sleep in different places.”

This procedure continued for the two remaining variables. Importantly, the remaining variables needed to be kept constant (controlled), in order to design a good test. If children assigned chickens the same food or nest, the experimenter told them: “Very good! We want to find out if the sleeping place makes a difference, so it’s correct that you gave both of the chickens the same food. This way, we can be sure that we only compare the sleeping place so we can find out if it makes a difference. And we are sure that our comparison is not influenced by something else.” If children assigned chickens different foods or nests, the experimenter told them: “You picked different foods for both chickens. We want to find out if the sleeping place makes a difference, so we actually need to give the chickens the same food so that the food doesn’t influence our comparison. Let’s change one of the foods so that the chickens eat the same food and we only compare the sleeping place.” Next, children were asked if they thought this set-up was a good test to find out if the sleeping place makes a difference in whether the chickens lay spotted or plain eggs and why or why not.

Finally, children repeated the task a third time, with the type of nest as the variable of interest (post-test). They were not given any feedback on their experiment design, as in the pre-test phase. Following the study, children were given a research certificate thanking them for their participation and informing them of the importance and role of “junior researchers.”

4.3.3.4 Coding

Usability issues. The video recordings of testings were watched to identify usability issues occurring during children’s interactions with both the tablet application and the interactive storybook. Some issues were unique to the task type and some issues overlapped. Interactions were considered issues if actions did not generate expected effects.

Selection from incorrect side (U1). The tasks were designed to proceed in a specified order, alternating left to right before moving down the page (Figure 4.6). This issue occurred when children attempted to select a variable from the incorrect side.

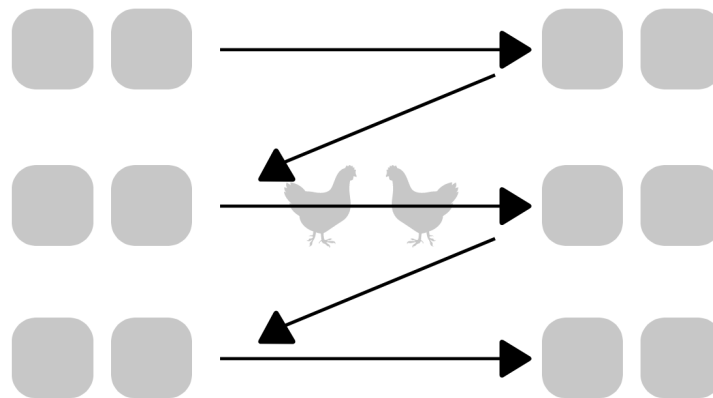


Figure 4.6. Predetermined order of the interaction for variable selection.

Element not yet selectable (U2). This issue occurred when children attempted to tap a button that was not yet able to be selected or to move a magnet before they were supposed to. In the tablet application, buttons were always highlighted with a blue background when they could be selected. Unhighlighted buttons could not be selected.

Selecting two things at once (U3). This issue occurred when children attempted to select two variables at the same time or move two magnets at once.

Reordering the magnets (U4). This issue occurred when children arranged the magnets in a different order than the initial set-up suggested.

Accidentally clicking (U5). This issue occurred when children accidentally clicked something, e.g., reaching to click a button at the top of the screen, but activating a button on the lower part of the screen with their palm/wrist.

Area not selectable (U6). This issue occurred when children clicked on an area that had no effect.

Incorrect action (U7). This issue occurred when children attempted to drag-and-drop an element rather than tap it.

Behavior issues. The video recordings of testings were watched to identify behavior issues occurring during children's interaction with the tasks.

Impatience (B1). This issue occurred when children wanted to proceed with the task before the experimenter was finished explaining something, or before the experimenter had activated the buttons to be selectable.

Hesitance/Confusion (B2). This issue occurred when children were unsure about how to proceed, where to tap, or how to change something. For example, if children hesitated to place a magnet because they were not sure where it should go or hesitated to click something because they were not sure what to click.

Help (B3). This issue occurred if a child verbally asked for help or how to do something, or if the experimenter had to intervene for the task to proceed.

Boredom (B4). This issue occurred if a child looked away from the task.

Interaction. The number of attempts a child made to tap or drag-and-drop in the application were recorded and compared to the number of actions those attempts were meant to initiate.

CVS performance. Children's CVS performance on the first and third task was coded by variable. They were given a score, correct or incorrect, for each variable assignment, which could total a maximum of three. For example, if the focal variable was food, children would have to vary this variable for one point and keep the other variables constant (sleeping location and nest type) for one point each. This coding scheme was developed based on coding schemes by van der Graaf and colleagues (2015) and Edelsbrunner (2017). Van der Graaf and colleagues assigned points for each correctly varied or correctly controlled variable and Edelsbrunner distinguished between correctly varying the focal variable in conclusive hypothesis testing and correctly controlling variables using CVS. Because we are investigating CVS abilities in an age group much younger than previously investigated and in which these abilities seem to be developing, it is important to look in more detail at how those abilities are developing. For example, children may first understand contrastive testing, then understand the need to control variables, but perhaps fail to control all variables at once, before finally understanding CVS as a need to control all variables other than the focal variable. Thus, this coding scheme allows us to see the development of the understanding of the need to control variables even when children have not yet mastered a full understanding.

Children were also asked if they thought their experiment design was a good test to find out if the focal variable made a difference for the type of eggs laid and why or why not. Their verbal responses were coded for references to theories (1), experimental design (2), or goals of the experiment (3): providing a theory for why one of the variables matters; referring to the design of the experiment, such as two variables being the same or being different; or referring to the goal of finding out why the eggs are different, or if the focal variable makes a difference.

4.3.4 Results

4.3.4.1 First Iteration

Five children participated in evaluating the first iterations of the application ($n = 3$; $M_{\text{age}} = 72.38$ months ($SD = 3.15$); range: 69.20 - 75.50 months, 1 girl) and the storybook ($n = 2$; $M_{\text{age}} = 74.38$ months ($SD = 0.35$); range: 73.77 - 75.00 months, 2 boys).

Usability. In the first iterations, five usability issues and four behavioral issues occurred (refer to Table 4.1). The most common usability issue, and one that affected both the application and the storybook, *selection from incorrect side*, occurred when children failed to recognize from which side (left or right) they should select a variable. The desired order was alternating from left to right as moving down the page. Children either attempted to select a variable for the second chicken from the same set as for the first chicken instead of switching sides, or they proceeded down the page to the next variable instead of selecting the same type of variable for the second chicken. This issue occurred 14 times, more often with the application ($M = 4.00$) than with the storybook ($M = 1.00$). The fact that issue occurred so often was surprising considering that in the application the currently active buttons were highlighted in blue and in the storybook the current variable was indicated with a blue arrow. To address this issue, we increased the size of the highlight in the application, so that the blue border would be more salient. For both tasks, the script was adapted to emphasize that there are two chickens, one on the left and one on the right, and that they each need a variable assigned to them from the options on each side.

The second usability issue, *not yet selectable*, occurred when children attempted to click on something before the button was activated or attempted to move a magnet before the experimenter had indicated they should. This issue occurred five times, more often with the application ($M = 1.33$) than with the storybook ($M = 0.50$). To address this issue

in the application and the storybook, an additional screen or page was added in between the trials. The experimenter could stop on this screen to fully explain the upcoming task before moving to the screen where variables were selectable.

Table 4.1

Average frequency of usability issues in the first iteration

Code	Description	Average Frequency (app)	Average Frequency (book)
U1	Selection from incorrect side	4.00	1.00
U2	Not yet selectable	1.33	0.50
U3	Selecting two variables at once	-	2.00
U4	Reordering magnets	-	1.50
U5	Accidentally clicking	0.33	-
B1	Impatience	1.33	0.50
B2	Confusion	0.67	1.00
B3	Help	3.67	1.50
B4	Boredom	0.33	0.50

The third usability issue, *selecting two variables at once*, was unique to the storybook and unique to one child. This issue occurred when the child selected two variables at once, one with each hand. The issue occurred four times. Because this issue was unique to one child, no adaptation was made to address this issue.

The fourth usability issue, *magnet reordering*, was also unique to the storybook and occurred when children attempted to reorder the magnets in an order different than what was indicated on the page. This issue occurred three times, ($M = 1.50$ times per

child). We did not consider this a serious issue and were curious if this issue would continue to occur, so no adaptation was made to address this issue.

The fifth usability issue, *accidentally clicking*, was unique to the application and occurred when a child clicked something they did not mean to click. This issue occurred only once.

Behavior. The first behavioral issue, *impatience*, occurred when a child attempted to proceed with the task before the experimenter was finished explaining something. This issue occurred five times, more often with the application ($M = 1.33$) than with the storybook ($M = 0.50$). It is possible that this is due to the application being more “fun,” and thus, children wanted to get to the next page more quickly so that they could continue interacting with the application. It is also possible that the design of the application motivated children to click on the buttons, while they were more hesitant to take action to move the magnets in the storybook setting. The adaptation for the usability issue, *not yet selectable*, the addition of transition screen or page between trials, can also address this issue. The experimenter could stop on this screen to fully explain the upcoming task before moving to the screen where variables were selectable.

The second behavioral issue, *confusion*, occurred when a child was unsure how to proceed or hesitated to complete an action. This issue occurred four times, less often with the application ($M = 0.67$) than with the storybook ($M = 1.00$). The adaptations made to the script, to emphasize that there are two chickens, one on the left and one on the right, and that they each need a variable assigned to them from the options on each side, should also address this issue.

The third behavioral issue, *help*, occurred when a child verbally asked for help or if the experimenter had to intervene for the task to proceed. This issue occurred 14 times, more often with the application ($M = 3.67$) than with the storybook ($M = 1.50$). No

changes were made to address this issue specifically, but the adaptations to improve clarity, emphasize side switching, and decrease impulsive actions, should also decrease the need for help.

The fourth behavioral issue, *boredom*, occurred when a child looked away from the activity or started talking about something else. This issue occurred twice, less often with the application ($M = 0.33$) than with the storybook ($M = 0.50$). No adaptations were made to address this issue.

4.3.4.2 *Second Iteration*

Eighteen children participated in evaluating the second iteration of the application ($n = 9$; mean age = 67.31 months ($SD = 6.07$); range: 59.00 - 79.87 months, 3 girls) and the storybook ($n = 9$; mean age = 64.23 months ($SD = 8.82$); range: 51.53 - 78.47 months, 7 girls). First, we will report the performance on the warm-up and the interaction with the application, before going into the issues in more detail as above.

For children who interacted with the application, they were first asked if they had previous experience with a tablet. Five of nine children had previously interacted with a tablet. Next, children completed a warm-up puzzle on the iPad. This required them to drag-and-drop puzzle pieces to their corresponding location. In general, children performed well on this task requiring only a few extra attempts to complete an action: two children required no extra attempts, three children required one, two children required two, and one child required six extra attempts to drag-and-drop the puzzle pieces ($M = 1.63$; $SD = 1.92$; range = 0-6). This finding suggests that the drag-and-drop motion is accessible to young children and could be used in future implementations of experimental tasks.

On average, in the training and three CVS tasks, children required 0.47 extra attempts to complete an action ($SD = .77$; range = 0-2). This could be because they clicked

the wrong button, clicked near a button but not on it, or did not use enough force for the button to register the tap. Buttons should be designed to be big enough so that children can easily click on them and children should also receive a warm-up to practice tapping on buttons with enough force to register the action.

The average duration of the interaction with the application was 7 minutes and 13 seconds ($SD = 0:29$; range = 6:28-7:55). The average duration of the interaction with the storybook was 9 minutes and 14 seconds ($SD = 1:40$; range = 7:37-13:07). For experimental research purposes, and also for in-school educational purposes, a shorter duration interaction with the task would save time and effort on behalf of the experimenter or teacher. That being said though, a child spending more time on a task is not always bad. This could be indicative of deeper thinking or longer verbal responses. Looking at specifically the story portions of the task, the introduction story lasted 19 seconds on average ($SD = 0:03$; range: 0:16-0:29) and the story about the different eggs lasted 10 seconds on average ($SD = 0:01$; range: 0:09-0:13).

In the second iterations, five usability issues and four behavioral issues occurred (refer to Table 4.2). The average frequency of usability issues was 2.56 ($SD = 1.88$) for the application and 4.78 ($SD = 2.17$) for the storybook. The average frequency of behavior issues was 4.44 ($SD = 2.60$) for the application and 5.44 ($SD = 2.51$) for the storybook.

Usability. *Selection from incorrect side* remained the most common usability issue; the average frequency of this issue decreased for the application ($M = 2.33$) but increased for the storybook ($M = 2.44$). This would suggest that increasing the size of the highlight in the application made the currently active variables more salient and decreased the confusion related to switching sides. However, adapting the script to emphasize that there were two chickens and that they each needed a variable assigned to them from the options on each side was not very successful in directing children's attention to the correct, active

variables. Because *selection from incorrect side* still remained an issue, perhaps the design of this interaction was not intuitive, and children spontaneously tried to interact with the tasks differently than we wanted them to according to the design. For example, it may have been that having the variables on both sides, one set for each chicken, was unnecessary and could be limited instead to just one set. It could also be that having the variables located on the outside edges of the screen and the chickens located in the middle of the screen did not place enough emphasis on the variables themselves.

Table 4.2

Average frequency of usability issues in the second iteration. Green indicates the frequency of the issue decreased; red indicates an increase; and yellow indicates no change between the two iterations.

Code	Description	Application	Storybook
U1	Selection from incorrect side	2.33	2.44
U2	Not yet selectable	0.33	0.22
U3	Selecting two variables at once	-	-
U4	Reordering magnets	-	2.11
U5	Accidentally clicking	-	-
U6	Area not selectable	0.44	-
U7	Incorrect action (Drag & drop)	0.44	-
B1	Impatience	0.44	0.56
B2	Confusion	1.00	2.44
B3	Help	3.00	2.67
B4	Boredom	0.33	0.22

Another possibility is that children could not see the variables on the right side of the screen since they used their right hand to make selections and frequently covered the right side of the screen with their forearm. Future iterations should investigate different interface designs to determine a more intuitive interaction. For example, the variables on the left side of the screen could disappear or grey out after one of them has been selected. This should make children notice that they can no longer select from those variables, and that there must be another set of variables, causing them to look to the right side of the screen where the second chicken and second set of variables are located.

The second usability issue with children clicking buttons or moving magnets that were *not yet selectable* decreased in both the application ($M = 0.33$), and the storybook ($M = 0.22$). The addition of an additional screen or page between the trials to allow the experimenter to fully explain the upcoming task before moving to the screen where variables were selectable appears to have helped children to not click or move the variables before they were supposed to.

The third usability issue, *selecting two variables at once*, did not occur in the second iteration. This issue was unique to one child in the first iteration and no children who participated in testing the second iteration did this.

The fourth usability issue, *magnet reordering*, occurred more frequently in the second iteration ($M = 2.11$) and eight out of nine children reordered the magnets at least once. No adaptations were made to address this issue after the first iteration. This issue is a potential problem, because when children organize the variables differently from trial to trial, it may be difficult for them to remember what they have done in previous trials or may cause them to focus less on the current trial and the current focal variable. This issue obviously does not occur in interactions with the application, because the variables appear in their corresponding location after being selected. Allowing the children too much

freedom in where to place magnets resulted in a frequently occurring issue that could be detrimental to their CVS abilities. Future iterations should clearly indicate where the magnets should be placed (for example, like puzzle pieces) and the interface should be designed so that the variables are located more intuitively. For example, children often placed the food at the bottom of the screen in front of the chicken, placed the nest type on top of the chicken, and placed the sleeping location above the chicken.

The fifth usability issue, *accidentally clicking*, did not occur in the second iteration, suggesting that this is not a common or systematic issue. Two new issues arose in the second iteration of the application: children clicking on an area of the screen that was *not selectable* and caused no effect ($M = 0.44$), and children performing an *incorrect action*, i.e., drag-and-drop instead of tap, ($M = 0.44$). The former issue occurred on static screens where the variables were introduced. The variables may have looked like buttons to children, who then decided to try to click on them. This could be addressed by making the variables look less like buttons, but this issue did not occur very frequently and probably does not need to be addressed specifically in future iterations. The latter issue of performing a drag-and-drop action instead of a tap also did not occur very frequently, however its occurrence may suggest that dragging and dropping the variables to the chickens may be a more intuitive interaction.

Behavior. For the behavioral issue *impatience*, the average frequency of this issue decreased for the application ($M = 0.44$) but increased slightly for the storybook ($M = 0.56$). The addition of an explanation page before the main tasks appears to have limited children's spontaneous actions before the appropriate time, at least in the application. It seems not to have been successful in the case of the storybook, perhaps because children grab a physical object, rather than tap a button, which may be more enticing and thus harder to inhibit. The behavioral issue *confusion* increased slightly in the

application ($M = 1.00$) and considerably in the storybook ($M = 2.44$). It is unfortunate that children's confusion increased in both interactions with the application and the storybook. The adaptations to the script seem to not improve clarity of the interaction enough, which would suggest that the interaction itself is still not intuitive, despite more detailed instructions.

The behavioral issue *help* decreased in the application ($M = 3.00$) and increased in the storybook ($M = 2.67$). The need for help was still very high for both the application and the storybook, at almost 3 times per child on average. This issue is clearly linked to children's confusion and shows that they were so confused that they were unable to proceed or required the experimenter to intervene in order to be able to continue interacting. Finally, the behavioral issue *boredom* did not change for the application ($M = 0.33$) and decreased in the storybook ($M = 0.22$). A low measure of boredom suggests that, even though children struggled to understand how to interact with the tasks, they were at least engaged in the task and did not allow their confusion or frustration to distract them from attempting to interact with the tasks.

Time course of issues. We were further interested if the frequency of issues decreased over the duration of the experiment, as children became more familiar with the interfaces. Figure 4.7 illustrates the average frequency of usability issues in the first and third trials of the testing and Figure 4.8 illustrates the average frequency of behavioral issues in the first and third trials of the testing. We chose to look specifically at the first and third trials because the second trial involved instruction and feedback from the experimenter. In trials 1 & 3, the children navigated through the task themselves, with limited interference.

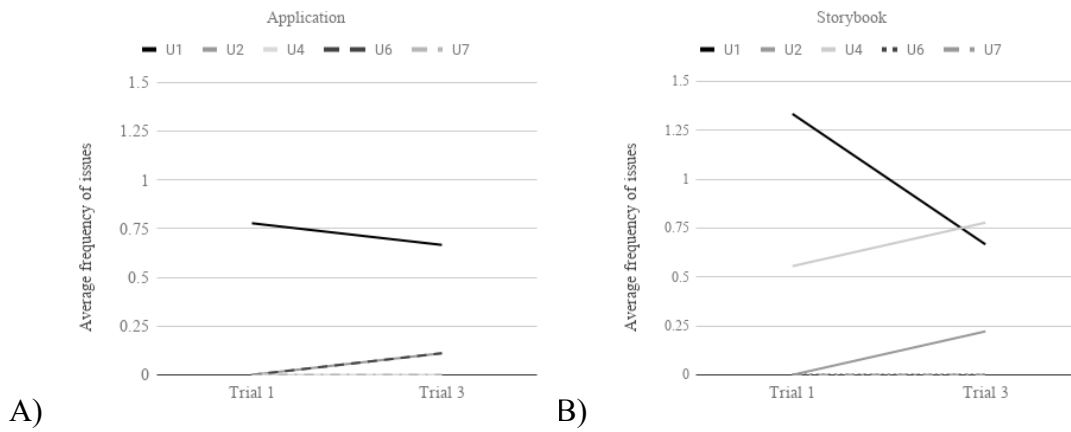


Figure 4.7. Average frequency of usability issues in interaction with A) application and B) storybook from first trial to third trial.

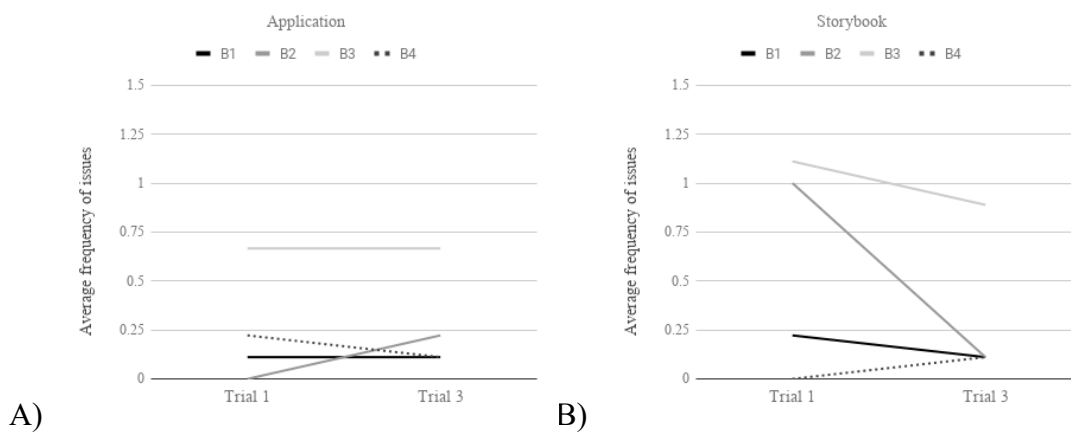


Figure 4.8. Average frequency of behavioral issues in interaction with application and storybook from first trial to third trial.

The usability issue, *selection from incorrect side*, decreased for both the application and the storybook between the first and third trial, so that it was occurring less than once per participant in the third trial. This indicates that with some experience with the tasks, children could better learn the correct interaction procedure. For the storybook, two usability issues increased from the first to the third trial, *element not yet selectable* and *reordering of the magnets*. The issue with a magnet not being selectable yet did not occur often, however the issue with reordering the magnets occurred frequently, on average almost once per participant in the third trial. As mentioned above, ordering the magnets differently in each trial could make it more difficult for children to keep track of their experimental design in previous trials. The fact that this issue increases over the course of the testing indicates that it becomes less clear where the magnets should be placed over a number of trials or that children decide over the course of the testing that certain variables should be placed in specific locations.

Behavioral issues with the application remained fairly stable from the first to the second trial, with need for *help* remaining the most frequent behavioral issue ($M = 0.67$). The storybook saw a large decrease in the issue of *confusion*, as children learned the interaction procedure. However, though there was also a decrease in children's need for *help*, it still remained high in the third trial ($M = 0.89$). The fact that children still required help in the third trial in both the interaction with the application and the storybook suggests that the design was not fully intuitive even after some experience with the tasks.

CVS Performance. In the first trial of the CVS task, children were asked to help Farmer Meyer find out if the type of food makes a difference in what type of eggs the chickens lay. To correctly design a test to answer this question, children should vary the focal variable, food, and keep the other two variables, sleeping location and nest type, constant. In other words, children should give one chicken corn and the other chicken

herbs, and both chickens should have the same sleeping location, inside or outside, and the same type of nest, a straw nest or a stick perch. In the second trial, children were asked to repeat the test but to vary a different focal variable, the sleeping location. They were given feedback on each of their variable assignments and instruction about performing a controlled test. In the third trial, they were asked to repeat the test again, this time varying the nest type as the focal variable. Children's CVS performance results on the first and third trials are reported below separately for the application and the storybook. Table 4.3 shows correlations between age, gender, task materials, tablet experience, and a CVS difference score and Explanation difference score generated by subtracting performance on the first trial from performance on the third trial. There is a significant correlation between task material and CVS performance, such that children who interacted with the storybook materials performed better on the third trial compared to the first trial.

Application. In the first trial, eight out of nine children (89%) correctly varied the focal variable. Two children (22%) controlled one of the control variables and two children (22%) controlled both control variables. One child designed a completely controlled experiment. One child set all variables to be the same and five children set all variables to be different.

These findings suggest that most children recognize the need to vary the focal variable in a test of a hypothesis. About a third of children also understood the need to control (at least one of) the remaining variables. However, only one child successfully did both of these actions to design a controlled experiment. Further, more than half of children varied all three variables, the equivalent of Tschirgi's (1980) "Change All" method.

Table 4.3

Correlations between age, gender, task materials, tablet experience, CVS difference score, and Explanation difference score

	1	2	3	4	5	6
1. Age	-					
2. Gender	-.03	-				
3. Task Material	-.21	.45	-			
4. Tablet Experience	.36	.06	-	-		
5. CVS Difference Score	.10	.00	-.49*	.27	-	
6. Explanation Diff. Score	.16	.00	.00	.73	.00	-

Note. * $p < .05$

Most children (67%) did not provide a relevant explanation for their experimental design. One child made a reference to a theory, “Because they sleep more often;” one child made a reference to the experimental design, “Because the two foods are different;” and one child made a reference to the goal of the experiment, “Because then he (FM) can know if it’s with spots or without.”

In the third trial, five out of nine children (56%) correctly varied the focal variable. Four children consistently varied the focal variable in both trials and four children who had correctly varied the focal variable in the first trial no longer did so in the third trial. Five children controlled more variables in the third trial than they had in the first trial. Three children controlled fewer.

Also in the third trial, no child designed a completely controlled experiment. Three children set all variables to be the same and three children set all variables to be different. Half of children showed persistence in their experiment design from the second trial to the third trial. This means that, after children received feedback and instruction on designing a

controlled experiment in the second trial, half of the children continued to vary the focal variable from the second trial, even though the focal variable was different in the third trial. Looking at children's CVS scores for each variable (0-3 points) in both trials reveals that three children performed equally well in the first and third trials, three children's performance deteriorated, and three children improved. A paired samples *t*-test revealed that performance did not change significantly between the first trial ($M = 1.56$) and the third trial ($M = 1.56$), $t(8) = 0.00$, $p = 1.00$.

Regarding children's explanations for their experiment design, two children made a reference to a theory, "Because then the eggs have dots or not;" and two children made a reference to the experimental design, "Because one is different and two are the same."

Storybook. In the first trial, seven out of nine children (78%) correctly varied the focal variable. One child (11%) correctly controlled both control variables. No child designed a completely controlled experiment. One child set all variables to be the same and seven children set all variables to be different. These findings suggest that most children recognize the need to vary the focal variable in a test of a hypothesis. However, those who correctly varied the focal variable, also varied all other variables. Performance on the first trial with the storybook was marginally worse than performance on the first trial with the application, $t(16) = -1.89$, $p = .08$.

Most children (56%) did not provide a relevant explanation for their experimental design. Two children made a reference to a theory, "Because it comes from food that some are spotted and some white;" and two children made a reference to the goal of the experiment, "Well then he can see if this chicken with the corn gets spotted eggs or this chicken with the herbs."

In the third trial, seven out of nine children (78%) correctly varied the focal variable. Six children consistently varied the focal variable in both trials. Six children

controlled more variables in the third trial than they had in the first trial. One child controlled fewer. An equal number of children correctly varied the focal variable in the first and third trials. Five children showed stable performance in varying the focal variable. More than half of the children controlled more variables in the third trial than in the first.

Also in the third trial, three children designed a completely controlled experiment. One child set all variables to be the same and two children set all variables to be different. Half of children showed persistence in their experiment design from the second trial to the third trial. They were unable to apply the general instruction of varying the focal variable to the third trial and instead continued to vary the specific focal variable from the second trial. Looking at children's CVS scores for each variable in both trials (0-3 points) reveals that four children performed equally well in the first and third trials and five children improved. A paired samples *t*-test revealed that performance improved significantly between the first trial ($M = 1.0$) and the third trial ($M = 2.0$), $t(8) = -3.00$, $p = .02$.

Regarding children's explanations for their experiment design, one child made a reference to a theory, "Because some have spots and some don't;" two children made a reference to the experimental design, "Because when it's the same they lay the same eggs and when it's different they lay different eggs;" and one child referred to the goal of the experiment, "Because now he (FM) can see if this (herbs) or this (corn) is the one (that makes spotted eggs)."

4.3.5 Discussion

The first goal of the present study was to develop and evaluate an educational tool for assessing and promoting young children's abilities in Control of Variables Strategy. To this end, we took an iterative approach, emphasizing a quick initial design process followed by an evaluation and a second design process to address issues found. This

iterative process allowed us to identify issues early and attempt to address them, however, it is an ongoing process and the final iteration reported here could undergo additional iterations to address issues that were not solved in the second iteration as well as new issues that arose.

Three main issues remained that were not sufficiently addressed or that required greater changes than was possible within the scope of this work. One of the biggest issues was caused by the expected interaction with the tasks that was, in the end, not intuitive. The desired interaction was for children to select variables in alternating order from left to right, then moving down the page. Children either attempted to select a variable for the second chicken from the same set as for the first chicken instead of switching sides, or they proceeded down the page to the next variable instead of selecting the same type of variable for the second chicken. Even after making changes to the script to emphasize this order and the two different chickens, *selection from incorrect side* remained an issue. It did decrease over the duration of the task, such that children seemed to learn over time what the expected interaction was. However, the fact that it was still occurring after multiple interactions suggests that the interaction itself was not intuitive.

Future iterations should investigate different interface designs to determine a more intuitive interaction. For example, the variables on the left side of the screen could disappear or grey out after one of them has been selected. This should make children notice that they can no longer select from those variables, and that there must be another set of variables, causing them to look to the right side of the screen where the second chicken and second set of variables are located. Another possibility would be to completely change the interaction to reduce the need for switching, i.e., for there to only be one set of variables to choose from, which would also further reduce cognitive load (Sweller, 1988).

Another main issue was when children indicated impatience or selected buttons or magnets before they were supposed to. The changes made to the script and the addition of extra page before the main tasks was not sufficient to clarify when a child was supposed to interact with the tasks. For the application, this could be improved by introducing audio cues to indicate when a child should perform an action. Further, having the instructions integrated as audio could also help children to wait until the audio is finished before continuing. For the storybook, this issue is more difficult to address, since the adaptations to the script were not sufficient. Perhaps a more engaging narrative would keep children focused on the experimenter before they are supposed to interact with the task (Robin, 2008).

Finally, in the storybook, the main issue was with children's confusion about where to place the magnets. Initially, in the first iteration, we thought that allowing children to place the magnets where they felt made the most sense and was not an issue. However, after further consideration, we believe that reordering the magnets differently between tasks likely makes it difficult for children to keep track of how they have set up previous experiments and consequently, they cannot learn from their previous efforts. This issue could be addressed in two ways: first, we could redesign the tasks so that the variables are presented in the order that they are most likely to be placed on the page (sleeping location on top, nest type in the middle, food type at the bottom); we could also redesign the workspace such that it is clear where each magnet goes, like puzzle pieces. The application has a clear advantage in this case, since the variables simply appear after they have been selected and always in the same location, standardizing the appearance of each test. This interaction reduces extraneous cognitive load (Sweller, 1988) and potentially scaffolds children's interactions with the application (Vygotsky, 1980).

Two additional findings should be considered in future iterations or new development of tasks: type and length of interaction. The present study investigated children's drag-and-drop behavior in a warmup and their tapping behavior throughout the interaction with the application. Tapping was easier for children to perform than drag-and-drop, however most children could perform the drag-and-drop action and seemed to become more comfortable performing this action over time. Some children even carried this action over to the application from the warm-up, suggesting that it might be a more intuitive way to assign variables to the chickens. Future work should consider the trade-off between these two actions, and if choosing to implement drag-and-drop, should include a longer training session during which children can practice this action. Tapping did not require any practice and children generally did not have issues with performing this action.

Finally, the length of the interaction is important to consider, especially for different use cases. The storybook had a longer interaction duration, two minutes longer than the interaction with the application. Two minutes may not seem like much, but in experimental testing situations, or even in school teaching situations, an extra two minutes per child can add up quickly. The storybook interaction was also longer because the actions themselves took longer and required more effort for children to pick up and place the magnets. However, the longer interaction may be beneficial to learning.

In the third trial, three children in the storybook condition designed a controlled experiment, while none in the application condition did so. More than half of the children showed improvement in their CVS score and none deteriorated, whereas children in the application condition deteriorated, remained stable, and improved equally. Further, although the same number of children in both conditions provided relevant explanations in the third trial, the explanations provided in the storybook condition ($M = 9.44$; $SD = 6.64$

words) were longer than those provided in the application ($M = 5.71$; $SD = 3.99$ words). Further investigation should consider this trade-off between greater duration of interaction as more effort for the adult and the child, versus increased time-on-task as better for learning.

Additionally, the story context surrounding the present task was rather short and possibly not very captivating. More focus on creating an engaging and interesting story behind the task could help children to feel more invested in the task and keep them cognitively involved with the task for a longer period of time (Robin, 2008).

The second goal of the present study was to assess children's CVS abilities and investigate the potential for promotion of these abilities. Children seem to inherently understand the need to contrast variables: most children in both conditions spontaneously manipulated the focal variable in the first trial, in line with previous findings that young children are capable of producing contrastive tests (e.g., Bullock & Ziegler, 1999). However, many of those children manipulated all variables, not keeping any variables constant. Half of the children consistently varied the focal variable, and more than half increased the number of variables they controlled from the first to the third trial. Overall, performance appears to be better in the storybook condition, with more children showing improvement (and fewer getting worse) in their CVS score. Further, three children in the storybook condition and none in the application condition designed a controlled experiment in the third trial. It could be, as discussed above, that the greater time spent on the task in the storybook condition led to better learning of the strategy. It could also be that physically interacting with the variables (placing the magnets) directed children's attention more successfully to the task. These possibilities need to be further investigated.

There was no difference in performance between the first and third trials for children who interacted with the application. On the one hand, interacting with the

application did not have a negative effect on performance, but on the other hand, we did not see improvement as we did in the children who interacted with the storybook. It is possible that the children were distracted by their experiences of interacting with the tablet application, though there was no relation between whether or not children had previous experience with a tablet and their performance on either the experiment design or explanations. Thus, it appears that children who had no previous experience with a tablet were not any more or less distracted by the novelty of interacting with a tablet.

An interesting pattern was discovered regarding the design of experiments in the second and third trials. Half of the children in each condition showed persistence in their experimental design from the second to the third trial, varying the focal variable from the second trial also in the third trial. In the second trial children received explicit instruction and feedback on how to design a controlled experiment in general, but in this case with the sleeping location as the focal variable. Many children seemed to apply the rule “make sure the variable we want to find out about is different” to the specific variable rather than transferring that rule to the new focal variable in the third trial. This issue was unexpected but could be addressed by creating a more appropriate transfer task which uses the same set up but in a different context or with different variables. For example, the new task could be to find out if the method of milking a cow has an effect on whether the milk is creamy or smooth (Figure 4.9). This would eliminate the possibility of simply copying the set-up from the previous task.

Finally, age was not significantly related to neither children’s performance on the CVS tasks nor their ability to provide relevant explanations for their experimental design. These results are, on the one hand, in line with the finding of no effect of age in children’s selection of a controlled test, but on the other hand, contrary to the finding that children’s justifications improved with age (Study 2b). The tasks in the present study are, however,

very different and, we would argue, more difficult. The lack of an age effect in relevant explanations could indicate that the difficulty increased too much and the tasks are too difficult for this age group.



Figure 4.9. Prototype of near transfer task using cows instead of chickens.

The present study identified a number of important factors to consider when designing educational tools for children and provides suggestions for future iterations of the CVS task. In the next study, we take into account some of these suggestions and look more closely at children's abilities both in interacting with the task and in performing CVS.

4.4 Study 5b: Continued Iterative Development of a Tablet Application for Training CVS

4.4.1 Statement of Collaboration

In this section we describe the study design, methods, and results of a study investigating the potential of a digital task for use in scientific research on children's abilities in Control of Variables Strategy. In terms of this study, I collaborated with Viktoriia Rakytianska, a bachelor student in Media Informatics at LMU. I developed the concept of the task for investigating CVS and we co-designed the material of the study i.e., the story and the script. Ms. Rakytianska was responsible for the final design and illustrations, as well as the implementation of the task as a tablet application. She collected the data under my supervision at a number of kindergartens in Munich. The work was relevant to her in terms of her bachelor thesis on design elements for applications for children (Rakytianska, 2019). Thereby, she conducted a preliminary evaluation of the user experience with the task based on observation (video recordings of testing sessions). In terms of this thesis, I report some results of Ms. Rakytianska's analyses, which have been analyzed again by me. In addition, for the evaluation of the tasks, I have reported children's need for help, their engagement with the tasks, and children's interactions with the application. For the evaluation of children's ability to use CVS following instruction, I took a more fine-grained approach of evaluating whether children vary focal variables and control the control variables.

4.4.2 Introduction

The first goal of the present study was to design an intuitive interface for a digital task to be used in scientific research on children's abilities in Control of Variables Strategy. To this end, we took into account the findings from Study 5a, that the interface

design was unintuitive, and children struggled to interact with the application in the expected way even after multiple interactions and instruction from the experimenter. To address this issue and further investigate the interface design, we developed two new interface prototypes. Both reduced the total number of variables, such that instead of two sets of each variable type on each side of the screen, there was only one set of each variable type in the middle of the screen, in an attempt to reduce extraneous cognitive load generated by too much information on the screen. In the first interface we moved the chickens from the middle of the screen to the sides of the screen, essentially flipping the location of chickens and variables. In the second interface, we flipped the orientation of the screen from landscape to portrait and moved the chickens to the top of the screen.

In addition to the interface design adaptations described above, we wanted to place a greater emphasis on the storytelling introduction to the task. Storytelling is a powerful teaching tool and an important opportunity for engaging learners (Dreon, Kerper, & Landis, 2011; Robin, 2008). Placing a problem in an interesting context and generating emotional connections to the problem is an effective way to increase children's engagement with a task.

The second goal of the study was to further investigate children's abilities in using the Control of Variables Strategy. Because of the low performance in generating controlled experiments in Study 5a, both in an initial test, as well as after instruction, we decided to explicitly tell children what to do in order to design a "good test." With this procedure, we investigated if children could follow instructions and complete the interaction correctly to design a controlled experiment.

4.4.3 Method

4.4.3.1 Participants

A total of 17 preschool children were included in the analysis ($M_{\text{age}} = 66.26$ months, $SD = 7.15$; range: 53.29 - 76.20 months, 11 girls). All participants were typically developing children of lower- to upper-middle class background from a large German city. Parental informed consent and child assent was obtained for all children before the study. Eight children participated in testing the first iteration of the application ($M_{\text{age}} = 67.41$ months, $SD = 6.05$; range: 60.70 - 74.67 months, 4 girls), nine children participated in testing the second iteration of the application ($M_{\text{age}} = 65.34$ months, $SD = 8.14$; range: 53.29 - 76.20 months, 7 girls).

4.4.3.2 Materials

The same Farmer Meyer story context as in Study 5a was used. The design and implementation of the application in the present study focused on the importance of character design and storytelling. Farmer Meyer was designed based on Disney (*Santa's Workshop*) and Pixar (*Up*) characters to look like a happy old man (Figure 4.10). The chickens were also designed as key characters that speak and interact with Farmer Meyer. The introduction story was elaborated to be more entertaining and capture children's attention. It also emphasized the task at hand, to find out why a chicken is laying spotted eggs. The script of the story can be found in Appendix J. Additionally, two different interfaces were designed and tested. The first was an interface similar to that of the application in Study 5a, however, the chickens were moved to the sides of the screen and the variables were moved to the middle to focus attention more on the critical issue of which variables to assign (Figure 4.11A). Additionally, to reduce the amount of information on the screen and to address the issue in the previous study in which children

had difficulty selecting variables when there were two sets of all three variables present at once, we presented only one set of only one variable at a time.

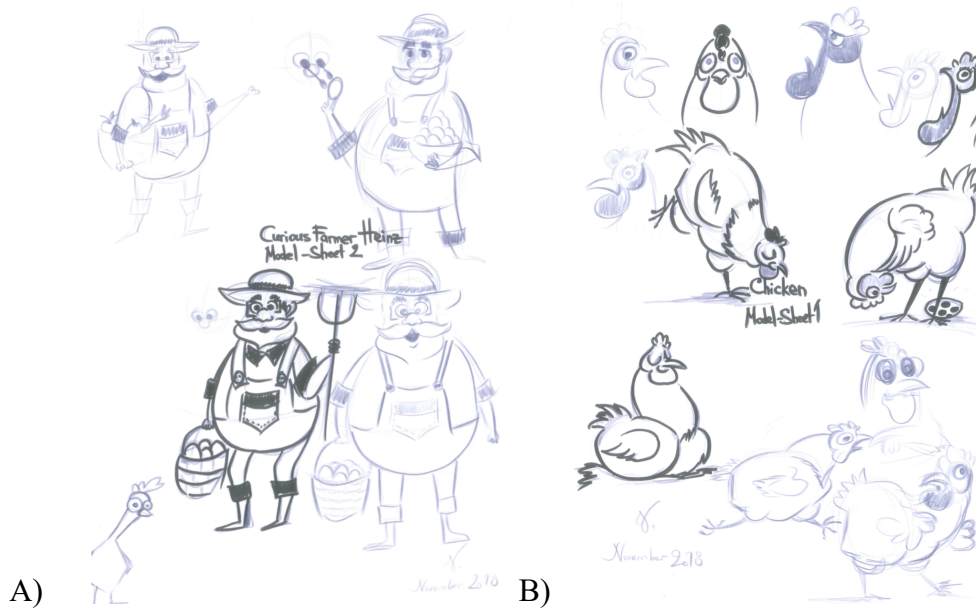


Figure 4.10. A) Farmer Meyer character design. B) Chicken character design. By Viktoriia Rakytianska (2019).

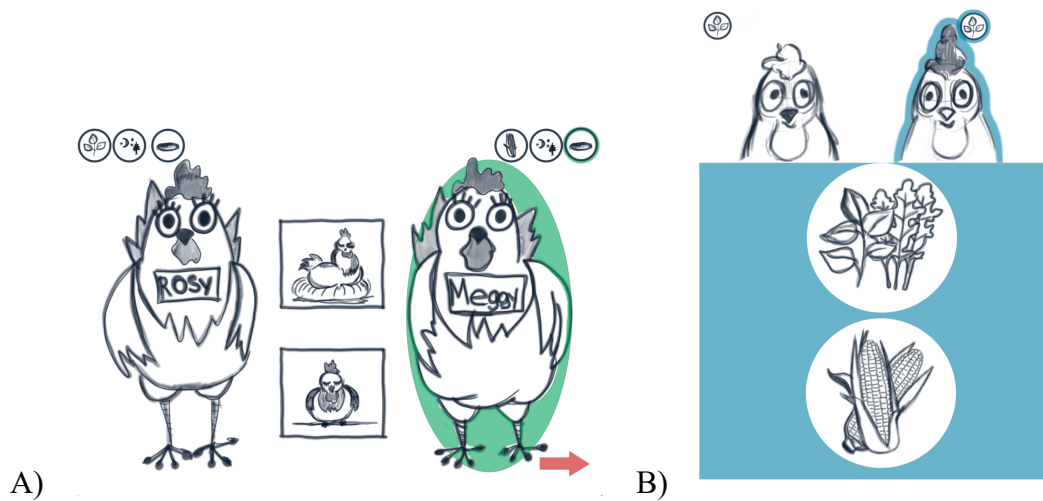


Figure 4.11. A) Interface 1 design. B) Interface 2 design

The second interface moved the two chickens to the top of the screen and placed the variables in the middle of the screen below the chickens (Figure 4.11B). Again, only one set of variables was presented at a time.

In both interfaces, the currently active chicken (to which they should assign a variable) was highlighted either with a colored background (Interface 1) or a colored outline (Interface 2). Once variables were selected, they appeared as small, round icons above the chicken for which they were chosen. Both interface designs were investigated initially, and the second design underwent a second iteration. For a detailed description of the design and development of the tasks, please refer to Rakytianska (2019).

4.4.3.3 Procedure

Data for this study were collected in February 2019. All sessions took place in local kindergartens in a separate, quiet room, and were video recorded. Children were tested individually in sessions lasting approximately 20 minutes. The experimenters were introduced to the children before the individual testings. When children entered the testing room, they were asked to sit at a low table and the experimenter sat across from them. The testing session began with a warm-up game in which children were asked to complete a simple matching puzzle. The purpose of the warm-up was to make the child comfortable with the experimenter and the testing environment prior to the main tasks. The session continued with the Farmer Meyer task. The task consisted of five phases: the introduction-story phase, the training phase, the question-story phase, the test phase, and the feedback phase. The test phase was repeated if children did not design a controlled experiment in the test phase (Figure 4.12).

The introduction-story phase began by introducing FM and his animals, particularly the chickens. The training phase introduced the three different variables one at a time and children could select a variable for each of two chickens. The purpose of this

phase was to familiarize children with the steps required to assign variables to the chickens. Next, the task continued with the question-story phase, in which the critical occurrence of Farmer Meyer's chickens producing two different types of eggs, spotted and plain, is presented. FM states that this difference could be due to the type of food, the sleeping location, or the type of nest. FM decides that he thinks that it is the type of food that makes a difference if his chickens lay spotted or plain eggs. Children were asked to help FM design a good experiment to find out if the food makes a difference in what type of eggs the chickens lay. The experimenter then told children explicitly: "To design a good experiment to find out if food makes a difference, we have to make sure the chickens get different foods, but everything else, the sleeping location and the nest type, should be the same." Children were asked if they understood what they had to do before continuing.



Figure 4.12. Procedure of the five phases of the task.

In the test phase, children could again assign each of the three variables to the two chickens, with the goal of producing a controlled experiment. If they did not produce a controlled experiment, the experimenter explained what was done incorrectly and reiterated how to design a good experiment. The children were then asked if they would like to try again, and if so, repeated the test phase a second time. Following the study, children were given a research certificate thanking them for their participation and informing them of the importance and role of "junior researchers."

4.4.3.4 Coding

Usability. The video recordings of testings were watched to identify usability issues occurring during children's interactions with the tablet application. Interactions were considered usability issues if actions did not generate expected effects.

Element not yet selectable. This issue occurred when children attempted to tap a button that was not yet activated. When this occurred, the correct button was briefly highlighted in blue to direct children to tap it. Once chickens were selected, they were outlined in the color corresponding to the current task. Once variables were selected, they appeared as icons above the chicken(s). Subcategories of this issue include:

Selecting the variable before the chicken. This issue occurred when children first attempted to select a variable before selecting a chicken rather than selecting the chicken before assigning it a variable.

Selecting the second chicken first. The order of the variable assignment was from left to right, first the chicken on the left, then the chicken on the right. This issue occurred when children attempted to select the chicken on the right first.

Area not selectable. This issue occurred when children attempted to tap on something that was not selectable. For example, on the introduction pages for the variables, the variables are images, not buttons.

Incorrect action. This issue occurred when children attempted to drag-and-drop an element rather than tap it.

Element does not activate. This issue occurred when children attempted to tap an element, but the element did not activate because their tap was not recognized. This could be due to the position of the finger or the force with which the child tapped.

Behavior. The video recordings of testings were watched to identify behavior issues occurring during children's interaction with the tasks.

Help. This issue occurred if a child verbally asked for help or how to do something, if the child looked questioningly at the experimenter, or if the experimenter had to intervene for the task to proceed.

Boredom. This issue occurred if a child looked away from the task.

Interest. Interest was measured in two ways: first by children's task-related spontaneous exclamations and second by their engagement with the experimenter, by looking at the experimenter particularly during the story phases.

Interaction. The number of attempts a child made to tap were recorded and compared to the number of actions those attempts were meant to initiate. Further, the ratio of children's taps to experimenter taps was compared as a measure of comfort with the task and ease of use.

CVS performance after instruction. Children's CVS performance in the test phase was coded by variable. They were given a score, correct or incorrect, for each variable assignment, which could total a maximum of three. For example, the focal variable was food, so children would have to vary this variable for one point and keep the other variables constant (sleeping location and nest type) for one point each. Please refer to the coding section of Study 5a for the detailed reasoning behind this coding scheme.

4.4.4 Results

4.4.4.1 First Iteration

Eight children participated in evaluating the first iterations of the application: Interface 1 ($n = 4$; mean age = 69.31 months (SD = 4.48); range: 65.10 - 74.67 months, 2 girls) and Interface 2 ($n = 4$; mean age = 67.92 months (SD = 5.78); range: 60.70 - 71.47 months, 2 girls).

Usability.

Interface 1. The duration of the interaction with the application through the first test phase ranged from 4:30 - 6:50 minutes ($M = 5:39$, $SD = 1:01$). The descriptives of the duration of each individual phase can be seen in Table 4.4.

Table 4.4

Length of each phase of the task for each interface and across iterations.

First iteration - Interface 1								
	Total to Test 1	Story 1	Training	Story 2	Question	Test 1	Explain	Test 2
Mean (SD)	5:39 (1:01)	0:53 (0:08)	1:48 (0:14)	0:41 (0:01)	0:34 (0:06)	1:06 (0:11)	0:34 (0:19)	0:46 (0:06)
Range	4:30-6:50	0:48-1:06	1:31-2:02	0:40-0:44	0:28-0:41	0:59-1:24	0:15-0:53	0:39-0:52
First iteration - Interface 2								
	Total to Test 1	Story 1	Training	Story 2	Question	Test 1	Explain	Test 2
Mean (SD)	5:16 (0:48)	1:05 (0:06)	1:39 (0:15)	0:45 (0:07)	0:28 (0:16)	0:59 (0:16)	0:39 (0:09)	0:32 (0:05)
Range	4:25-6:03	0:59-1:14	1:23-1:58	0:39-0:57	0:13-0:44	0:36-1:15	0:33-0:46	0:28-0:36
Second Iteration - Interface 2								
	Total to Test 1	Story 1	Training	Story 2	Question	Test 1	Explain	Test 2
Mean (SD)	6:33 (1:00)	0:59 (0:08)	1:58 (0:21)	0:43 (0:19)	0:47 (0:12)	0:59 (0:21)	0:39 (0:15)	0:52 (0:13)
Range	4:21-7:35	0:47-1:17	1:32-2:29	0:31-1:29	0:21-1:02	0:31-1:39	0:16-0:58	0:30-1:08

In the first iteration of Interface 1, three usability issues occurred. The most common usability issue, *selecting the variable before the chicken*, occurred when children attempted to assign a variable to a chicken by selecting the variable first, rather than the

chicken to which they wanted to assign the variable. This issue occurred 13 times ($M = 3.25$, $SD = 3.40$). This issue suggests that it was somewhat unintuitive to first have to activate one of the chickens to receive the variable. However, we wanted to ensure that children recognized there were two different chickens and that the variables children choose were then assigned to only one of the chickens at a time.

The second usability issue, *area not selectable*, occurred when children clicked on an area of the screen which was not selectable, for example the variables on the introduction pages. This issue occurred four times ($M = 1.00$, $SD = 1.15$) and was common across both interfaces. We addressed this issue by making the variables on the variable introduction screen look less like buttons. We removed the encompassing circle and background. We also removed the variable introduction pages in between each variable selection task during the test phase.

The third usability issue, *element does not activate*, occurred when children attempted to tap on a button, but the button did not activate due to the position of the finger or the force with which the child tapped. This issue occurred nine times ($M = 2.25$, $SD = 2.87$). This issue was also common across both interfaces. To address this issue, we made the arrow larger on some screens and changed it to a small star on other pages.

Interface 2. The duration of the interaction with the application through the first test phase ranged from 4:25 - 6:03 ($M = 5:16$, $SD = 0:48$). The descriptives of the duration of each individual phase can be seen in Table 4.4.

In the first iteration of Interface 2, three usability issues occurred. The most common usability issue, *area not selectable*, occurred when children clicked on an area of the screen which was not selectable, for example the variables on the introduction pages. This issue occurred 13 times ($M = 3.25$, $SD = 3.86$). This issue was addressed as described above.

The second usability issue, *element does not activate*, occurred when children attempted to tap on a button, but the button did not activate due to the position of the finger or the force with which the child tapped. This issue occurred seven times ($M = 1.75$, $SD = 2.22$). This issue was addressed as described above.

The third usability issue, *incorrect action*, occurred when children attempted to drag-and-drop an element rather than tap it. This issue occurred twice ($M = 0.50$, $SD = 1.00$). We did not make any changes to address this issue, however it could indicate that a drag-and-drop motion may be more intuitive for assigning variables to the chickens.

Behavior.

Interface 1. In the first iteration of Interface 1, children required *help* 50 times ($M = 12.50$, $SD = 6.56$). They were *bored or distracted*, as measured by looking away from the task, three times ($M = 0.75$, $SD = 0.50$). In the first iteration of Interface 1, children showed *interest* with task-related spontaneous exclamations 11 times ($M = 2.75$, $SD = 3.10$), and by looking at the experimenter, particularly during the story phases, 14 times ($M = 3.50$, $SD = 3.42$).

Interface 2. In the first iteration of Interface 2, children required *help* 55 times ($M = 13.75$, $SD = 8.18$). They were *bored or distracted*, as measured by looking away from the task, 12 times ($M = 3.00$, $SD = 1.41$). In the first iteration of Interface 2, children showed *interest*, with task-related spontaneous exclamations 13 times ($M = 3.25$, $SD = 8.18$), and by looking at the experimenter, particularly during the story phases, 15 times ($M = 3.75$, $SD = 3.20$).

Interaction.

Interface 1. Overall, the application required 50 taps to navigate through the story and complete the training and first test phase. In both story phases, the experimenter mainly initiated page turns. One child turned a page themselves. In the training and first

test phase, children completed an average of 74% of the required taps (range: 61-92%).

Children performed 28 additional unnecessary taps ($M = 7.00$, $SD = 6.88$, range = 2-17).

Interface 2. Overall, the application required 38 taps to navigate through the story and complete the training and first test phase. In both story phases, the experimenter mainly initiated page turns. One child turned five pages themselves. In the training and first test phase, children completed an average of 74% of the required taps (range: 46-100%). Children performed 24 additional unnecessary taps ($M = 6.00$, $SD = 6.68$, range = 0-14).

4.4.4.2 *Second Iteration*

Nine children participated in evaluating the second iteration of the application (mean age = 65.37 months ($SD = 8.14$); range: 53.29 - 76.20 months, 7 girls). The duration of the interaction with the application through the first test phase ranged from 4:21 - 7:35 ($M = 6:00$, $SD = 1:00$). The descriptives of the duration of each individual phase can be seen in Table 4.4. The average frequency of issues of the first and second iteration of Interface 2 can be seen in Table 4.5. We determined that the first interface still presented too much information on the screen and over-emphasized the chickens rather than the variables, so we continued with a second iteration of only the second interface.

Usability. In the second iteration of Interface 2, three usability issues occurred. The most common usability issue, *selecting the variable before the chicken*, occurred when children attempted to assign a variable to a chicken by selecting the variable first, rather than the chicken to which they wanted to assign the variable. This issue occurred 23 times ($M = 2.56$, $SD = 2.24$). This issue occurred slightly less often here than in the first iteration of Interface 1, where the same type of interaction was required. This type of interaction was not required in the first iteration of Interface 2, but because we wanted to ensure that children recognized there were two different chickens and that the variables children

choose were then assigned to only one of the chickens at a time, we implemented this interaction in the second iteration of Interface 2. In a third iteration not evaluated here, we addressed this issue by implementing a drag-and-drop interaction that will no longer require children to first select a chicken. They can simply select the variable and drag it to the desired chicken.

Table 4.5

Average frequency of usability issues in the first and second iteration of Interface 2. Green indicates an improvement and red indicates a deterioration between the two iterations.

Code	Description	Average Frequency 1st Iteration	Average Frequency 2nd Iteration
U1	Area not selectable	3.25	0.22
U2	Element does not activate	1.75	-
U3	Incorrect action	0.50	-
U4	Incorrect selection order	3.25*	2.56
U5	Incorrect distribution order	-	1.33
B1	Help	13.75	14.44
B2	Boredom	3.00	1.11
B3	Interest (exclamations)	3.25	3.33
B4	Interest (looking)	3.75	3.44
I1	Unnecessary taps	6.00	2.67

Note. *Frequency of issue in first iteration of Interface 1. The interaction that produced this issue was not included in the first iteration of Interface 2.

The second usability issue, *selecting the second chicken first*, occurred when children attempted to select the chicken on the right first, rather than the chicken on the left. The correct order was left to right. This issue occurred 12 times ($M = 1.33$, $SD = 1.58$)

and arose for the first time during this iteration. This issue was also addressed in a third iteration by allowing variables to be assigned to the chickens in any order.

The third usability issue, *area not selectable*, occurred when children clicked on an area of the screen which was not selectable, for example the variables on the introduction pages. This issue occurred twice ($M = 0.22$, $SD = 0.67$). This issue was minimized greatly from the first iteration, so no further changes were made.

In the second iteration of Interface 2, there were no issues with *elements not activating*. The changes, which made the arrow larger on some screens or changed it to a small star on other pages, appear to have successfully addressed the issue of the elements being appropriately sized and shaped for children to activate without issue.

Behavior. In the second iteration of Interface 2, children required *help* 130 times ($M = 14.44$, $SD = 9.53$). They were *bored or distracted*, as measured by looking away from the task, ten times ($M = 1.11$, $SD = 2.32$). Children showed *interest*, with task-related spontaneous exclamations 30 times ($M = 3.33$, $SD = 7.07$), and by looking at the experimenter, particularly during the story phases, 31 times ($M = 3.44$, $SD = 2.01$).

Interaction. Overall, the application required 46 taps to navigate through the story and complete the training and first test phase. In both story phases, the experimenter initiated page turns. No children attempted to turn the pages themselves during the story phases. In the training and first test phase, children completed 75% of the required taps (range: 71-80%). Children performed 24 additional unnecessary taps ($M = 2.67$, $SD = 1.73$, range = 1-6).

Preference for particular variables. Children did not prefer particular levels of the variables over others (i.e., corn vs herbs; all p -values $> .50$).

CVS performance after instruction. In the test phase of the CVS task, children were instructed to design an experiment to find out if the type of food makes a difference

in whether the chickens lay spotted or plain eggs. They were told specifically to make sure the two chickens received different types of food, but the same sleeping location and the same type of nest. Here we report the data from the full sample of 16 children. One child did not complete the second test phase due to a technical issue with the application. We will first report the performance in the first test phase. If children did not assign the variables correctly in the test phase, the experimenter gave feedback, repeated the instructions, and children repeated the test phase. No children declined the opportunity to complete a second test phase. Age was not related to CVS score in either test phase.

In the first test phase, 94% of children (15 children) correctly varied the focal variable. Eleven of those children (73%) varied all three variables. Thirty-one percent of children (5 children) correctly controlled at least one control variable and 25% (4 children) correctly controlled both variables. The four children (25%) who designed controlled tests did not complete a second test phase. In the second test phase, 83% of children (10 children) correctly varied the focal variable. (Two children who had correctly varied the focal variable in the first test no longer did so, and one child who had not varied it in the first test did so in the second test.) Five of those children (50%) varied all three variables. Fifty-eight percent of children (7 children) correctly controlled at least one control variable and 50% (6 children) correctly controlled both variables. Thirty-three percent (4 children) correctly designed controlled tests.

Overall, 25% of children could follow instructions to design a controlled test after receiving instruction once, an additional 25% of children could do so after receiving instruction twice, and 50% of children could not design a controlled test even after receiving explicit instructions twice.

4.4.5 Discussion

The first goal of the present study was to improve upon the design of a digital educational tool for assessing children's abilities in Control of Variables Strategy. Taking into account the findings from Study 5a that children struggled to interact with the application in the expected way even after multiple interactions and instruction from the experimenter, we developed two new interface prototypes. Both reduced the total number of variables, such that instead of two sets of each variable type, one on each side of the screen, there was only one set of each variable type in the middle of the screen, in an attempt to reduce extraneous cognitive load generated by too much information on the screen, as well as to reduce confusion and the potential for error. In the first interface we moved the chickens from the middle of the screen to the sides of the screen, essentially flipping the location of chickens and variables. In the second interface, we flipped the orientation of the screen from landscape to portrait and moved the chickens to the top of the screen. We determined that the first interface still presented too much information on the screen and over-emphasized the chickens rather than the variables, so we continued with a second iteration of only the second interface.

Two main usability issues arose, again, from children performing actions in an unexpected order, suggesting that the desired interaction was still not completely intuitive. Children struggled to pre-select the chickens before assigning them variables, and occasionally selected the chicken on the right first, rather than moving from left to right as expected. These issues show the importance of designing an application in a flexible manner to accommodate different types of interactions. We have addressed this issue in the final version of the application by implementing a drag-and-drop interaction which allows children to first select the variable and then drag it to the desired chicken. We also

made it possible to assign a variable to either chicken first. Further investigation is needed to assess the effectiveness of these changes.

The issue with children attempting to interact with elements that were not buttons was successfully addressed by making those elements look less like buttons. By removing the circular background behind the variables on introduction pages, we were able to greatly reduce the number of instances of children trying to click on those images. This issue and the solution emphasize the importance of designing for function. The design of elements should reflect the corresponding action (or lack of action) associated with them.

The issue of children unsuccessfully attempting an action, such as trying to click on the arrow to move to the next page but being unable to activate it, was addressed by adjusting the size, and sometimes shape, of the element. After making these changes, no children had issues activating elements. This issue illustrated the importance of designing for specific target groups. Young children have small fingers, but their motor skills are not as refined as those of adults, so it is important to design elements in a way that it is easy for young children to successfully select them.

Overall, children still required a significant amount of help when interacting with the application. Specifically, they were most often uncertain about clicking on the arrows or stars to move to the next page. The addition of an audio cue or perhaps motion in those elements could increase confidence that selecting those elements is the correct next step to progress through the application. In general, children seemed more interested than bored with the application. There were many instances of children spontaneously commenting on the chickens or the different variables, and children often looked at the experimenter while she was telling the story.

The second goal of the study was to further investigate children's abilities in using the Control of Variables Strategy. Because of the low performance in generating

controlled experiments in Study 5a, both in an initial test, as well as after instruction, we decided to explicitly tell children what to do in order to design a “good test.” With this procedure, we investigated if children could follow instructions and complete the interaction correctly to design a controlled experiment.

In the first test trial, 94% of children correctly varied the focal variable after instruction. This is better than performance in the third trial of Study 5a, in which 83% correctly varied the focal variable after instruction. These results are also in line with results from van der Graaf and colleagues (2015) in which all children in their sample ($M_{\text{age}} = 63$ months) were able to correctly design a test that manipulated one variable at least once in four trials with feedback. In both the present study and Study 5a, a majority of children varied all three variables (69%; 67%). That the majority of children still do this after receiving explicit instructions to keep the other variables constant is surprising. These results show the preference for contrastive testing or trying out everything at once. This may represent a difference between a scientific mindset and an “engineering” mindset (Schauble, Glaser, Raghavan, & Reiner, 1991), in which children focus on optimizing the outcome and as such, use different strategies like manipulating many things at once simply to see if anything has an effect before performing more targeted exploration. Indeed, due to time or money constraints in real-world contexts, it is not always feasible to change things one at a time (Zimmerman, 2007).

Twenty-five percent of children could follow instructions to design a controlled test after receiving instruction once and an additional 25% of children could do so after receiving instruction twice. Thus, 50% of children could design a controlled test after receiving instruction twice, which is similar to performance in the study by van der Graaf and colleagues (2015), in which 52% of children could correctly design a controlled test with three variables in at least one of four trials (after receiving feedback three times).

Future work could investigate if additional trials of the task in the present study would further increase performance. In this case, there would need to be a discussion about the type of materials used to investigate CVS abilities in young children, if performance on such as task as described in the present study is better than physical interactions with multiple variables as in the ramp task (van der Graaf et al., 2015).

Considering the results of both Study 5a and Study 5b, we can tentatively suggest that a tablet application as described could be used to assess CVS abilities in preschoolers, though the final version of the application should also be compared to a non-technological assessment as was done in Study 5a. The application could also potentially be used to promote CVS abilities through instruction and feedback, with the addition of more thorough training in the use of the application and more near-transfer task versions to prevent the carryover seen when children make multiple attempts to design experiments with the same variables. Future iterations could implement audio recordings of the story sections to standardize across all children and introduce audio cues for proceeding to the next phase. Audio recordings of the feedback for controlled and confounded experiments could also be implemented, however, the advantage of having a researcher or teacher give feedback is that it can easily be adjusted in the moment based on how well the child seems to understand the task and the feedback. A comparison of these two forms of feedback and how they relate to learning CVS could be investigated.

One issue with teaching CVS by having children produce experiments is the extra cognitive load placed on children by having to keep track of and manipulate variables themselves. As discussed at the beginning of the chapter, some research suggests that direct instruction and demonstration may be an effective way to communicate the Control of Variables Strategy to learners. The next section will discuss the development and assessment of a video tutorial for teaching preschoolers CVS.

4.5 Study 6: The Development of a Video Tutorial for Training CVS

4.5.1 Statement of Collaboration

In this section we describe the study design, methods, and the results of a study investigating both the potential of a video tutorial for use in scientific research and the effect of training through this task on children's abilities in Control of Variables Strategy. In terms of this study, I collaborated with Nicolai Schork and Marcel Schubert, bachelor students in Media Informatics at LMU. I developed the concept of the task for teaching CVS, based on an adaptation of the plane task (Bullock & Ziegler, 1999) and the Lego scientific reasoning tasks (described in Chapter 2). We co-designed the material of the study i.e., the story and the script. Mr. Schork and Mr. Schubert were responsible for the final design and implementation of the tutorial and the website on which it was hosted. They collected the data under my supervision at a number of kindergartens in Munich. The work was relevant to them in terms of their bachelor theses on guidelines for designing an educational video tutorial for children (Schork, 2018) and the effect of animation on children's acquisition of CVS (Schubert, 2018). Thereby, Mr. Schork conducted a preliminary evaluation of children's engagement while watching the tutorial based on observation (video recordings of testing sessions). Mr. Schubert carried out a preliminary analysis of the effect of the tutorial on children's CVS abilities. In terms of this thesis, I bring together the results of both of these bachelor theses, to look at how both engagement and animation are related to children's performance on the CVS tasks. I report some results of Mr. Schork's and Mr. Schubert's analyses, which have been analyzed again by me. In addition, I look further at children's engagement throughout the duration of the tutorial to categorize children into different levels of engagement.

4.5.2 Introduction

The first goal of the present study was to design a video tutorial to be used in scientific research on children's abilities in Control of Variables Strategy. We developed a story based on Bullock and Ziegler's (1999) airplane task, in which children have to determine whether or not the shape of the nose of an airplane influences how fast it flies, and designed it to appeal to young children. We also considered a number of important elements of digital storytelling design. Digital storytelling is also a powerful tool for increasing engagement and influencing learning (Robin, 2008). In the 1990's, the Center for Digital Storytelling, now ("StoryCenter"), indicated seven important elements of digital storytelling to consider when designing digital media for education. The College of Education at the University of Houston ("The 7 Elements of Digital Storytelling") has extended those original seven elements to ten elements, which we will describe below. Creators of digital stories should consider what is the overall purpose of the story and what is the narrator's point of view. They should consider using dramatic questions to keep the viewer's attention and which will be answered by the end of the story. The choice of content is important, and specifically, emotional content can help connect to an audience.

Regarding the design details of the digital story itself, the voice should be clear and powerful, the pacing of the narrative should be appropriate to the content and can vary, slowly or quickly, to support the progress of the story. An audio soundtrack should be meaningful and associated with the message of the story. The quality of the images, video, and other multimedia material should be appropriate for the content and audience. Creators should consider what amount of story detail to include and recognize when there is too little or too much detail. Finally, creators should consider the type and style of language to use and take care to be grammatically correct. Considering these elements in

the design of digital stories should help audiences connect to the content, which should, in turn, help viewers learn from the content.

In addition, we wanted to investigate the role of animation and its influence on engagement and learning. Animation is defined as a rapidly changing series of images that suggest a movement to the viewer (Rieber & Kini, 1991). They are further divided into three different types: transformations that illustrate changes in form, translations that illustrate changes in position, and transitions that illustrate the appearance or disappearance of objects (Lowe, 2003). Animations can be used to display a process and to present information in a way that reduces cognitive load (Höffler & Leutner, 2007).

Further, animations are considered attractive and intrinsically motivating for learners (Bétrancourt, 2005), and animation is an indicator of programming that is interesting to children (Zosh, Lytle, Golinkoff, & Hirsh-Pasek, 2016). A meta-analysis on the effects of animation revealed an advantage of animation over static pictures for learning, and that animation is even more effective when the animation is representational (and not just decorative), when the animation is highly realistic, and when the knowledge to be learned is procedural (Höffler & Leutner, 2007). Further, studies with adults have revealed better learning with a persona that was animated than when it was static (Baylor & Ryu, 2003; Mayer & DaPra, 2012). The advantage of animation is that information can be presented sequentially, in the correct order, and can take advantage of motion and storytelling (van der Meij & van der Meij, 2014).

The second goal of the present study was to investigate the effect of watching the tutorial on children's performance on the Lego CVS task. We developed two versions of the tutorial, an animated and a static version. Children performed one trial each of the three tasks of the Lego scientific reasoning tasks, then watched one of the tutorials, then performed a second trial each of the three tasks.

4.5.3 Method

4.5.3.1 Participants

A total of 18 preschool children were included in the analysis ($M_{\text{age}} = 62.89$ months, $SD = 9.18$; range: 49 - 79 months, 12 girls). Two additional children were tested but excluded due to language comprehension problems or incomplete data due to technical failure. All participants were typically developing children of lower- to upper-middle class background from a large German city. Parental informed consent and child assent was obtained for all children before the study. Nine children were assigned to the animated tutorial group ($M_{\text{age}} = 60.44$ months, $SD = 8.85$; range: 50 - 75 months, 6 girls) and nine children were assigned to the static tutorial group ($M_{\text{age}} = 64.11$ months, $SD = 9.35$; range: 49 - 79 months, 6 girls).

4.5.3.2 Materials

The warm-up and color vision tests described in Chapter 2 were used. The Lego scientific reasoning tasks described in Study 2b (Chapter 2) was used. The task was split into two equal parts to create a pre-test and a post-test, each consisting of one trial each of the confounded evidence task, the 2-variable task, and the 3-variable task.

To explain the Control of Variables Strategy, the tutorial presents two incorrect instances of CVS use, points out the mistakes, and subsequently corrects them. The context was adapted from the airplane task by (Bullock & Ziegler, 1999). The narrator introduces two characters, Max and Anna, who want to determine whether the shape of an airplane nose influences how fast it flies. They have competing theories: Max believes angular planes fly faster while Anna believes rounded planes fly faster. They conduct a series of experiments to find out if the shape of the airplane matters. In a first experiment, each character builds an airplane. They correctly manipulate the variable in question: Max's plane has an angular nose and Anna's has a rounded nose. However, there are two

other variables to consider and these are also varied. Max builds his plane out of wood with a jet fuel engine and Anna builds her plane out of metal with a propeller engine (Figure 4.13A). They then fly their planes and Max's plane is faster. He celebrates winning the race and learning that an angular plane flies faster than a rounded plane.

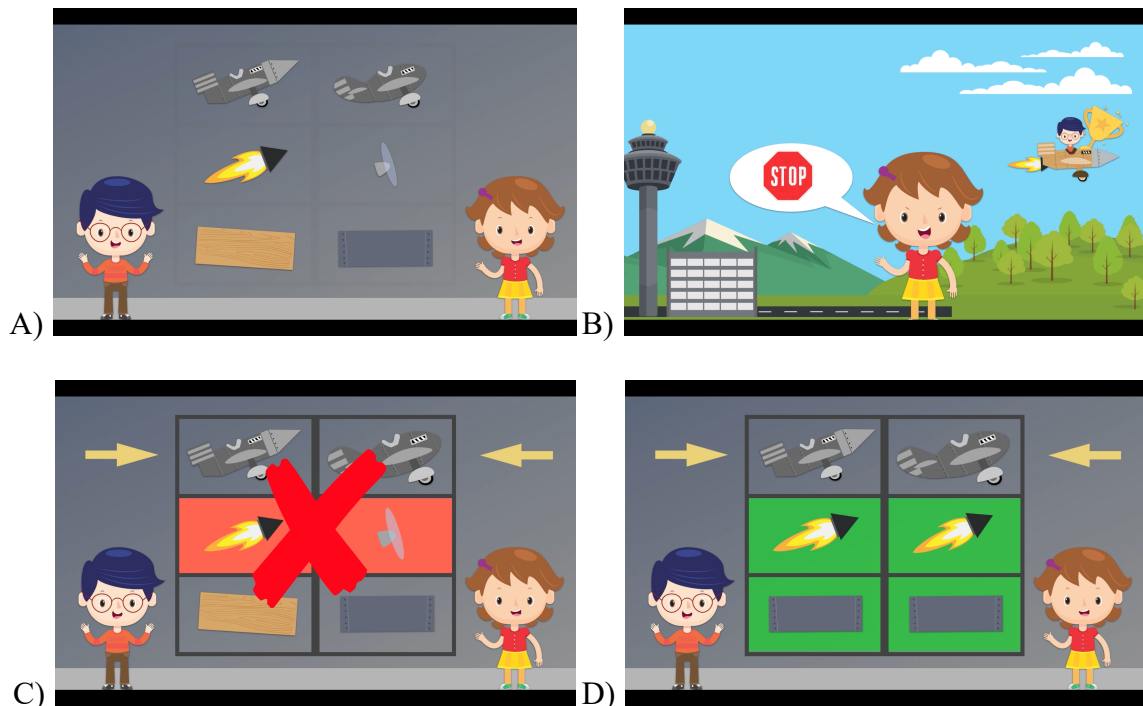


Figure 4.13. A) The materials of the two planes, B) the interjection after the race, C) the indication of a bad experiment, D) the instruction for a good experiment.

Anna then yells “Stop!” and the narrator asks, “Why did Anna yell “Stop!”?, Is something wrong with their experiment?” (Figure 4.13B). After a pause, the narrator continues to explain that the experiment was not fair because Max has a jet engine and Anna has a propeller engine. They can’t know if an angular or rounded plane is faster, because the engines are different. Anna rebuilds her plane to also have a jet engine and then they fly them again (Figure 4.13C). This time, Anna’s plane is faster. She celebrates learning that a rounded plane flies faster than an angular plane.

Max then yells “Stop!” and the narrator asks, “Do you know why Max yelled “Stop!”?, Do you know what is wrong?” The narrator points out that Anna’s plane is made of metal while Max’s is made of wood. He reminds the viewer that Anna and Max want to find out if an angular or rounded plane is faster, and this is why they already changed the type of engine so they both have the same one. But they also have to change the type of material. Max rebuilds his plane out of metal and the narrator emphasizes that now they have a fair experiment; if they want to compare angular and rounded planes, it is important that all the other things are the same (Figure 4.13D). They fly the planes one last time, and Anna’s plane is faster. The narrator concludes that they can now say a rounded plane flies faster. Because they only changed the shape of the plane, they conducted a fair experiment. The storyboard, including the script and frames from the tutorial, can be found in Appendix K.

Two versions of the tutorial were created, one static and one animated. Both videos included the same content, e.g., graphics, speaker, and music, but the animated version had 34 instances of animation that were not present in the static version. Examples of animation include thought bubbles appearing above the characters’ heads, the planes flying, and the variables being changed. For a detailed description of the design and development of the tutorial, please refer to Schork (2018) and Schubert (2018)

The CVS tutorial was presented as a full-screen video (5 minutes, 35 seconds) on a laptop (screen size: 18 x 28.5 cm, resolution: 2,560 x 1,600 pixels) placed in front of the child. The laptop keyboard was covered to prevent children from pressing any keys. The sessions were video recorded using two cameras. The internal laptop camera was used to record children’s faces while they watched the tutorial. A second camera and tripod were placed farther away to capture the full testing environment, children’s behavior, and their performance on the Lego scientific reasoning tasks.

4.5.3.3 Procedure

Data for this study were collected between February - March 2018. All sessions took place in local kindergartens in a separate, quiet room, and were video recorded. Children were tested individually in sessions lasting approximately 25 minutes. The experimenters were introduced to the children before the individual testings. When children entered the testing room, they were asked to sit at a low table and the experimenter sat across from them.

The testing session began with a warm-up game in which children were asked to complete a simple matching puzzle. The purpose of the warm-up was to make the child comfortable with the experimenter and the testing environment prior to the main tasks. Children were also given a color vision test to detect any color vision deficiencies. The procedure of the main tasks can be found in Figure 4.14. The session continued with three trials of the Lego scientific reasoning tasks; one trial each of the interpretation of confounded evidence task (ICE), the 2-variable task, and the 3-variable task.



Figure 4.14. Procedure of the seven phases of the task.

Next, children watched the CVS tutorial. They watched the video uninterrupted while the experimenter sat back from the table and looked at papers. Half of the children viewed the static version of the tutorial and the other half viewed the animated version. Following the tutorial, children completed the second half of the Lego CVS task, comprising a second trial of the 2-variable task, followed by a second trial of the 3-variable task, and concluded the session with a second trial of the ICE task. The detailed

procedure of the Lego scientific reasoning tasks can be found in Appendix E. Following the study, children were given a research certificate thanking them for their participation and informing them of the importance and role of “junior researchers.”

4.5.3.4 Coding

Coding of the Lego scientific reasoning tasks conducted as described in Study 1 (Chapter 2). The videos of children’s faces during the tutorial were coded for positive and negative reactions, adapted from Read et al. (2002) and Guo et al. (2014), as a measure of engagement. Positive reactions included smiling, laughing, excited bouncing, positive vocalizations, as well as indicators of concentration, such as furrowed brows. Negative reactions included frowning, shrugging, negative vocalizations, as well as signs of boredom, such as fiddling or playing around. If reactions could not be clearly determined as positive or negative, they were coded as neutral. The type and length of reactions was coded and used to categorize children as engaged, neutral, or unengaged.

Additionally, the number of times a child looked away from the screen and the length of time spent looking away from the screen were coded as a measure of attention (Chapman, 1997). The length of looks away from the screen were further distinguished as glances less than two seconds and glances longer than two seconds.

4.5.4 Results

Below we report the descriptive results of children’s engagement and their performance on the CVS task, looking at the animated and static tutorial separately.

4.5.4.1 Animated Tutorial

Nine children watched the animated version of the tutorial. Twelve positive reactions (from five children) and four negative reactions (from three children) were observed throughout the tutorial. Four children displayed only positive reactions throughout, two children displayed only negative reactions throughout, one child

displayed both positive and negative reactions throughout, and two children never displayed any reaction. Overall, four children were considered engaged, two were neutral, and three were not engaged.

The reactions of participants as measured over the duration of the tutorial indicate a clear pattern of mostly positive reactions in the first half of the tutorial and mostly negative reactions in the second half of the tutorial (Figure 4.15). The first negative reaction occurs after 160 seconds, and after 230 seconds, only negative reactions occur. Around 150s into the tutorial is the first explanation portion of the video, when the experimental set-up is discussed. At around 220 seconds, the second explanation portion occurs. Four children never looked away from the tutorial. Five children looked away at some point, for a total of 12 times ($6 < 2s$, $6 \geq 2s$). The average time spent looking away was approximately five seconds (ranging from $< 1s$ to $25s$; Figure 4.16).

CVS performance. Recall that we had children perform the Lego scientific reasoning tasks (from Study 2b) in two parts as a pre-test and a post-test. In this way, we could investigate if the tutorial had an effect on children's performance in this knowledge-lean CVS assessment task. In the Interpretation of Confounded Evidence task (ICE), knowledge claim performance (correctly responding that they do not know which bricks make the box light up) was 67% in the first trial and 63% in the second trial. Performance improved for one child (13%), deteriorated for one child (13%), and remained stable for six children (75%; one child did not provide a post-test ICE response). The one child who showed improvement had been categorized as not engaged. In the ICE task in Study 2b, 23% of children showed a deterioration in performance and 9% showed an improvement in performance between the first trial and the second trial, with only the CVS tasks in between. In the present study, the fact only one child showed a deterioration in performance (13%) suggests that though there does not seem to be an improvement in ICE

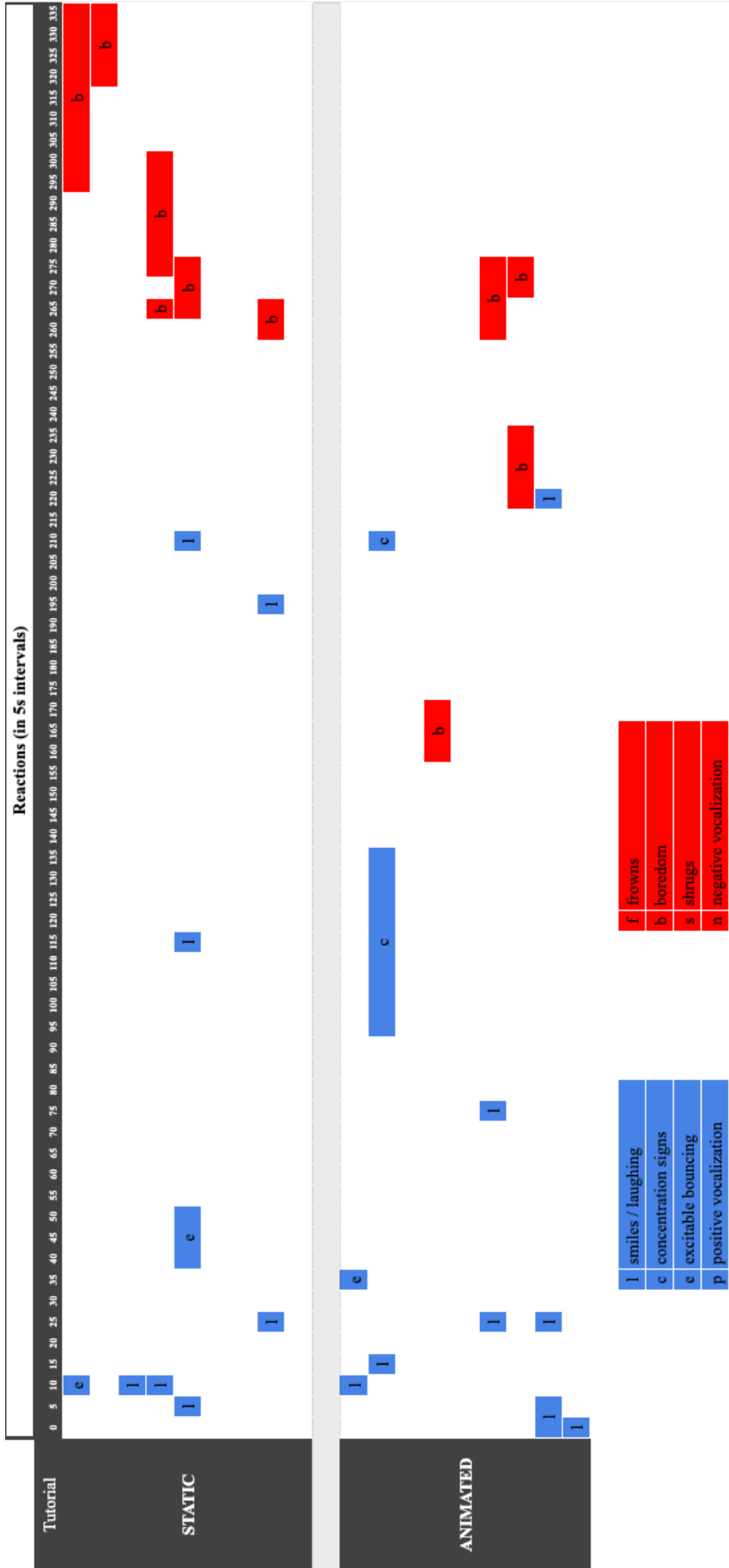


Figure 4.15. Frequency and duration of childrens reactions to the tutorial. Positive reactions are blue and negative reactions are red.

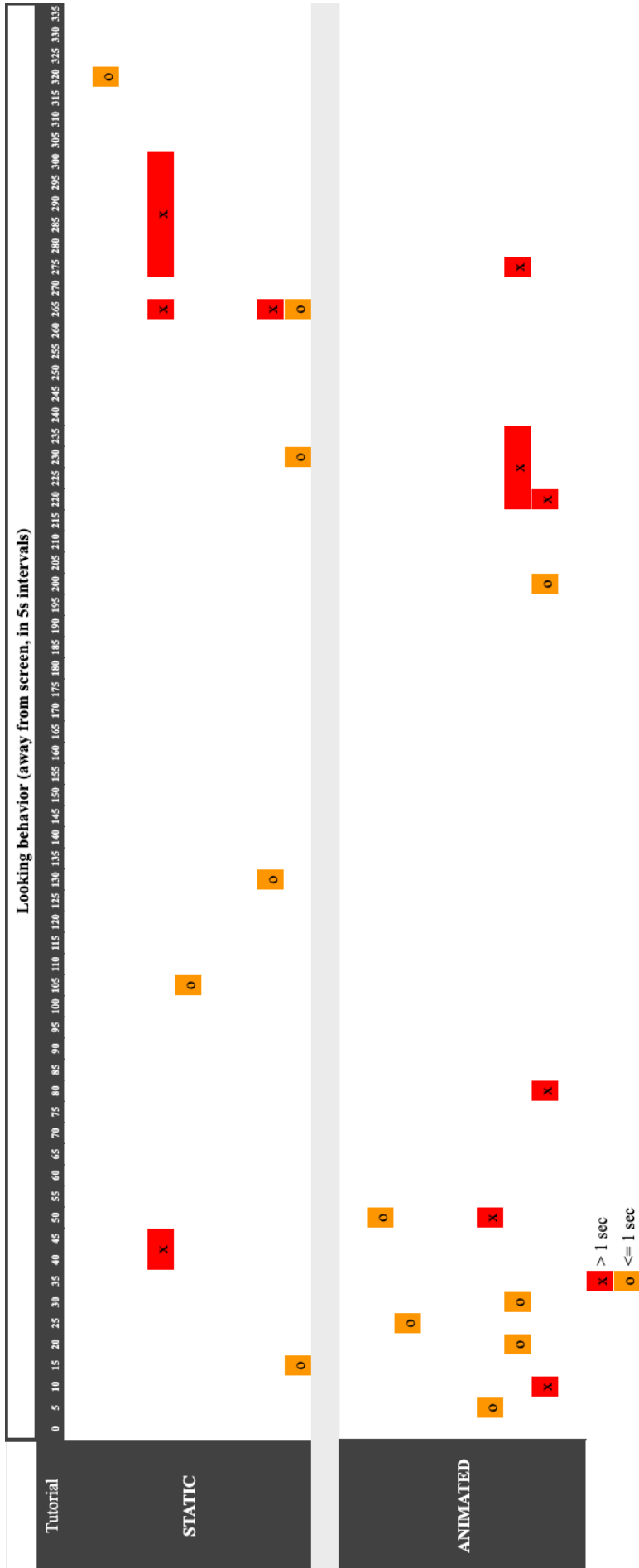


Figure 4.16. Frequency and duration of childrens looks away from the screen for the duration of the tutorial.

due to the tutorial, most children were stable in their performance, and the number of children improving was equal to the number of children showing a decline in performance.

In the 2-variable CVS task, choice performance (selecting the correct, controlled test) was 44% in the first trial and 78% in the second trial. Performance improved for four children after watching the tutorial (44%), deteriorated for one child (11%), and remained stable for four children (44%). Of the four children who showed improvement, one had been categorized as not engaged, one as neutral, and two as engaged. This pattern of performance was better than that in Study 2b, in which 16% of children showed improvement and 25% showed a decrease in performance.

In the 3-variable task, performance was 56% in both trials. Performance improved for two children (22%), deteriorated for two children (22%), and remained stable for five children (56%). Of the two children who showed improvement, one had been categorized as not engaged and one as engaged. This pattern of performance was similar to that in Study 2b, in which 21% of children showed improvement and 30% showed a decrease in performance. If we combine all three scores into a general scientific reasoning score, this reveals that two children improved their performance: one child improved from 1 to 2 points and one child improved from 0 to 3 points. Two children declined in their performance: one child decreased from 3 to 2 points and one child decreased from 2 to 1 point. Finally, four children maintained a stable performance of 2 points in both the pre- and post-test.

4.5.4.2 Static Tutorial

Nine children watched the static version of the tutorial. Nine positive reactions (from five children) and six negative reactions (from five children) were observed throughout the tutorial. One child displayed only positive reactions throughout, one child displayed only negative reactions throughout, four children displayed both positive and

negative reactions throughout, and three children did not display any reaction. Overall, two children were considered engaged, three were neutral, and four were not engaged.

The reactions of participants as measured over the duration of the tutorial indicates a clear pattern of mostly positive reactions in the first half of the tutorial and mostly negative reactions in the second half of the tutorial (Figure 4.15). The first negative reaction occurs after 255 seconds and from then on, only negative reactions occur. This also occurs during the explanation portion of the video, when the experimental set-up is discussed. Four children never looked away from the tutorial. Five children looked away at some point, for a total of 11 times ($7 < 2s$, $4 \geq 2s$). The average time spent looking away was approximately five seconds (ranging from $< 1s$ to $32s$; Figure 4.16).

CVS performance. In the Interpretation of Confounded Evidence task, knowledge claim performance (correctly responding that they do not know which bricks make the box light up) was 56% in the first trial and 22% in the second trial. Performance deteriorated for three children (33%) and remained stable for six children (66%). In the 2-variable CVS task, choice performance (selecting the correct, controlled test) was 78% in the first trial and 56% in the second trial. Performance deteriorated for two children (22%) after watching the tutorial and remained stable for seven children (78%). In the 3-variable task, performance was 78% in the first trial and 56% in the second trial. Performance improved for one child (11%), deteriorated for three children (33%), and remained stable for five children (56%). The one child who showed improvement had been categorized as engaged. If we combine all three scores into a general scientific reasoning score, this reveals that five children declined in their performance: two children decreased from 3 to 2 points, one child decreased from 3 to 1 point, and one child decreased from 2 to 1 point. Four children maintained a stable performance: two children with 2 points and two children with 1 point in both the pre- and post-test.

4.5.5 Discussion

Overall, the animated tutorial seems to have kept children engaged more and improved CVS performance more than the static tutorial. Children who watched the animated tutorial showed more positive reactions, and the total duration of positive reactions was also longer when compared to the static tutorial. Children showed fewer negative reactions, and the total duration of negative reactions was also shorter for the animated tutorial group compared to the static tutorial group. Based on their behavior, four children from the animated tutorial group and two children from the static tutorial group were categorized as engaged. Children who watched the animated tutorial were more likely to show improvement in all of the CVS tasks. In fact, only one child from the static tutorial group showed improvement in one of the CVS tasks. These findings are in line with what we would expect based on the literature on animation and engagement presented at the beginning of this study. The animations included in the tutorial served to capture and hold children's attention, which in turn helped them to learn the content of the tutorial and apply this to the second trials of the CVS tasks.

Interestingly, the first negative reactions emerged more than a minute and a half earlier in the animated tutorial group compared to the static tutorial, and exclusively negative reactions also occurred 25 seconds earlier in the animated tutorial group. The three children driving this result were categorized as not engaged: they were the only unengaged children and also the only children to show any negative reactions in the animated group. It is possible that these three children struggled to understand the content and thus stopped paying attention as a result. A larger sample size would be needed to investigate this result further and determine if this is due more to individual characteristics as opposed to the nature of the animated version of the tutorial.

The use of animation in this educational tutorial, which focused on teaching the control of variables strategy to young children, seems to be effective in both keeping children engaged and paying attention, and in improving their performance in an unrelated CVS task. Importantly, though, the results also suggest that the length of the tutorial was too long, with children losing interest after approximately four minutes. The tutorial should be adapted to shorten the length while maintaining the content of the CVS instruction.

This study is only a first step in investigating the effectiveness of animation for engagement and learning in young children. A larger experimental study would be necessary to determine if the trends revealed in this study can also be found in a larger sample and if they are significant. However, these preliminary results suggest a promising possibility to use direct instruction through video tutorial to promote young children's CVS abilities.

4.6 General Discussion & Recommendations

The first study discussed in this chapter presented the iterative development of an educational tool for assessing and promoting young children's abilities in Control of Variables Strategy (4.1). The second study continued with the development of the educational tool, addressing some issues discovered in the first study, and improving upon the design and structure of the task (4.2). The third study discussed the development and use of a video tutorial for instructing children in the use of CVS (4.3). This discussion section aims to bring together the results of these three studies and consider their theoretical and methodological implications for research with preschool children, as well as for early childhood education.

The findings of Study 5a presented in this chapter revealed the importance of an iterative design process for testing products with children and identifying issues early on

in the process. The study presented two versions of an educational tool: a paper-based storybook and a digital application. In regard to the storybook, we found that the average interaction time was approximately two minutes longer than that of the digital application, and the biggest difficulty children faced was related to choosing and placing magnets on the storybook pages. Children placed the magnets in different locations for each trial. The application had an advantage regarding variable placement because the variables appeared when clicked and always in the corresponding location. Further, looking at usability issues overall, we found that the application had fewer issues than the storybook. However, children who interacted with the storybook showed greater improvement in the number of variables they controlled after instruction. Further, after the storybook interaction, more children were able to produce a controlled test in the third trial than in the group who interacted with the application. This could be due to the longer interaction time, or the physical manipulation of the objects, potential distraction with the application, or this could be a result of individual differences in performance. Further investigation with a larger sample size would be required to unravel this finding.

Despite the trend toward better performance with the storybook, we proceeded in Study 5b with the further development of the application. Our goal in developing a tablet application as a tool for teaching CVS was to increase the ease of use for research and education purposes. The storybook required twelve printed and laminated sheets of paper and over 40 individual magnet pieces, which can easily be damaged or lost, especially in a classroom environment. The advantage of a tablet application is that it is contained within a compact mobile device, which is generally relatively durable.

Study 5a revealed a number of design issues that were addressed within the study as a second iteration, but a few issues remained. Primarily, these were the unintuitive alternating-order interaction and the need for a more engaging narrative around the task.

Study 5b presented in this chapter addressed the first issue by designing two new interfaces to present the chickens and the variables. The interface that was ultimately chosen for further development vastly reduced the amount of information present on the screen at one moment with the goal of focusing attention on the current choice and reducing extraneous load as a result of the overload of information.

However, the interaction of choosing variables for the chickens still presented an issue as it required children to do this in a specific order, first selecting the chicken and then the variable, and also proceeding from left to right across the screen. This issue and the similar issue from the first study emphasized the importance of designing for flexible interactions to allow children to interact with the application in the way they prefer and to avoid that children feel like they have done something wrong in choosing to interact in a particular way that the application does not account for. The final solution to this issue was to implement a drag-and-drop interaction allowing children to physically assign a variable to a chicken by dragging it to the corresponding chicken and to be able to do this starting from the left or right, as we believe this interaction will be more intuitive. Although the drag-and-drop interaction was slightly more difficult, as shown in the first study, children improved quickly with practice, and some even spontaneously transferred this interaction to the new task context. For this reason, we believe the drag-and-drop interaction would be appropriate for this age group with some pre-task practice.

The second issue of an engaging narrative was also addressed by placing the CVS task in an interesting context and generating emotional connections to the problem to increase children's engagement with the task. The story was rewritten with more emotional language, and one of the chickens was given an active role in the story. The characters were also redrawn to express more emotion. As we learned in the third study, designing for engagement is important and potentially beneficial to learning.

Looking at the CVS performance of children in both of these studies, we saw that most children correctly varied the focal variable, but many had trouble correctly controlling the other variables. Indeed, many children varied all three variables, essentially changing everything for the second chicken. This behavior may be due to curiosity, the desire to try everything out first, or a desire to produce “fair” experiments in which every variable is assigned, and none are left out. This issue could be addressed with an extended introduction period in which children could interact with all the variables individually before the main tasks.

In Study 5b we saw about a quarter of children could design a controlled test after one instance of instruction, and an additional quarter of children could do so after the second instance of instruction. These findings are in line with those of van der Graaf and colleagues (2015), who dynamically assessed children’s CVS ability by giving feedback after each test designed in four trials per variable level. They found that approximately half of four- to six-year-olds could design a controlled test with three variables in at least one of four trials with feedback after each trial. These results suggest that the tablet application developed in these studies could be used for assessing children’s abilities in CVS and potentially training these abilities through repeated instruction. Additional transfer tasks should be developed to avoid the carryover of experiment design and to keep children engaged and excited about helping Farmer Meyer and his animals.

Study 6 developed a video tutorial for teaching children the Control of Variables Strategy and found that the use of animation in digital storytelling was effective in keeping children engaged, as measured through positive reactions and looking time. The animated version of the tutorial improved performance in transfer CVS tasks, suggesting the tutorial could be useful in promoting CVS skills in preschool children. However, the tutorial was a

little too long and should be shortened to ensure children are engaged for the full duration of the tutorial and do not miss any of the instruction.

4.6.1 Recommendations

Bringing together the results of these three studies, we can recommend a number of factors to consider when designing educational tools for research with or education for preschool children. It is important to aim for the sweet spot of engagement, such that a tool is engaging and playful, immersing children in the task, but at the same time ensuring that the content is relevant and that other features are not distracting. This can be done by keeping in mind cognitive architecture, for example, referring to Cognitive Load Theory and principles of multimedia learning to decrease extraneous cognitive load and increase germane cognitive load. In the case of the interface design in Studies 5a and 5b, in each iteration, we continuously reduced the number of objects present on the screen at once to reduce the distractions and extraneous load.

One should consider the warm-up or training that is used prior to testing a product and ensure that it is appropriate and relevant. As we saw in Study 5a, children warmed-up with a puzzle application that featured a drag-and-drop interaction, and some children transferred this interaction over to our application, even though we did not implement a drag-and-drop interaction at that time. If the application only features a tapping interaction, the warm-up should also present this type of interaction.

Regarding the type of interaction in general, it is important to consider what is an intuitive interaction for the particular expected actions. Drag-and-drop was slightly more difficult than a tapping interaction, but because we wanted children to assign variables to the chickens, the act of selecting a variable and dragging it to the chicken is probably more intuitive than having to first select a variable and then select the chicken to which you

want to assign the variable. Further, interactions should be designed to be flexible to allow children to perform actions in any order they desire.

When designing objects and buttons within an application, it should be clear which objects are buttons and that they are clickable, and objects that are not buttons should not look like they can be clicked to trigger an effect. In other words, design should reflect the corresponding effect. This is especially important when designing for children because they could get demotivated quickly by attempting to perform an action that has no effect or think that they are doing something wrong. In the same line, buttons should be designed to ensure that they are of an appropriate size and shape to allow the target user to easily click them. For children, this means they should be big enough that children can easily select them despite their not-fully-developed fine motor skills.

The length of a task should be kept in mind. We saw that children began to lose interest in a video tutorial after about four minutes, but at the same time, remained engaged in the tablet application for around ten minutes. The difference, of course, is that the video tutorial required passive observation, which, for young children with a short attention span, can only last so long. However, when children are engaged in a task that requires them to do something, to perform actions, then they can remain focused longer, especially if they are enjoying their interaction with the task.

The above-mentioned recommendations should be considered when designing educational tools for young children. However, they are just a small selection of factors when it comes to designing for education, and for each new product developed, there will be new issues unique to that product. Therefore, it is important to take an iterative approach, to identify issues quickly, and to include children in the design process as much as possible to end up with the best possible final product.

The results of testing the application and tutorial presented here with preschool children suggest that they can be used to assess and potentially promote children's understanding and use of the Control of Variables Strategy. They could, however, benefit from further iteration and development to address the remaining issues identified through these studies.

5 General Discussion

The ability to reason scientifically is critical not only throughout one's education but also for active participation in modern society (Bromme & Goldman, 2014). However, mature scientific reasoning is difficult, and both children and adults struggle with many aspects of reasoning scientifically, for example, distinguishing between hypotheses and evidence or designing controlled experiments (Kuhn et al., 1988). For these reasons, researchers, educators, and organizations place great emphasis on teaching the skills involved in scientific reasoning throughout education, usually beginning in middle-elementary school (UNESCO, 2014). In contrast to the struggles children and adults have with reasoning scientifically, very young children and even infants show precocious causal reasoning abilities (Muentener & Bonawitz, 2018). For example, children can use covariation evidence to make causal inferences (Gopnik et al., 2001), they are sensitive to the informativeness of evidence (Cook et al., 2011), and they can intervene on causal systems to gain information (Gweon & Schulz, 2008). Many of these causal reasoning abilities seem related to scientific reasoning abilities and may be possible precursors to the development of scientific reasoning.

However, scientific reasoning requires a metaconceptual understanding of the distinction between theory and evidence, which allows children to recognize that a theory can be tested and revised, or that evidence can support or undermine a hypothesis (Kuhn, 1989, 2002; Kuhn & Franklin, 2007). Further, it requires the ability to generate hypotheses and then generate evidence to test those hypotheses. Thus, though children may spontaneously isolate or even control variables in causal reasoning assessments (Cook et al., 2011; van Schijndel et al., 2015), it is unclear if they are intentionally seeking knowledge through those behaviors, for example, by generating and testing particular hypotheses during exploration (Kuhn, 2002). All of these abilities require reflection and

thus rely on metacognitive abilities, such as the understanding of one's own knowledge and ignorance, which are developing around five years of age (Bullock et al., 2009; Perner, 1991; Rohwer et al., 2012; Sodian & Bullock, 2008; Wimmer & Perner, 1983).

To investigate how scientific reasoning develops in early childhood, this thesis brought together the literatures on causal reasoning and scientific reasoning, explored preschool children's abilities in scientific reasoning, examined scientific reasoning ability in relation to other cognitive abilities, and finally, took first steps toward determining if and how preschool children can be taught to control variables. This final chapter summarizes and discusses the findings presented in previous chapters, presents both theoretical and practical implications of these findings, and recommends future research directions to further elucidate the development of scientific reasoning in early childhood.

5.1 Summary of results

Children's and adults' abilities in scientific reasoning are affected by a number of factors, such as their prior belief or prior knowledge about task content, the outcomes of experiments, or the level of difficulty of the tasks (e.g., Kuhn et al., 1988; Tschirgi, 1980). It is possible that studies using such tasks underestimate children's abilities and, instead, highlight their struggles with difficult tasks. This could provide an explanation for the discrepancy between "precocious" causal reasoning ability, which is typically measured through knowledge-lean or decontextualized tasks, and deficient scientific reasoning, which is more commonly measured with tasks using scientific or everyday content. Such tasks can be quite complex in terms of task design or requirements to manipulate variables, and children may have prior beliefs or knowledge about the content.

To investigate preschool children's abilities in scientific reasoning, we developed novel knowledge-lean tasks using theblicket detector paradigm (Gopnik & Sobel, 2000) to limit the influence of prior knowledge or beliefs on preschool children's abilities in

scientific reasoning. Specifically, we assessed their ability to recognize when evidence is confounded and that, as a result, they cannot know something conclusively, as well as their ability to recognize a controlled test of a hypothesis using the Control of Variables Strategy (CVS). Critically, and to distinguish from causal reasoning, we included a task that required children to reflect on their own ignorance resulting from confounded evidence and tasks that required children to specifically test a hypothesis by selecting a controlled test. In this way, we could take advantage of and build upon children's precocious causal reasoning abilities but target the scientific reasoning abilities that are not typically assessed with knowledge-lean paradigms.

In summarizing and discussing the results, we will begin with Study 1, which investigated the stability of children's scientific reasoning abilities using a novel knowledge-lean task before continuing with Studies 2a and 2b, which represented more robust assessments of children's scientific reasoning abilities in one session. We will then continue with Study 4, which investigated the structure and correlates of scientific reasoning in preschool, again using the knowledge-lean tasks, and relate this Study to Studies 2a and 2b. We will then discuss adults' abilities on these same knowledge-lean tasks (Study 3). Finally, we will discuss Studies 5a and 5b, which developed a tablet application for assessing and teaching CVS, and Study 6, which developed a video tutorial for teaching CVS.

5.1.1 Study 1: Stability of preschoolers' scientific reasoning abilities

In Study 1, we investigated the stability of children's (three- to six-year-olds) scientific reasoning abilities over a short interval. Children performed one trial each of the Interpretation of Confounded Evidence task (ICE), the 2-variable CVS task, and the 3-variable CVS task in a first session and then performed a second trial each in a session two weeks later.

5.1.1.1 Stability

In all three tasks, the robust performance as measured by correct responses plus correct justifications was stable between the two sessions. However, the selection of a 2-variable controlled test improved significantly in the second session, suggesting that even without explicit instruction and with only the feedback from the task itself, children showed improvement in recognizing a controlled test with two variables. This is consistent with studies showing that repeated interactions with systems for designing experiments improved children's performance in designing controlled tests (Kuhn, 2007b; Schauble, 1990).

5.1.1.2 ICE tasks

Further, taking into account performance across the two sessions, two-thirds of the children provided correct knowledge claim responses indicating they could not know which of the bricks made the box light up in the ICE task at least once and a quarter of the children provided a robust ICE response indicating why they could not know which bricks made the box light up at least once.

5.1.1.3 CVS tasks

In the CVS tasks, over 80% of the children selected a controlled test with two variables at least once and over 60% selected a controlled test with three variables at least once, both significantly more than expected due to chance. We also found that older children in this sample performed better than younger children on the CVS tasks.

5.1.1.4 CVS justifications

Children provided relatively few relevant justifications for their test selections, with less than a quarter of all justifications being considered relevant. Again, older children in this sample performed better than younger children in providing relevant justifications.

5.1.2 Studies 2a & 2b: Preschoolers' scientific reasoning abilities

Following Study 1, we made a number of improvements to the materials and protocol resulting in an intermediate version of the tasks in Study 2a and a final version of the tasks in Study 2b. Specifically, in Study 2a, we changed the wording of the test selection and justification questions from “Which stick do you want to pick and why?” to “Which stick is the best to find out if the X brick makes the box light up and why is that the best stick to find that out?” to place an emphasis on the testing of a hypothesis and that they should choose in relation to the hypothesis and not based on preference. In Study 2b, we improved the interpretation question from “Now do you know if the X brick makes the box light up?” to “Is the X brick a lighter, not a light, or can you not know?” In Studies 2a and 2b, we investigated children’s scientific reasoning abilities in one session with two trials of each of the three tasks.

5.1.2.1 ICE tasks

Looking at children’s performance across the two trials, 55% to 70% of the children provided correct knowledge claim responses indicating they could not know which of the bricks made the box light up in the ICE task at least once and 28% to 40% of the children provided a robust ICE response indicating why they could not know which bricks made the box light up at least once.

Children performed worse on the ICE task in Study 2b than in Study 2a. This may be explained by the addition of another combination in the familiarization before the main tasks. We introduced a combined stick of four bricks in the familiarization and had children identify which of the four bricks were lighters as a memory check. This may have then transferred over to the ICE task in which we asked children if they could know for sure which of four bricks in a novel stick were lighters. Consequently, children may have remembered from the familiarization that they could identify the lighters and may have

thought that they should also be able to do that in the ICE task, even though these were novel bricks. If this is the case, then performance on the ICE task in Study 2b is an even more robust measure since the children would have had to overcome an even stronger tendency to claim that they know which bricks make the box light up.

This explanation is also supported by the fact that we found an effect of age on the ICE task for the first time in Study 2b, which would suggest that indeed the task may have been more difficult and consequently older children were more likely to succeed on it. In both studies performance declined from the first trial to the second trial, possibly as a result of fatigue or because children felt they should have learned something over the course of the experiment and felt they should now know which bricks made the box light up.

5.1.2.2 CVS tasks

In the CVS tasks, 82% to 86% of the children selected a controlled test with two variables at least once, and 68% to 73% selected a controlled test with three variables at least once, similarly to the level of performance in Study 1 (83%, 62%). However, statistically, we observed opposite patterns of performance between Studies 2a and 2b. In Study 2a, performance on the 3-variable CVS task was significantly better than expected due to chance, while this was not the case for the 2-variable task. In Study 2b, performance on the 2-variable CVS task was significantly better than expected due to chance while performance on the 3-variable task was marginally better than expected due to chance.

Based on the literature suggesting that task difficulty should increase with an increasing number of variables (Tschirgi, 1980), and considering that the 3-variable task also presented three different choice options, we would expect that the 2-variable task should be easier than the 3-variable task, as we saw with the performance in Study 2b.

Thus, the pattern of performance in Study 2a is unexpected. Especially since there are only two choices and one of those choices provides no information relevant to the hypothesis, it is surprising that children would perform more poorly on this task compared to the objectively more difficult 3-variable task. This finding could suggest that the 2-variable task is not tapping into the abilities we intend to measure with it. In Studies 2a and 2b, we did not find any effect of age on the CVS tasks, as in Study 1.

5.1.2.3 CVS justifications

In Study 2a, children still provided relatively few relevant justifications for their test selections, with about a quarter of all justifications being considered relevant. In Study 2b, significantly more relevant justifications were provided (almost 40%). The only other difference between Studies 2a and 2b (other than the familiarization) was the change to the interpretation question to make it clearer. Thus, it is possible that by rewording the interpretation question, to make children reflect on whether the X brick was a lighter or not or if they could not know, subsequently facilitated children's ability to provide a relevant justification for their choice of test. As a reminder, the original interpretation question asked if children could now know whether the X brick was a lighter or not, while the improved question asked children if the X brick was a lighter, not a lighter, or if they could not know.

In both Study 2a and Study 2b, and as in Study 1, we found an effect of age on children's justifications, such that older children were better able to provide relevant justifications for their choices. Finally, in Study 2b, we found that children were more likely to provide relevant justifications for the 3-variable CVS task than for the 2-variable CVS task. This pattern is also somewhat unexpected but could be explained by the fact that the 3-variable sticks contained three bricks and consequently two control variables in

the case of the correct stick, thus, this may have facilitated children's ability to refer to one or both of the control bricks in justifying their choice.

5.1.3 Study 3: Adults' scientific reasoning abilities

Because research has also shown that not all adults are fully competent in scientific reasoning and that some adults struggle with many of the same issues as children, such as reasoning fallacies or confirmation bias (Kuhn et al., 1988; Kuhn et al., 1995), we wanted to investigate adult's scientific reasoning abilities using the same knowledge-lean task as in the studies summarized above. In addition, having adults complete these scientific reasoning tasks can serve as a validation of their use as a measure of scientific reasoning ability.

5.1.3.1 *Spontaneous responses*

Adults performed well in providing initial responses to all of the tasks: all participants correctly recognized they could not know which bricks made the box light up in the ICE task, 93% of participants selected a controlled test with two variables, and 95% of participants selected a controlled test with three variables. Where adults seemed to struggle in these tasks was in providing elaborated explanations or justifications for their responses.

5.1.3.2 *ICE explanations*

In the ICE task, 85% of participants received at least one point for their explanation, providing at least one relevant statement regarding their inability to know which bricks made the box light up. However, ~8% were unable to provide a valid explanation, and another ~8% actually provided an incorrect statement, such as claiming that none of the bricks were lighters. These explanations were related to participants' age, such that older participants provided higher quality explanations than younger participants.

5.1.3.3 *CVS justifications and interpretations*

In the CVS tasks, participants had two opportunities to provide justifications: for their selection of a test and for their interpretation of the outcome of their test. In the 2-variable CVS task, of participants who selected the correct test, 68% received at least one point for a valid justification. However, one participant was unable to provide a valid explanation and another ~30% of participants actually provided an incorrect justification, such as claiming that if their test made the box light up, they could know that the control brick was a lighter. When interpreting the outcome of their test, 10% of participants made an incorrect inference from the evidence. Surprisingly, and in contrast to the pattern found for the ICE explanations, 2-variable CVS justifications were negatively related to age such that older participants provided lower quality justifications than younger participants. In the 3-variable CVS task, of participants who selected the correct 97% received at least one point for a valid justification, providing at least one relevant statement regarding control of variables, with only one participant unable to provide a valid explanation. When interpreting the outcome of their test, 5% of participants made an incorrect inference from the evidence. 3-variable CVS justifications were positively related to age, such that older participants provided higher quality justifications than younger participants. Performance on the ICE explanation positively predicted both 3-variable justification and interpretation scores.

Interestingly, performance on the 3-variable task was better than performance on the 2-variables task across all measures (choice, justification, and interpretation). Overall though, the scores for the ICE explanation and CVS justifications and interpretations were relatively low, reflecting that participants did not provide fully elaborated explanations. However, participants provided higher quality interpretations compared to justifications in the CVS tasks, revealing that it was easier to explain the inferences one could draw from

the outcome of an experiment than to explain why one chose a specific experimental design.

5.1.3.4 Other findings

Five individuals forgot about the fact that the original stick had already been placed on the box and lit up. Ten individuals mentioned that they were assuming that bricks of the same color had the same effect on the box. Three individuals mentioned that they had to assume that the sticks of three performed in the same way as the sticks of two, because they had not seen a stick of three previously.

5.1.4 Study 4: Structure and correlates of scientific reasoning in preschool

In Study 4, we investigated the structure of scientific reasoning and its relation to other cognitive abilities. We investigated children's (four-year-olds) scientific reasoning abilities using the same knowledge-lean tasks, except for the exclusion of the interpretation question, which asked children to interpret the outcome of the experiment. We shortened the task to avoid fatigue since this task was just one of many in a ~80 minute session. In addition to the scientific reasoning tasks, we also measured intelligence, language abilities, executive functioning, and Theory of Mind.

5.1.4.1 ICE tasks

Looking at children's performance across the two trials, 68% of the children provided correct knowledge claim responses indicating they could not know which of the bricks made the box light up in the ICE task at least once and 20% of the children provided a robust ICE response indicating why they could not know which bricks made the box light up at least once.

5.1.4.2 CVS tasks and CVS justifications

In the CVS tasks, 88% of the children selected a controlled test with two variables at least once, and 58% selected a controlled test with three variables at least once, both

significantly more than expected due to chance. Children provided very few relevant justifications for their test selections, with just 11% of all justifications being considered relevant. We replicated the effects of Study 2b that children provided more relevant justifications for controlled tests than for contrastive tests.

5.1.4.3 Factors affecting performance

We found a significant effect of gender on the CVS tasks and a marginally significant effect of gender on the ICE task, such that girls performed better. Further, we found that robust performance on the first trial of the ICE task was negatively related to children's ability to provide relevant justifications. This result was unexpected because we expected that having an understanding of the inconclusiveness of confounded evidence would positively predict the ability to select a controlled test and to explain that selection. However, it is possible that children who provided a robust explanation early on (in the first ICE trial) were then no longer motivated to provide elaborated justifications for the remaining tasks.

5.1.4.4 Correlates of scientific reasoning

To investigate the relation between scientific reasoning and executive functioning and Theory of Mind, we generated a sum score from the ICE task and both CVS tasks, including responses and justifications. We found that inhibition and Theory of Mind both contributed significantly to a multiple regression model predicting scientific reasoning. Further, we found that Theory of Mind appeared to be necessary in order to be successful in scientific reasoning: more than half of children showed mastery of Theory of Mind but not of scientific reasoning, while less than 2% of children showed the opposite pattern of mastery of scientific reasoning but not of Theory of Mind. Scientific reasoning was not related to age, intelligence, language, working memory, planning, or cognitive flexibility.

5.1.5 Studies 5a, 5b, & 6: Promotion of scientific reasoning with digital training tools

Because of the importance of scientific reasoning and the findings of beginning abilities in scientific reasoning in preschool, we wanted to further investigate the possibility of promoting scientific reasoning in preschoolers (four- to six-year-olds). We chose to develop digital materials in the form of a tablet application and a video tutorial to teach scientific reasoning since the use of technology and the consumption of digital content is becoming more and more common among young children. We considered cognitive architecture, instructional design, and multimedia learning, as well as storytelling and animation in designing our materials (Bétrancourt, 2005; R. E. Mayer, 2014; Robin, 2008; Sweller, 1988; Vygotsky, 1980). We took an iterative approach in developing the application and engaged children as testers of the application, gathering feedback on usability through observation and implementing this feedback in further iterations.

5.1.5.1 Iterative evaluation

In Studies 5a and 5b we investigated the design and development of a tablet application for assessing and training Control of Variables Strategy abilities in preschool children. In Study 5a, we compared both usability and CVS ability with a tablet application and a paper-based storybook. The comparison of the tablet application and the storybook revealed that after two iterations, the application had fewer usability issues than the storybook and that, in some cases, issues with the storybook increased in the second iteration. In Study 5b, we continued the iterative development of the tablet application and assessed children's ability to design a controlled test after instruction. We found the second iteration of this version of the application was also able to reduce the overall number of usability issues with the application. The results of both of these studies

emphasize the importance of iterative design and testing, particularly with preschool children. Repeated improvements and testing are generally beneficial for reducing issues, but because new iterations can also introduce new problems, this process should undergo a number of cycles. This process is not specific to digital tools, but also useful for any research or educational materials.

5.1.5.2 Study 5a

In Study 5a, we found that the storybook interaction took over two minutes longer on average than the application interaction, likely as a result of the need to select and place magnets and additionally due to children's hesitation to perform those actions. Further, the practicality of the storybook interaction was brought into question considering the need for many materials, specifically over 40 individual magnets, which could be easily lost in testing situations or in schools. The application has a clear advantage here in that the tasks are contained within one tablet, and tablets have been shown to be pretty robust for use with preschool children. Looking at indicators of impatience, both mediums performed equally well. Children using the storybook seemed to be slightly more confused but required an equal amount of help from the experimenter as children using the application. We did not record instances of boredom or engagement, but we qualitatively determined that the interaction was not engaging enough. In the next Study (4.2), we focused on storytelling both in the script and in the illustrations of the characters.

CVS performance. After training, the storybook interaction produced better performance in the third trial compared to the application. Three children in the storybook group were able to design a controlled experiment in the third trial, while none of the children in the application group were able to do so. Thus, training with the storybook was more effective than training with the application, and training with the application had no effect on performance. Children contrasted the focal variable 75% of the time. However,

children also often varied all of the variables about half of the time. Children provided relevant justifications about 40% of the time. Half of the children showed persistence from the training trial to the third test trial, carrying over the specific instructions regarding varying a particular variable but not generalizing this to the new hypothesis.

5.1.5.3 Study 5b

In this study, we observed children's interest and engagement in terms of looking at the experimenter while she explained something or spontaneously discussing the events in the application. In addition, we included observations of boredom as measured by children looking away from the application. Overall, children were more interested ($M = 3.77$) than bored ($M = 1.11$) and seemed to enjoy their interactions with the application. Further, when children did not successfully design a controlled task, they received feedback and were asked if they wanted to try again. No children declined the opportunity to complete a second trial.

CVS performance. In Study 5b, we gave children direct instructions on how to design a controlled experiment, specifically how to design a controlled experiment to test the hypothesis that type of food affected whether chickens laid spotted or plain eggs. We found that after one instance of instruction 94% of children initially contrasted the focal variable. However, 73% of children also varied all of the variables. A quarter of children could design a controlled experiment and after a second instance of instruction an additional quarter of children could design a controlled experiment. Thus, half of the children could design a controlled experiment with three variables after two instances of instruction.

5.1.5.4 Study 6

In Study 6, we created a video tutorial for teaching children CVS and assessed if different versions of the tutorial had any impact on children's performance on the

scientific reasoning tasks from Chapters 2 and 3. One version of the tutorial was static and the other version included over 30 instances of animation throughout. We found that children who viewed the animated tutorial were considered more engaged, as measured by having more positive reactions and fewer distracted glances away from the tutorial. Additionally, children who viewed the animated tutorial were more likely to show improvement in the scientific reasoning tasks. Of the children who watched the animated tutorial, two children showed improvement, two children showed deterioration, and four children showed stable performance. Of the children who watched the static tutorial, no children showed improvement, five children showed deterioration, and four children showed stable performance.

5.2 Discussion

5.2.1 Preschooler's knowledge-lean scientific reasoning

Studies 1, 2a, 2b, and 4 all investigated preschool children's scientific reasoning abilities with a novel knowledge-lean task. The results of these four studies suggest that preschoolers show a nascent understanding of the inconclusiveness of confounded evidence and a beginning ability to recognize a controlled test of a hypothesis with both two and three variables.

5.2.1.1 Control of Variables Strategy

The beginning ability to recognize a controlled test of a hypothesis was present in children aged three to six, with a majority of children able to recognize controlled tests at least once in two trials in tasks with both two and three variables. Around 35-45% of children could consistently select a controlled test with two variables, and around 17-24% of children could consistently select a controlled test with three variables. Thus, though fewer children showed consistency in this ability, there seems to be some early, stable ability to recognize controlled tests. These findings provide support for recent advances in

the literature suggesting that preschoolers can select a conclusive test of a hypothesis (Koerber & Osterhaus, 2019; Piekny & Maehler, 2013) and have some understanding of Control of Variables Strategy (van der Graaf et al., 2015) even without any instruction or support. One potential limitation of the present tasks is that the correct choice is also the one that most resembles the stick children have seen placed on the box. Thus, children may be able to solve the CVS tasks simply by selecting the stick most perceptually similar to the test stick. Other researchers have tackled this issue of perceptual similarity with size, shape, and color and found that children did not solve their CVS tasks on the basis of perceptual similarity (Walker, Goel, Nyhout, & Ganea, 2019). This finding makes us optimistic that this was also not the case in our tasks; however, future research could address this issue by introducing another test choice with perceptual similarity equal to that of the controlled test. Further, the results of Study 4, that the CVS tasks are not only related to inhibition, but also to Theory of Mind, suggests that it is not solely perceptual similarity driving the performance on recognizing a controlled test.

We found an unexpected pattern in Study 2a that children performed better on the 3-variable CVS task than on the 2-variable CVS task. In Study 3, adults showed a similar pattern, performing better on the 3-variable CVS task than on the 2-variable CVS task in selecting the controlled test, justifying the test, and interpreting the outcome. Further, age was negatively related to adult's justifications in the 2-variable CVS task. These findings seem to suggest that the 2-variable task is not simply an easier version of the 3-variable task. Instead, it seems to sometimes confuse both children and adults, and thus, may not be measuring scientific reasoning in the way we expect, if at all.

The ability to justify one's selection of a controlled test was developing from three to six years, with older children better able to provide relevant justifications for their selection of a controlled test. Further, children were more likely to generate relevant

justifications for controlled than for confounded tests. This could suggest that children indeed have a more explicit understanding of CVS, and this results in their ability to both choose the correct test and provide an explanation for it. It could also suggest, however, that if children are driven to make the correct selection based on perceptual similarity as described above, they also provide justifications referring to perceptual similarity, which also happens to be what we consider a relevant justification by referring to control variables. There is not a good way to control for this, since relevant justifications would require such comparisons. To illustrate, when we look at the justifications adults provided for their test selection, they were very similar to those children provided, for example, “because it has the same top and bottom Legos.”

5.2.1.2 Design of knowledge-lean CVS tasks

We designed the CVS tasks to be appropriate for preschoolers in a number of ways. First, by designing a novel knowledge-lean task, we reduced the potential influence of prior knowledge or beliefs about the task content, which have been shown to affect children’s reasoning (e.g., Kuhn et al., 1988). Children should have no preconceptions about what makes the box light up. We also chose to design the CVS tasks as a recognition or selection task, so that children had to recognize a controlled test of the hypothesis from a number of options presented to them, rather than to design a controlled test themselves, as selection has been shown to be easier than production (e.g., Bullock & Ziegler, 1999). Further, when designing the choices of a test, we chose not to include the hypothesized cause, the X brick, to avoid that children attempt to reproduce the positive effect of the box lighting up simply by choosing a test which included X brick, since children have been shown to prefer to produce effects rather than test hypotheses (e.g., Tschirgi, 1980). Finally, we limited the number of variables to two or three in the CVS

tasks to minimize the cognitive load associated with increased information processing of additional variables (e.g., Tschirgi, 1980).

However, though these were all decisions we made in designing the task to be appropriate, one must also consider what those decisions mean for the results. First, in choosing to use a knowledge-lean task, we re-enter the discussion about scientific reasoning as a domain-general or domain-specific ability. As presented at the beginning of this thesis, some research supports the claim that scientific reasoning, and particularly the control of variables strategy, is a domain-general process skill that can be learned in any domain and transferred to other domains (e.g., Daxenberger et al., 2018; Inhelder & Piaget, 1958; Kuhn, 2002). However, in designing a knowledge-lean task to specifically avoid any prior knowledge, we have to acknowledge that the content of scientific reasoning tasks does indeed affect performance and that having prior knowledge or prior beliefs about task content could either hinder or support performance. That being said, to make a scientific reasoning task appropriate for preschool children, we believe the use of a knowledge-lean task is justified to assess underlying, unadulterated scientific reasoning abilities in a way that is analogous to how causal reasoning is often investigated. Further, since we believe these abilities are domain-general, we could additionally investigate differences in performance on corresponding knowledge-rich contexts to more precisely determine the effect of task content on reasoning abilities. Finally, it would be important to investigate whether this developing knowledge-lean scientific reasoning ability is relevant for scientific reasoning in meaningful real-world contexts.

Design of CVS selection tasks. In using a selection task, we have chosen a more lenient measure of scientific reasoning ability. However, we believe this is appropriate to determine a baseline of beginning scientific reasoning ability in preschool at perhaps the earliest detectable level. Similarly, in providing choices for a test of a hypothesis that do

not include the hypothesized cause, we take the first step of introducing a contrastive test for the child and eliminate the possibility to attempt to produce the positive effect by testing the hypothesized variable. This reduces the level of difficulty but does not allow us to assess children's ability when they have to choose between testing a hypothesis or producing an effect. Further, the fact that the effect that occurs, the box lighting up, could be considered "positive" could also present an issue. Particularly, children were disappointed when their choice did not make the box light up, even though that was the ideal outcome to generate a conclusive test. To avoid this reaction and use the desire to produce an effect to our advantage, we could reverse the effects of the bricks to be inhibitory and, instead, ask children to find out if a particular brick was stopping the box from lighting up.

5.2.1.3 Interpretation of Confounded Evidence

To be successful in the ICE task, after children observe confounded evidence of a light effect, they must correctly claim that they cannot know which bricks caused the effect. They can additionally explain why, for example, because the bricks are stuck together, or they did not see them individually. The understanding of one's own ignorance as a result of confounded evidence in the ICE task was more difficult than the selection of a controlled test. Further, unlike the CVS tasks in any of the studies, performance on the ICE task was related to age in Study 2b, which represented the most robust measure of the ability to recognize what one does not know. Thus, as children undergo cognitive development throughout the preschool years, they become more able to recognize when they can or cannot know something. This supports recent research on children's ability to handle uncertainty when making causal inferences, which found that this ability was developing between the ages of four and seven, with only older children showing above chance-level performance (Sobel et al., 2017). The ability to reflect on one's own

knowledge is likely related to metacognitive abilities, which are also undergoing development around this same time (e.g., Rohwer et al., 2012). The findings of the ICE task also support recent findings showing that a majority of five- to six-year-olds could correctly claim a lack of knowledge as a result of confounded evidence (Köksal-Tuncer et al., 2019). When we narrowed our sample to that age range as well, we also found that a majority of five- to six-year-olds could correctly claim a lack of knowledge as a result of confounded evidence in the ICE task. Together, these results seem to suggest that by the age of five, children have an appreciation for the unformativeness of confounded evidence.

5.2.2 Structure and correlates of scientific reasoning

The fact that the ICE task showed a developmental trend while the CVS tasks did not, as well as the lack of effect of ICE on performance on the CVS tasks, may suggest that these abilities are separate components of scientific reasoning or that the abilities needed for ICE may be developing naturally, while the abilities needed for CVS may not. In Study 4, we found that the ICE tasks and the 3-variable CVS tasks were, in fact, related, but they also correlated with different cognitive abilities. The ICE tasks were related to intelligence, planning, and Theory of Mind, while the 3-variable CVS tasks were related to inhibition and Theory of Mind. These results would suggest that ICE and CVS are separate but related components of scientific reasoning.

When we looked at the tasks together as a general measure of scientific reasoning, we found that scientific reasoning was related to both inhibition and Theory of Mind. The relation between inhibition and scientific reasoning is in line with previous findings of the same relation in a number of other studies (Bauer & Booth, 2019; Kwon & Lawson, 2000; Osterhaus et al., 2017; van der Graaf et al., 2016, 2018). The positive relation with inhibition makes sense both in terms of the ICE task and the CVS tasks. In the ICE task,

children have to inhibit a natural bias to answer in the affirmative in order to correctly claim that they cannot know something. In the CVS tasks, children have to inhibit the incorrect choices in order to select the correct choice representing a controlled test of a hypothesis. It is possible, then, that children could fail the CVS tasks, not because they do not have an understanding of CVS, but because they possess a poor level of inhibitory control.

The positive relation with Theory of Mind is also in line with the literature showing that false belief understanding and (advanced) Theory of Mind abilities are related to several aspects of scientific reasoning, for example, experimentation, understanding evidence, and justifications (Astington et al., 2002; Klein, 1998; Koerber & Osterhaus, 2019; Osterhaus et al., 2017; Piekny et al., 2013; Sodian et al., 2016). In the ICE task, children need a metacognitive understanding of their own ignorance and to distinguish between their beliefs and the evidence they observed. In the CVS tasks, children need an understanding of alternative possibilities, i.e., the box could light up or not light up when their choice is placed on it. They also have to distinguish between the hypothesis (that one particular brick makes the box light up) and the evidence that different experimental design choices would produce. Further, we found evidence suggesting that mastery of Theory of Mind appears to be necessary for mastery of scientific reasoning and not the reverse, as about half of children showed mastery of Theory of Mind but not scientific reasoning, while 2% of children showed mastery of scientific reasoning but not mastery of Theory of Mind.

Scientific reasoning was not related to intelligence or language. Previous studies have found a relation between scientific reasoning and intelligence (Bullock et al., 2009; Haslbeck et al., 2018; Koerber et al., 2015; Koerber & Osterhaus, 2019; Mayer et al., 2014; Piekny et al., 2013). The lack of a relation to intelligence may be explained by the

intelligence measure used. In the present study, we used a subscale that measured children's general knowledge. The scientific reasoning tasks in the present study are knowledge-lean, thus, it may be expected that they do not relate to a measure of intelligence that is knowledge-rich. The studies that found a relation between intelligence and scientific reasoning used every day or scientific content tasks, for example, the plane task (Bullock & Ziegler, 1999). Our finding of no relation to a knowledge-rich intelligence measure would further suggest that our tasks are truly knowledge-lean.

The lack of a relation between scientific reasoning and language is somewhat surprising considering the literature showing that a number of different language measures have been shown to be related to scientific reasoning, including reading comprehension (Koerber et al., 2015; Mayer et al., 2014; Osterhaus et al., 2017; Wagensveld et al., 2015), language receptiveness (Koerber & Osterhaus, 2019), vocabulary (van der Graaf et al., 2018; Wagensveld et al., 2015), and grammatical abilities (van der Graaf et al., 2016, 2018). We used a language measure that assessed children's grammar, specifically their understanding of morphological rules, which required children to produce the plurals of real and made-up nouns. This language measure was also related to all other tasks, except the Content False Belief task. The finding of no relation between language and scientific reasoning may suggest that the language of the questions in the scientific reasoning tasks was not so difficult that it required extensive language processing.

However, this finding may also suggest that children did not need to rely on their language abilities or the understanding of the questions to correctly solve at least the CVS tasks. Instead, children may have been able to solve the CVS tasks using a lower level perceptual similarity matching strategy as discussed previously, though the relation to Theory of Mind would suggest that they were not only relying on perceptual matching, but also possibly on an understanding of the relation between the hypothesis and the evidence

their test would produce. Moreover, the fact that the scientific reasoning tasks did not relate to language despite the requirement for children to explain or justify can be discussed in terms of the frequency with which children provided such justifications. For example, in the CVS tasks, only 11% of all verbal responses were considered relevant to CVS. Many children simply responded with “I don’t know” or not at all. Thus, another explanation for the lack of relation between children’s scientific reasoning scores and language could be in the fact that the scientific reasoning scores did not include much in terms of language production.

Finally, scientific reasoning was not related to working memory, planning, or cognitive flexibility. Previous research has also found no relation to cognitive flexibility (van der Graaf et al., 2016, 2018), and the present scientific reasoning tasks did not really require children to engaging in shifting. The relation between working memory and scientific reasoning is inconclusive in the literature, with some studies finding a relation and some not, and particularly when there is a distinction between verbal and non-verbal working memory (Piekny et al., 2013; van der Graaf et al., 2016; 2018). However, van der Graaf and colleagues found that verbal working memory was related to scientific reasoning and our working memory task, the Backwards Digit Span, is also a measure of verbal working memory. It could be that the tasks were not so difficult that they required much from working memory or that the Backwards Digit Span was particularly difficult and, thus, did not result in much variability.

Finally, the literature did not present any findings regarding a relation between planning and scientific reasoning, and we also found no relation, but it would seem that planning could be important for scientific reasoning. Planning was related to the ICE measures suggesting perhaps that the ability to think ahead or organize one’s thoughts

supported children's ability to correctly recognize that they could not reach a conclusion on the basis of confounded evidence.

Study 4 found an effect of gender on the scientific reasoning tasks, specifically on the first trial of the 2-variable task, in which girls were more likely to select the correct choice than boys. This finding was surprising considering we found no effects of gender in Studies 1, 2a, or 2b. Further, most studies have shown that gender is not related to performance on scientific reasoning measures (Astington et al., 2002; Bauer & Booth, 2019; Bullock et al., 2009; Koerber et al., 2015; Sodian et al., 2016; van der Graaf et al., 2018), and studies that have found gender differences generally show that boys perform better on measures of science achievement and reasoning throughout school (J.-T. Kuhn & Holling, 2009; Saçkes et al., 2011). The effect of gender in our tasks may be unrelated to children's scientific reasoning abilities. Taking a closer look at Trial 1 (Figure 2.4), it is possible that children's (specifically girls') color preferences affected their responses: the correct choice included a purple block. That being said, we attempted to design the materials in a way that would not influence the choice of a correct or incorrect response, and in Studies 2a and 2b in which these task materials were counterbalanced there was no influence of specific task materials on children's performance. Interestingly, the effect of gender on the ICE tasks was approaching significance ($p = .07$) in the same direction, such that girls performed better. This would suggest that perhaps the effect of gender is something more than just an effect of task materials.

5.2.3 Adults' scientific reasoning abilities

5.2.3.1 Initial responses

Adults performed very well in their initial responses to all three scientific reasoning tasks: recognizing their ignorance as a result of confounded evidence and selecting controlled tests of a hypothesis with two and three variables. All adults were able to correctly claim

that they could not know which of the four bricks in the stick made the box light up. This level of performance is impressive, considering previous findings showing that most adults were fairly certain they knew something even when they could not on the basis of evidence (Kuhn, 2007a). In that case, though, adults were asked to make judgements about what factors had an effect on ticket sales. Thus, their previous knowledge or beliefs about the task content likely affected how certain they felt about reaching a conclusion. These two findings provide support for the theory that the task content as well as prior knowledge or beliefs about the task content can affect performance. When those factors are removed, here in the form of knowledge-lean tasks, adults can recognize the state of their knowledge based on the evidence they observe.

In the CVS tasks, though adults performed well, they did not show full competence even in these simple knowledge-lean tasks. Thus, not all adults have developed stable and robust scientific reasoning abilities. That being said, overall performance was better (~94%) compared to previous findings that around 80% of adults could select a controlled test of a hypothesis (Bullock & Ziegler, 1999), again suggesting that the knowledge-leanness of the present scientific reasoning tasks better supported adults' ability compared to knowledge-rich tasks. Finally, approximately 93% of adults were able to correctly interpret the outcome of their test, either claiming that the X brick was a lighter, in the case of a controlled test, or claiming that they could not know if the X brick was a lighter, in the case of a confounded test, showing that a majority of adults were able to make sense of the evidence they both observed and produced, and relate this to the hypothesis. This performance is better than previous findings showing that anywhere between 20% - 70% of adults were able to provide valid inferences in knowledge-rich tasks (Kuhn et al., 1995; Schauble, 1996). However, not all adults were fully competent in this assessment, showing again that not all adults have developed robust scientific reasoning abilities.

5.2.3.2 *Explanations, justifications, and interpretations*

Despite good performance in the initial responses, adults struggled to provide elaborated explanations for why they could not know, in the case of the ICE task, and for why they selected a particular test, in the case of the CVS tasks. In the ICE task, 85% of adults could provide an explanation for their ignorance of the cause of the light effect, but the remaining adults either could not provide an explanation or, instead, provided an incorrect statement about the bricks. This finding was especially surprising considering the protocol for this task: we showed participants the stick of bricks and told them that they were stuck together and could not be taken apart, while pulling on them to emphasize this fact. Thus, we had provided participants with one of the potential answers to receive a point on this task: that the bricks could not be isolated, so it is surprising that they could not generate this response. When looking at the incorrect responses, some participants claimed there must be two lighters, which was based on the familiarization when two lighters were combined but was not necessarily the case. They could know there was at least one lighter present, but not if there were more than one. This example illustrates that adults attempt to use prior knowledge, which they had just learned, to make sense of new information, instead of interpreting the new information independently. In real-world contexts, it, of course, often makes sense to use prior experiences or knowledge to deal with new information, but individuals must be aware of the risk of not considering new information independently before trying to relate it to existing knowledge.

In the CVS tasks, adults had two opportunities to provide responses to explain their initial responses: they should explain their choice of a test and their interpretation of the outcome of the experiment. In the 2-variable task 65% of adults received at least one point for their justification, meaning that 35% could not provide any valid justification for why they selected a particular test. Performance was better in the 3-variable task, with

almost all adults (97.5%) able to provide at least one statement providing a valid justification for their test. It is surprising that adults struggled so much to provide a valid reason for their test selection in the case of the 2-variable task. This finding suggests that perhaps they are over-thinking the task, or that the 2-variable version, which we assumed would be easier, was, in fact, more confusing. This result also provides valuable information regarding the use of the task with children, that if the 2-variable task is confusing for adults may also be confusing for children, and thus, perhaps the 3-variable task is a better assessment of control of variables abilities. Looking at the quality of the justifications, adults received relatively few points for their justifications showing that they did not elaborate or try to explain in a way that referred to all of the considerations they could have made in selecting their test. For example, they could have referred to the focal or control variables, to potential effects, and to the potential resulting inferences. Participants mainly referred to the control variables, followed by the absence of the focal variable, followed by the possibility that the box does not light up, and what that would mean for the hypothesis. Participants also provided higher quality justifications in the 3-variable compared to the 2-variable task, again suggesting that the 2-variable task was more difficult or more confusing.

Finally, looking at adults' interpretations of the outcome of their experiment, they were more able to explain their conclusion than to explain their test selection. This is to be expected considering that, for a test selection, one must reason about potential outcomes to make a selection, but for an interpretation, one has already seen the evidence and only has to reason about reality and not hypothetical outcomes. This finding is relevant for using the tasks with children as well, such that they may also be better able to interpret or explain the outcome than why they selected a test. This hypothesis would also be supported by the causal reasoning literature showing that children can make inferences on

the basis of covariation information (e.g., Gopnik et al., 2001; Schulz & Gopnik, 2004; Sobel & Kirkham, 2006). Moreover, participants again provided higher quality interpretations in the 3-variable compared to the 2-variable task, providing further evidence that the 2-variable task was more difficult or more confusing.

5.2.3.3 Other factors affecting performance

The finding that all three tasks were related to age, but that the ICE task and the 3-variable CVS task were positively related while the 2-variable CVS task was negatively related to age could speak further to the theory that the 2-variable task was confusing. It is also interesting that the ICE task and the 3-variable CVS task were related to age but not to the participants' level of education, suggesting that the ability to explain and justify is still developing with age but not as a result of continued education. Previous findings have shown that college-educated adults perform better on scientific reasoning tasks than non-college-educated adults (Amsel & Brock, 1996; Kuhn, 2007a), however, the present results suggest that even within a sample of college-educated adults, there is still variation in ability related to age. In addition, the relation between the ICE task and the 3-variable task replicates the finding from Study 4 with preschoolers that those two tasks were related.

Finally, a few additional findings warrant discussion. First, that a number of participants initially selected an incorrect test but then self-corrected when trying to explain why they had selected that test. Thus, the process of (self-)explanation was beneficial to recognizing a controlled test of a hypothesis, a robust finding in the literature (e.g., Chi, Glaser, Reimann, Lewis, & Bassok, 2005; Rittle-Johnson et al., 2008). Second, that adults stated a number of assumptions they had to make, for example, that the same colors have the same effects and that the sticks of three perform in the same manner as the sticks of two or four. The fact that adults felt they needed to state these assumptions

suggests that they should instead be made clear by the experimenter or by the evidence participants observe in familiarization. Further, we cannot know if children make these same assumptions, thus, making them clear would also be beneficial for children. Finally, a few adults forgot that the test stick placed on the box prior to the hypothesis had made the box light up. This observation was surprising, and also critical, as this information is essential for choosing a test and for reaching a conclusion after the outcome of the experiment. Again, this could have implications for studies with children, that perhaps they also forget this initial effect, but do not verbalize that they have forgotten. This issue could be addressed by helping participants to keep track of what has or has not made the box light up.

5.2.4 Developing training tools for preschoolers

The literature on promotion of scientific reasoning and specifically CVS abilities has shown that training can be effective across all ages, with different methods of instruction and different mediums (Schwichow et al., 2016). We wanted to investigate if this robust effect is also found in younger children. We developed digital materials in the form of a tablet application and a video tutorial to teach scientific reasoning. Because the use of technology and the consumption of digital content is becoming more and more common among young children, it would be appropriate to have research-based educational content available to parents and educators in a form that is accessible and easy to use. We considered cognitive architecture, instructional design, and multimedia learning, as well as storytelling and animation in designing our materials (Bétrancourt, 2005; R. E. Mayer, 2014; Robin, 2008; Sweller, 1988; Vygotsky, 1980). We took an iterative approach in developing the application and engaged children as testers of the application, gathering feedback on usability through observation, and implementing this feedback in subsequent iterations.

This process of iterative development allowed us to reduce overall usability issues. Limiting usability issues is important to ensure that the tools we develop are measuring the cognitive abilities in which we are interested, and that those abilities are not masked by difficulties arising from the tasks or the task materials. Thus, it is critical that the usability of products, whether as research tools or educational tools, whether digital or paper-and-pencil, is optimized to ensure that the data obtained from them is both meaningful and accurate.

During the iterative process of designing both the tablet applications and the video tutorial, we considered principles of instructional design and multimedia design to guide our decisions. For example, we attempted to limit and reduce extraneous cognitive load (Sweller, 1988) due to too much information being present on the screen at once, in the case of multiple sets of variables for selection in the application. We additionally used some principles of multimedia learning for design, as we presented information both visually on the screens and verbally in instructions and explanations (R. E. Mayer, 2002). For example, we used the *spatial and temporal contiguity principles* to structure the presentation of verbal information at the same time as the visual information is present and to present related items grouped with each other. We used the *coherence principle* to guide the exclusion of extraneous material. We did not present any text, both because preschoolers are just beginning to learn how to read and because the *modality principle* suggests that verbal information should be presented as narration rather than text to take advantage of both modalities. We allowed children the opportunity to learn and practice the required actions in advance in the applications (*pre-training principle*). We wrote the scripts to be conversational in style (*personalization principle*). Finally, we used the *segmenting principle* (R. E. Mayer & Pilegard, 2005) to guide our structuring of the information into digestible chunks. In addition to these principles of cognitive architecture

and multimedia learning, we also embraced the concept of scaffolding to support children in learning the Control of Variables Strategy (Vygotsky, 1980). Based on the findings of precocious causal reasoning abilities and beginning abilities in scientific reasoning, learning CVS should be within children's Zone of Proximal Development.

In addition to considering the above design principles, we wanted to design our tasks to be engaging so that children would be more likely to benefit from the instruction (Lieberman et al., 2009). The interaction with the applications seemed to be engaging for children. For example, in Study 5b children showed more indications of interest than of boredom and when asked if they wanted to complete a second test trial, no children declined. This indicated either that children were enjoying the interaction or that they felt they had to continue, though the purpose of the warm-up was to get children comfortable with the experimenter and the testing situation and to inform them that they could stop at any time. There is, of course, the danger of crossing the line from engagement to distraction. In Study 5a there were some cases of impatience or attempts to select something that was not yet selectable. This impatience could be either due to boredom or to over-excitement to continue interacting with the application without paying attention to the requirements of the task.

Furthermore, it is important that the tasks are educational rather than just fun. Basing their design both on research about design principles and research about cognitive development should improve the likelihood that the tasks are, in fact, educational. One criterion for evaluating products that claim to be educational was to assess if the tasks are about winning or about learning (McManis & Gunnewig, 2012). The CVS tasks in the applications place an emphasis on finding something out and conducting good experiments, however, the tutorial conflated the characters' desire to win the race with the desire to produce a good experiment. This framing could be adjusted to place more

emphasis on the experimentation processes. However, it would be important to ensure that the resulting version is still as captivating as watching two characters compete to win a race. In the case of Study 6 and the development of the video tutorial, we investigated the effect of animation on both engagement and learning. Animations are considered motivating and have been shown to positively affect learning (Bétrancourt, 2005; Höffler & Leutner, 2007). We found that children who watched an animated tutorial were more likely to be categorized as engaged and also more likely to show improvement in scientific reasoning tasks compared to children who watched a static tutorial.

Finally, when designing and evaluating products for children it is important to include children in both the design and evaluation process. We enlisted children as *testers* (Druin, 1999) for the applications to test them in an early stage of development when observations and feedback can still be implemented into the final product. We mainly used observation as our evaluation tool given the young age of our participants. As we have seen throughout all of the studies presented here, preschoolers have difficulty verbalizing their explanations and justifications in a robust way. This also makes it difficult to use evaluations that require children to verbalize, such as the Think Aloud method (Nielsen, 1994). Though other evaluation tools also exist, they present different issues when used with children. For example, the Fun Toolkit (Read et al., 2002) includes a Smileyometer in which children can select from a set of smileys to indicate how they enjoyed their interaction with a product. This tool presents two potential issues, one, that children are biased to select the more positive responses, and two, that children may not understand that this tool is meant to evaluate the produce and instead simply pick whichever smileys they like best. For these reasons, we video recorded children's interactions and were able to code their actions or reactions as measures of usability and engagement.

5.2.5 Assessing and promoting CVS in preschoolers with digital tools

The findings of the studies in Chapter 4 showed that, when preschool children are asked to produce a controlled experiment rather than select a controlled test, their success is limited. These findings are in line with research showing that production tasks are more difficult than selection tasks (Bullock et al., 1999). With the tablet applications a majority of children were able to produce contrastive tests that varied the focal variable. This finding is consistent with the literature showing that contrastive testing is easier than controlled testing (Bullock et al., 1999). However, many children also varied all of the variables, producing two completely different tests. This behavior may be a result of wanting to try everything or wanting to be “fair” in using all the variables, a finding that is also in line with the literature showing that elementary school children struggle to produce controlled tests (Bullock et al., 1999).

Few children designed a controlled test after instruction in Study 5a, but in Study 5b half of the children could produce a controlled test after two instances of instruction, which is in line with other findings of preschool children’s abilities in CVS after feedback (van der Graaf et al., 2015). Further, it suggests that there is potential in using direct instruction to promote CVS abilities in preschoolers. In the test trial of Study 5a, half of the children showed persistence from the training trial to the third test trial, carrying over the specific instructions regarding varying a particular variable but not generalizing this to the new hypothesis. This would seem to indicate that children understood the instructions but did not understand that they needed to adapt their strategy to the new hypothesis. This result indicates the importance of choosing an appropriate transfer task. For example, if we had provided a new task with the same structure but different variables, we could have avoided that children carry over the instructions regarding a specific variable.

When comparing the materials of the task in Study 5a, either a paper-based interactive storybook or a tablet application, we found that, after instruction, three children in the storybook group were able to design a controlled experiment in the third trial while none of the children in the application group were able to do so. This could be a result of children not being comfortable interacting with the tablet or because the application was distracting rather than engaging during the training and/or in the last trial. Thus, training with the storybook was more effective than training with the application and training with the application had no effect on performance. Because of these differences both in the assessment and training, the final iteration of the application from Study 5b should be compared to a non-digital assessment.

Finally, in Study 6 we found that children benefited from direct instruction in the form of a video tutorial, particularly when the tutorial included animations designed to draw attention to critical content. Children who watched the animated tutorial were categorized as more engaged and also showed some improvement on the knowledge-lean scientific reasoning tasks, while children who watched the static tutorial were less engaged and showed no improvement and greater deterioration on the scientific reasoning tasks. In the ICE tasks, performance was stable for children in the animated tutorial condition but deteriorated in the static tutorial condition. In the 2-variable CVS tasks children who watched the animated tutorial improved from chance level to above chance level similarly to the improvement in performance in Study 1, but within one session rather than over two weeks. Children's performance on the 3-variable tasks was also stable. Children who watched the static tutorial showed deteriorated performance in the second trial in both the 2- and 3-variable tasks. These results would suggest a benefit of animation for at least maintaining if not improving performance, which is line with the literature showing that animated content is motivating and can positively affect learning (Bétrancourt, 2005;

Höffler & Leutner, 2007). However, non-animated content appeared to negatively affect performance. Furthermore, it is interesting that a tutorial with knowledge-rich material would affect performance on a knowledge-lean task. This could speak to the domain-generalty of the Control of Variables Strategy and scientific reasoning (Klahr, Zimmerman, & Jirout, 2011; Kuhn, 2002).

The results of Studies 5a, 5b, and 6 suggest that there is potential in promoting scientific reasoning abilities in preschool children and highlight the importance of designing tools to be developmentally appropriate, engaging, and accurate. These results are, however, limited by the small sample sizes of these studies. While they represent initial investigations, the findings reported here need to be further investigated with larger sample sizes to ensure that the sample is representative and that effects are not a result of sampling error.

5.3 Implications

5.3.1 Theoretical implications

The finding of early scientific reasoning abilities in preschoolers using a knowledge-lean task and a paradigm commonly used to investigate causal reasoning can serve as a bridge between these two literatures. On the one hand the scientific reasoning literature has shown that young children have limited abilities in scientific reasoning (e.g., Kuhn et al., 1988), even if those abilities are more developed than previously thought (e.g., Zimmerman, 2007). On the other hand, the causal reasoning literature shows that very young children and infants show precocious abilities in determining cause-effect relations and performing informative interventions (e.g., Gopnik et al., 2001; Schulz & Gopnik, 2004; Cook et al., 2011), abilities which may serve as building blocks for later scientific reasoning. The proposed distinction between these abilities is the development of a metaconceptual understanding of theories, beliefs, and hypotheses as different from

evidence, and the metacognitive ability of intentionally seeking knowledge through actions, such as experimentation (Kuhn, 2002).

The knowledge-lean scientific reasoning tasks presented in this thesis required children to metacognitively reflect on their own state of knowledge and to test a hypothesis by selecting a controlled test. Thus, the finding of nascent abilities in these tasks and the relation of these tasks to Theory of Mind, suggests a developmental progression starting with basic causal reasoning abilities in infancy and early childhood, followed by beginning scientific reasoning abilities in preschool coinciding with the development of Theory of Mind. This progression then continues with the development into more mature scientific reasoning abilities in late childhood and continues throughout schooling and into adulthood, though not all adults reach full competence.

Further, it is important to consider why children and adults have been shown to struggle with scientific reasoning and to take a closer look at the tasks with which we assess scientific reasoning, especially with children. Specifically, we should consider the task content and whether individuals have prior knowledge or beliefs about it, the difficulty of the task in terms of having to recognize or produce valid tests, and the outcomes of experimentation. This is not to say that the finding that children and adults struggle under certain circumstances is not valuable information about their reasoning processes. But, if we consider that children do have basic abilities, then we can look at these struggles not as an indicator that they are not capable of scientific reasoning, but instead as an opportunity to build upon the basic scientific reasoning abilities shown here, to ensure that children develop mature scientific reasoning abilities across a broad range of constraints.

5.3.2 Practical implications

Practically, the finding of early scientific reasoning abilities in preschool, and the finding of some effect of training tools for CVS with preschoolers, suggest that it may be possible to begin promoting these abilities earlier in education than is currently the case. Especially considering that not all adults currently reach full competence in scientific reasoning (Bullock & Ziegler, 1999; Kuhn et al., 1988; Tschirgi, 1980), and considering how important scientific reasoning is to actively participating in society and helping to shape its future (Bromme & Goldman, 2014; Trilling & Fadel, 2009), more emphasis on promoting these abilities early on and throughout education could result in more mature scientific reasoners in society.

The relation of scientific reasoning to Theory of Mind and the finding that Theory of Mind seems to be necessary (but not sufficient) for scientific reasoning suggests that to help promote scientific reasoning abilities, we should also focus on promoting metacognitive abilities (e.g., Amsel et al., 2008; Kuhn et al., 2008). The robust findings of the effectiveness of teaching the Control of Variables Strategy (Schwichow et al., 2016) and the tentative trend toward positive effects of training in preschoolers with a tablet application and a video tutorial suggest that preschool children could benefit from exposure to instruction in scientific reasoning. Further, with robust tools that can be contained within a tablet, such training could occur in preschool or daycare settings, as well as in the home or other informal learning settings, as children already spend time with tablets or watching educational videos (Rideout, 2017; A. Smith et al., 2018).

5.4 Future Directions

The findings of the present studies present many opportunities for future research. To begin with the findings of children's beginning abilities in scientific reasoning as measured by knowledge-lean tasks assessing recognition of controlled tests, an initial next

step would be to investigate if success in the tasks as shown here is a result of a perceptual matching strategy. Ideally, further research would replicate the results of Walker and colleagues (2019) that children do not solve control of variables tasks on the basis of perceptual similarity. This could be accomplished by adapting the task to include an additional choice with the same perceptual similarity to the test stick as the correct, controlled test. Specifically, we could create a test stick with four bricks, XYZO, which makes the box light up and in which X is again the hypothesized cause. We could then show children a second stick ABCO, which does not make the box light up. From these first observations, children should be able to conclude that O is not a lighter. Then we could provide children with choices to test the hypothesis that X is a lighter. First, JYZ, which varies the X brick and keeps two of the original bricks constant (Y & Z) and does not make the box light up. Because children know that O is not a lighter, they could conclude that X is a lighter. For the other option to be equally perceptually similar but not able to conclusively test the hypothesis, it should include JZO and also not make the box light up. With this option, children cannot know whether it was X or Y that made the box light up originally. With this procedure, if children simply select a test based on perceptual similarity, there should be no difference between which test they select, though one is a conclusive test and one is not.

Additional adaptations to the current scientific reasoning tasks could provide opportunities to broaden our understanding of early scientific reasoning abilities. For example, children could be asked to generate their own hypothesis about what makes the box light up before selecting a test of that hypothesis. In the current tasks, we always presented children with a hypothesis. On the one hand this likely makes the task easier by removing the additional effort of generating a hypothesis, but on the other hand it may have reduced children's motivation to test that hypothesis since it was not their own.

Indeed, a number of children spontaneously generated their own hypotheses after seeing the test stick make the box light up. Thus, it is also possible that children ignored the hypothesis that we presented them and instead tested their own. As some children have no difficulty in saying which brick they think made the box light up, we could have them generate their own hypothesis to test. With this procedure, we could investigate whether children are more successful testing their own or other's hypotheses and could investigate another aspect of scientific reasoning, hypothesis generation.

Further, instead of having children select a test of a hypothesis from already existing options, we could ask children to produce their own tests of a hypothesis. We could teach children that the box only works with sticks of three bricks and after showing them the light effect of the test stick, we could present them with choices of individual bricks that they would have to combine into a stick of three to produce a controlled test. This would vary the difficulty of the task and distinguish between selecting and producing a controlled test.

As discussed previously, the protocol and materials for the scientific reasoning tasks underwent a few iterations from Study 1 to 2b, thus, it would also be important to conduct another assessment of stability on these tasks with the final, improved versions of the tasks from Study 2b. In Study 1 we saw improvement from Session 1 to Session 2 without any instruction. This type of improvement is in line with studies showing that repeated interactions with systems for designing experiments improved children's performance in designing controlled tests (Kuhn, 2007b; Schauble, 1990). However, the initial performance in Session 1 was lower than in the other studies, possibly due to the materials or wording of the questions.

Specifically, in the improved versions we changed the wording of the test selection and justification questions from "Which stick do you want to pick and why?" to "Which

stick is the best to find out if the X brick makes the box light up and why is that the best stick to find that out?” to place an emphasis on the testing of a hypothesis, and that they should choose in relation to the hypothesis rather than based on preference. We also improved the interpretation question from “Now do you know if the X brick makes the box light up?” to “Is the X brick a lighter, not a lighter, or can you not know?” because responses to the first version of the question were unclear, as children sometimes responded only yes, but then added an explanation that implied the X brick was not a lighter. Thus, we could not be sure if their response was indicative of their knowledge (Yes, I know; No, I don’t know) or of the brick’s category (Yes, it’s a lighter; No, it’s not a lighter). Thus, it would be interesting to observe if the same type of improvement between sessions is seen with the improved tasks.

In addition, Study 4 presented data from the first measurement point of a longitudinal study and revealed the relations between scientific reasoning and inhibition and Theory of Mind. The subsequent development of scientific reasoning over time within individual children from four to five-and-a-half years will further illuminate how these abilities develop and relate to other cognitive factors. Moreover, Study 4 only presented a small portion of the many cognitive measures assessed in this first measurement point. Thus, further analysis of these additional data will provide an even more detailed picture of the structure and correlates of scientific reasoning in early childhood, for example, in relation to causal reasoning, as well as other scientific reasoning, language, and intelligence measures.

Finally, the studies in Chapter 4 represented only initial steps in assessing the possibility to promote CVS abilities in preschool children. The findings showing potential for success in training should encourage future research to further explore this possibility on a larger scale, taking into account the design and usability recommendations for

improving the existing tasks. For example, future research could assess the effect of instruction with the tablet application using a transfer task with different variables. Additionally, audio instructions could be added into the application to standardize the instructional material across all children and to investigate if receiving instruction from a person or from the tablet affects children's acquisition of CVS. The present studies represent repeated measures designs in which we compare performance in a pre-test to performance in a post-test. Future research could compare training with the application or tutorial to a non-training control group. For example, in the case of the video tutorial, there is an endless supply of video content claiming to be educational that could be used as an alternative, such that children still watch something, but do not receive explicit CVS training. Lastly, as we saw that a knowledge-rich tutorial had an impact on performance on knowledge-lean tasks, it could be interesting to investigate the opposite direction, if training on knowledge-lean tasks impacts performance on knowledge-rich tasks.

5.5 Conclusion

The aim of this thesis was to explore scientific reasoning abilities in preschool children, the relation between those abilities and other cognitive factors, and the potential for training scientific reasoning in preschool. As a result, we can conclude that preschool children possess nascent scientific reasoning abilities, specifically a metacognitive understanding of the state of their knowledge and a recognition of the inconclusiveness of confounded evidence, as well as a recognition of a controlled test of a hypothesis. We can further suggest a developmental progression from basic abilities in causal reasoning in infancy and early childhood, to basic abilities in scientific reasoning coinciding with the development of Theory of Mind, to later mature scientific reasoning in late childhood and into adulthood. Finally, we can tentatively propose the potential to promote scientific reasoning abilities in preschool using direct instruction with digital tools.

References

- American Association for the Advancement of Science (AAAS). (2009). Conference Homepage. *Vision and Change in Undergraduate Biology Education: A View for the 21st Century*
- Amsel, E., & Brock, S. (1996). The development of evidence evaluation skills. *Cognitive Development, 11*(4), 523–550.
- Amsel, E., Klaczynski, P. A., Johnston, A., Bench, S., Close, J., Sadler, E., & Walker, R. (2008). A dual-process account of the development of scientific reasoning: The nature and development of metacognitive intercession skills. *Cognitive Development, 23*(4), 452–471.
- Anand, S., & Krosnick, J. A. (2005). Demographic Predictors of Media Use Among Infants, Toddlers, and Preschoolers. *The American Behavioral Scientist, 48*(5), 539–561.
- Apple. (2019). Retrieved August 2019, from Apple website:
<https://www.apple.com/education/products/>
- Astington, J. W., Pelletier, J., & Homer, B. (2002). Theory of mind and epistemological development: The relation between children's second-order false-belief understanding and their ability to reason about evidence. *New Ideas in Psychology, 20*(2-3), 131–144.
- Ayres, P., & Sweller, J. (2005). The Split-Attention Principle in Multimedia Learning. *The Cambridge Handbook of Multimedia Learning, 135–146*.
- Baillargeon, R. (2002). The acquisition of physical knowledge in infancy: A summary in eight lessons. In U. Goswami (Ed.), *Blackwell Handbook of Childhood Cognitive Development* (pp. 47–83). Blackwell publishing.

- Baillargeon, R. (2004). Infants' Physical World. *Current Directions in Psychological Science*, 13(3), 89–94.
- Bauer, H. H. (1994). *Scientific literacy and the myth of the scientific method* (Vol. 22). University of Illinois Press.
- Bauer, J.-R., & Booth, A. E. (2019). Exploring potential cognitive foundations of scientific literacy in preschoolers: Causal reasoning and executive function. *Early Childhood Research Quarterly*, 46, 275–284.
- Baylor, A. L., & Ryu, J. (2003). The Effects of Image and Animation in Enhancing Pedagogical Agent Persona. *Journal of Educational Computing Research*, 28(4), 373–394.
- Becker, D. (2018). *Design and Evaluation of an Educational Application for Preschoolers* (Master; A. Moeller, Ed.). Ludwig-Maximilians-Universität München.
- Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development*, 77(2), 413–426.
- Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, 93(1), 26–55.
- Bétrancourt, M. (2005). The Animation and Interactivity Principles in Multimedia Learning. *The Cambridge Handbook of Multimedia Learning*, pp. 287–296.
- Bloom, B. S. (1956). *Taxonomy of educational objectives. Vol. 1: Cognitive domain* (pp. 20–24). New York: McKay.
- Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3), 322–330.

- Bredenkamp, S., & Copple, C. (1997). *Developmentally Appropriate Practice in Early Childhood Programs. (Revised Edition)*. National Association for the Education of Young Children.
- Bromme, R., & Goldman, S. R. (2014). The public's bounded understanding of science. *Educational Psychologist, 49*(February 2015), 59–69.
- Bryant, P., Nunes, T., Hillier, J., Gilroy, C., & Barros, R. (2013). The importance of being able to deal with variables in learning science. *International Journal of Science and Mathematics Education, March*, 1–19.
- Budke, A., & Meyer, M. (2015). *Fachlich argumentieren lernen: Die Bedeutung der Argumentation in den unterschiedlichen Schulfächern*. Münster: Waxmann.
- Bullock, M. (1991). *Scientific reasoning in elementary school: Developmental and individual differences*. Max-Planck-Inst. for Psychological Research, Munich (West Germany).
- Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. *The Developmental Psychology of Time*, 209–254.
- Bullock, M., Sodian, B., & Koerber, S. (2009). Doing Experiments and Understanding Science: Development of Scientific Reasoning from Childhood to Adulthood. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood* (pp. 183–208). Psychology Press.
- Bullock, M., & Ziegler, A. (1999). Scientific reasoning: developmental and individual differences. In *Individual development from 3 to 12: Findings from the Munich Longitudinal Study* (pp. 38–60). Cambridge University Press.
- Bustamante, A., Greenfield, D., & Nayfeld, I. (2018). Early Childhood Science and Engineering: Engaging Platforms for Fostering Domain-General Learning Skills. *Education Sciences, 8*(3), 144.

- Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development, 7*(2), 213–233.
- Carey, S. (1987). Theory change in childhood. *Piaget Today*, 141–163.
- Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.
- Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C. (1989). “An experiment is when you try it and see if it works”: A study of grade 7 students' understanding of the construction of scientific knowledge. *International Journal of Science Education, 11*(5), 514–529.
- Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology, 87*(4), 299–319.
- Carroll, J. M. (2004). Beyond fun. *Interactions, 11*(5), 38–40.
- Carver, S., & Shrager, J. (2012). Is development domain specific or domain general? a third alternative. In S. Carver & J. Shrager (Eds.), *The journey from child to scientist: Integrating cognitive development and the education sciences* (p. 305). Washington D.C., USA: American Psychological Association.
- Case, R. (1974). Structures and Strictures: Some Functional Limitations on the Course of Cognitive Growth. *Cognitive Psychology, 6*, 544–573.
- Chapman, P. M. (1997). *Models of engagement: Intrinsically motivated interaction with multimedia learning software*. University of Waterloo.
- Chen, S.-M. (2009). Shadows: young Taiwanese children's views and understanding. *International Journal of Science Education, 31*(1), 59–79.
- Chen, Z. (2012). *The learning of science and the science of learning: The role of analogy*.

- Chen, Z., & Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098–1120.
- Chen, Z., Mo, L., Klahr, D., Tong, X., Qu, C., & Chen, H. (2011). Learning to test hypotheses: Kindergartners and elementary school children’s acquisition of scientific reasoning strategies. *Manuscript in Preparation*.
- Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73–105.
- Chi, M. T. H., Glaser, R., Reimann, P., Lewis, M. W., & Bassok, M. (2005). Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cognitive Science, 13*(2), 145–182.
- Ching-Ting Hsin, Ming-Chaun Li, & Chin-Chung Tsai. (2014). The Influence of Young Children’s Use of Technology on Their Learning: A Review. *Journal of Educational Technology & Society, 17*(4), 85–99.
- Chinn, C. A., Buckland, L. A., & Samarapungavan, A. (2011). Expanding the Dimensions of Epistemic Cognition: Arguments From Philosophy and Psychology. *Educational Psychologist, Vol. 46*, pp. 141–167.
- Chiong, C., & Shuler, C. (2010). Learning: Is there an app for that. *Investigations of Young Children’s Usage and Learning with Mobile Devices and Apps. New York: The Joan Ganz Cooney Center at Sesame Workshop, 13–20*.
- Clark, C. A. C., Nelson, J. M., Garza, J., Sheffield, T. D., Wiebe, S. A., & Espy, K. A. (2014). Gaining control: changing relations between executive control and processing speed and their relevance for mathematics achievement over course of the preschool period. *Frontiers in Psychology, 5*, 107.
- Clements, D. H. (2002). Computers in Early Childhood Mathematics. *Contemporary Issues in Early Childhood, 3*(2), 160–181.

- Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.
- Cohen, L. B., & Amsel, G. (1998). Precursors to infants' perception of the causality of a simple event. *Infant Behavior and Development*, Vol. 21, pp. 713–731.
- Cohen, L. B., Amsel, G., Redford, M. A., & Casasola, M. (1998). The development of infant causal perception. *Perceptual Development: Visual, Auditory, and Speech Perception in Infancy*, 167–209.
- Cohen, L. B., & Oakes, L. M. (1993). How infants perceive a simple causal event. *Developmental Psychology*, Vol. 29, pp. 421–433.
- Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349.
- Coull, G. J., Leekam, S. R., & Bennett, M. (2006). Simplifying second-order belief attribution: What facilitates children's performance on measures of conceptual understanding?. *Social Development*, *15*(3), 548-563.
- Couse, L. J., & Chen, D. W. (2010). A Tablet Computer for Young Children? Exploring its Viability for Early Childhood Education. *Journal of Research on Technology in Education*, Vol. 43, pp. 75–96.
- Crawley, A. M., Anderson, D. R., Wilder, A., Williams, M., & Santomero, A. (1999). Effects of repeated exposures to a single episode of the television program Blue's Clues on the viewing behaviors and comprehension of preschool children. *Journal of Educational Psychology*, *91*(4), 630.
- Croker, S., & Buchanan, H. (2011). Scientific reasoning in a real-world context: The effect of prior belief and outcome on children's hypothesis-testing strategies. *The British Journal of Developmental Psychology*, *29*, 409–424.

- Crowley, K., Callanan, M. A., Jipson, J. L., Galco, J., Topping, K., & Shrager, J. (2001). Shared scientific thinking in everyday parent-child activity. *Science Education*, 85(6), 712–732.
- Davis, H. L., & Pratt, C. (1995). The development of children's theory of mind: The working memory explanation. *Australian Journal of Psychology*, Vol. 47, pp. 25–31.
- Daxenberger, J., Csanadi, A., Ghanem, C., Kollar, I., & Gurevych, I. (2018). Domain-Specific Aspects of Scientific Reasoning and Argumentation. In F. Fischer, C. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific Reasoning and argumentation: The Roles of Domain-Specific and Domain-General Knowledge* (pp. 34–55). Routledge.
- Dean, D., Jr., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education*, 91(3), 384–397.
- Dede, C. (2010). Comparing frameworks for 21st century skills. *21st Century Skills: Rethinking How Students Learn*, 20, 51–76.
- De Vries, E., Lund, K., & Baker, M. (2002). Computer-mediated epistemic dialogue: Explanation and argumentation as vehicles for understanding scientific notions. *The Journal of the Learning Sciences*, 11(1), 63–103.
- Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, 64, 135–168.
- Dickey, M. D. (2005). Engaging by design: How engagement strategies in popular computer and video games can inform instructional design. *Educational Technology Research and Development: ETR & D*, 53(2), 67–83.
- Din, F. S., & Calao, J. (2001). The effects of playing educational video games on kindergarten achievement. *Child Study Journal*, 31(2), 95–103.
- Donker, A., & Markopoulos, P. (2002). A comparison of think-aloud, questionnaires and interviews for testing usability with children. In *People and Computers XVI-Memorable Yet Invisible* (pp. 305-316). Springer, London.

- Donker, A., & Reitsma, P. (2004). Usability Testing With Young Children. *Interaction Design and Children*, 43–48.
- Dreon, O., Kerper, R. M., & Landis, J. (2011). Digital storytelling: A tool for teaching and learning in the YouTube generation. *Middle School Journal*, 42(5), 4–10.
- Driver, R., Leach, J., Millar, R., & Scott, P. (1996). *Young people's images of science*. Philadelphia, PA. Open University Press.
- Druin, A. (1999). Developing Cooperative New Technologies Inquiry: for Children with Children. *Human-Computer Interaction*, Vol. 14, pp. 592–599.
- Druin, A. (2002). The role of children in the design of new technology. *Behaviour & Information Technology*, 21(1), 1–25.
- Dubé, A. K., & McEwen, R. N. (2017). Abilities and affordances: factors influencing successful child–tablet communication. *Educational Technology Research and Development: ETR & D*, 65(4), 889–908.
- Dumas, J. S., & Redish, J. (1999). *A Practical Guide to Usability Testing*. Intellect books.
- Dunbar, K., & Klahr, D. (2012). Scientific Thinking and Reasoning. *Oxford Handbooks Online*.
- Dunbar, K., & Klahr, D. (2013). Developmental differences in scientific discovery processes. In *Complex information processing* (pp. 129–164). Psychology Press.
- Dunn, J., & Munn, P. (1987). Development of justification in disputes with mother and sibling. *Developmental Psychology*, Vol. 23, pp. 791–798.
- Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education*.
- Edelsbrunner, P. A. (2017). *Domain-General and Domain-Specific Scientific Thinking in Childhood: Measurement and Educational Interplay*. ETH Zurich.

- Elliott, A., & Hall, N. (1997). The impact of self-regulatory teaching strategies on “at-risk” preschoolers’ mathematical learning in a computer-mediated environment. *Journal of Computing in Childhood Education*, 8.
- Erb, C. D., & Sobel, D. M. (2014). The development of diagnostic reasoning about uncertain events between ages 4-7. *PloS One*, 9(3).
- Erduran, S. (2007). Methodological Foundations in the Study of Argumentation in Science Classrooms. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in Science Education: Perspectives from Classroom-Based Research* (pp. 47–69). Dordrecht: Springer Netherlands.
- Fernbach, P. M., Macris, D. M., & Sobel, D. M. (2012). Which one made it go? The emergence of diagnostic reasoning in preschoolers. *Cognitive Development*, 27(1), 39–53.
- Fischer, F., Chinn, C., Engelmann, K., & Osborne, J. (Eds.). (2018). *Scientific Reasoning and Argumentation: The Roles of Domain-specific and Domain-general Knowledge*. Routledge.
- Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., ... Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, 4, 28–45.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive--developmental inquiry. *The American Psychologist*, 34(10), 906.
- Flavell, J. H. (2000). Development of children’s knowledge about the mental world. *International Journal of Behavioral Development*, 24(1), 15–23.
- Fransen, S., & Markopoulos, P. (2010). Let Robots Do the Talking. *9th International Conference on Interaction Design and Children*, 59–68.

- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research*, 74(1), 59–109.
- Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: performance of children 312–7 years old on a stroop-like day-night test. *Cognition*, 53(2), 129–153.
- Gilbert, J. (2005). *Catching the knowledge wave?: The knowledge society and the future of education*. Nzcer Press Wellington.
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. *Psychological Review*, Vol. 111, pp. 3–32.
- Gopnik, A., & Graf, P. (1988). Knowing How You Know: Young Children's Ability to Identify and Remember the Sources of Their Beliefs. *Child Development*, 59(5), 1366–1371.
- Gopnik, A., & Meltzoff, A. N. (1997). *Learning, development, and conceptual change*. Words, thoughts, and theories. Cambridge, MA, US: The MIT Press.
- Gopnik, A., & Schulz, L. (2007). *Causal Learning: Psychology, Philosophy, and Computation*. Oxford University Press, USA.
- Gopnik, A., & Sobel, D. M. (2000). Detectingblickets: how young children use information about novel causal powers in categorization and induction. *Child Development*, 71(5), 1205–1222.
- Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, 37(5), 620–629.

- Gopnik, A., & Wellman, H. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, 138(6), 1085–1108.
- Gopnik, A., & Wellman, H. M. (1992). Why the Child's Theory of Mind Really Is a Theory. *Mind & Language*, 7(1-2), 145–171.
- Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 257–293). Cambridge University Press.
- Gould, J. D., & Lewis, C. (1983). Designing for usability---key principles and what designers think. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '83*.
- Gropen, J., Clark-Chiarelli, N., Hoisington, C., & Ehrlich, S. B. (2011). The importance of executive function in early science education. *Child Development Perspectives*, 5(4), 298–304.
- Grygier, P. (2008). *Wissenschaftsverständnis von Grundschulern im Sachunterricht*. Julius Klinkhardt.
- Gweon, H., & Schulz, L. (2008). Stretching to learn: Ambiguous evidence and variability in preschoolers exploratory play. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, 570–574.
- Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, 107(20), 9066–9071.
- Hall, L., Hume, C., & Tazzyman, S. (2016). Five Degrees of Happiness. *Proceedings of the The 15th International Conference on Interaction Design and Children - IDC '16*.
- Hammond, J., & Gibbons, P. (2005). What is scaffolding. *Teachers' Voices*, 8, 8–16.

- Hanna, L., Ridsden, K., & Alexander, K. (1997). Guidelines for usability testing with children. *Interactions*, 4(5), 9–14.
- Harris, P. L., German, T., & Mills, P. (1996). Children’s use of counterfactual thinking in causal reasoning. *Cognition*, 61(3), 233–259.
- Haslbeck, H., Lankes, E. M., Fritzsche, E. S., Kohlhauf, L., & Neuhaus, B. J. (2018). How Do Kindergarten and Primary School Children Justify Their Decisions on Planning Science Experiments? *International Society of the Learning Sciences*, 1601–1602.
- Hassenzahl, M., & Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & Information Technology*, 25(2), 91–97.
- Haßler, B., Major, L., & Hennessy, S. (2016). Tablet use in schools: a critical review of the evidence for learning outcomes. *Journal of Computer Assisted Learning*, 32(2), 139–156.
- Herodotou, C. (2018). Mobile games and science learning: A comparative study of 4 and 5 years old playing the game Angry Birds. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, 49(1), 6–16.
- Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015). Putting Education in “Educational” Apps: Lessons From the Science of Learning. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, 16(1), 3–34.
- Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark. *Educational Psychologist*, 42(2), 99–107.

- Hofer, T., Hauf, P., & Aschersleben, G. (2004). *Wahrnehmung von zielgerichteten Handlungen bei 6 Monate alten Säuglingen. Ein Vergleich von TV-vs. Live-Präsentationen.*
- Höffler, T. N., & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction, 17*(6), 722–738.
- Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development, 9*(1), 45–75.
- Inhelder, B., & Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence: An Essay on the Construction of Formal Operational Structures.* Psychology Press.
- International Organization for Standardization. (2018). *Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts* (No. 9241-11:2018).
- Iordanou, K. (2016). Developing Epistemological Understanding in Scientific and Social Domains through Argumentation. *Zeitschrift Für Pädagogische Psychologie, 30*(2-3), 109–119.
- Keil, F. C. (1989). *Concepts, kinds, and conceptual development.* Cambridge, MA: MIT Press.
- Kind, P. M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *Journal of Research in Science Teaching, 50*(5), 530–560.
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educational Psychologist, 41*(2), 75–86.
- Kittredge, A. K., Klahr, D., & Willows, K. (2015). Thinking about play: Young children's spontaneous experiments and scientific reasoning. *Paper Presented at the Workshop*

on Digital Assessment and Promotion of Children's Curiosity, Interaction Design and Children.

Klahr, D., & Dunbar, K. (1988). *Cognitive Science: A Multidisciplinary Dual space search during scientific reasoning.* 48(768320842).

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: a developmental study. *Cognitive Psychology*, 25(1), 111–146.

Klahr, D., & Nigam, M. (2004). The Equivalence of Learning Paths in Early Science Instruction: Effects of Direct Instruction and Discovery Learning. *Psychological Science*, 15(10), 661–667.

Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational Interventions to Advance Children's Scientific Thinking. *Science*, 333(6045), 971–975.

Klein, P. D. (1998). The Role of Children's Theory of Mind in Science Experimentation. *Journal of Experimental Education*, 66(2), 101–124.

Köksal-Tuncer, Ö., & Sodian, B. (2018). The development of scientific reasoning: Hypothesis testing and argumentation from evidence in young children. *Cognitive Development*, 48(June 2017), 135–145.

Köksal-Tuncer, Ö., Sodian, B., & Legare, C. H. (2019). Young children's metacognitive awareness of confounded evidence. Manuscript in preparation. Ludwig-Maximilians-Universität München.

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The Development of Scientific Thinking in Elementary School: A Comprehensive Inventory. *Child Development*, 86(1), 327–336.

Koerber, S., & Osterhaus, C. (2019). Individual Differences in Early Scientific Thinking: Assessment, Cognitive Influences, and Their Relevance for Science Learning. *Journal of Cognition and Development*, 1–24.

- Koerber, S., Sodian, B., Kropf, N., Mayer, D., & Schwippert, K. (2011). Die Entwicklung des wissenschaftlichen Denkens im Grundschulalter [The Development of Scientific Thinking in Elementary]. *Zeitschrift Fur Entwicklungspsychologie Und Padagogische Psychologie*, 43, 16–21.
- Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss Journal of Psychology: Official Publication of the Swiss Psychological Society Schweizerische Zeitschrift Fur Psychologie = Revue Suisse de Psychologie*, 64(3), 141–152.
- Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. MIT Press.
- Krathwohl, D. R., & Anderson, L. W. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. *Theory Into Practice*, 41(4), 212.
- Kristen, S., Thoermer, C., Hofer, T., Aschersleben, G., & Sodian, B. (2006). Skalierung von "Theory of Mind"-aufgaben. *Zeitschrift Fur Entwicklungspsychologie Und Padagogische Psychologie*, 38(4), 186–195.
- Kuczynski, L., & Kochanska, G. (1990). Development of children's noncompliance strategies from toddlerhood to age 5. *Developmental Psychology*, 26(3), 398.
- Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, 96(4), 674–689.
- Kuhn, D. (1991). *The Skills of Argument*. Cambridge University Press.
- Kuhn, D. (2000). Theory of mind, metacognition, and reasoning: A life-span perspective. *Children's Reasoning and the Mind*, 301–326.

- Kuhn, D. (2002). What is Scientific Thinking and How Does It Develop? In U. Goswami (Ed.), *Blackwell Handbook of Childhood Cognitive Development* (pp. 371–393). Blackwell Publishers Ltd.
- Kuhn, D. (2007a). Jumping to Conclusions: Can people be counted on to make sound judgments? *Scientific American Mind*, 44–51.
- Kuhn, D. (2007b). Reasoning about multiple variables: Control of variables is not the only challenge. *Science Education*, 91(5), 710–726.
- Kuhn, D. (2010). Teaching and learning science as argument. *Science Education*, 94(5), 810–824.
- Kuhn, D., Amsel, E., & O’Loughlin, M. (1988). *The development of scientific thinking skills*. Academic Press.
- Kuhn, D., & Crowell, A. (2011). *Argumentation as a Path to the Thinking Development of Young Adolescents*. ERIC Clearinghouse.
- Kuhn, D., & Dean, D. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development: Official Journal of the Cognitive Development Society*.
- Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*.
- Kuhn, D., & Franklin, S. (2007). The Second Decade: What Develops (and How). *Handbook of Child Psychology*.
- Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Journal of Social and Biological Systems*, 60(4).
- Kuhn, D., Goh, W., Iordanou, K., & Shaenfield, D. (2008). Arguing on the computer: A microgenetic study of developing argument skills in a computer-supported environment. *Child Development*, 79(5), 1310–1328.

- Kuhn, D., Jordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: What needs to develop to achieve skilled scientific thinking? *Cognitive Development*, 23(4), 435–451.
- Kuhn, D., & Pearsall, S. (1998). Relations between metastrategic knowledge and strategic performance. *Cognitive Development*, 13(2), 227–247.
- Kuhn, D., & Pearsall, S. (2000). Developmental Origins of Scientific Thinking. *Journal of Cognition and Development: Official Journal of the Cognitive Development Society*, 1(1), 113–129.
- Kuhn, D., & Phelps, E. (1982). The Development of Problem-Solving Strategies. *Advances in Child Development and Behavior*, 1–44.
- Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-Domain Development of Scientific Reasoning. *Cognition and Instruction*, 9(4), 285–327.
- Kuhn, J.-T., & Holling, H. (2009). Gender, reasoning ability, and scholastic achievement: A multilevel mediation analysis. *Learning and Individual Differences*, 19(2), 229–233.
- Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology*, 43(1), 186–196.
- Kushnir, T., & Gopnik, A. (2018). Young Children Infer Causal Strength From Probabilities and Interventions. *Psychological Science*, 16(9), 678–683.
- Kwon, Y.-J., & Lawson, A. E. (2000). Linking brain growth with the development of scientific reasoning ability and conceptual change during adolescence. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 37(1), 44–62.

- Lazonder, A. W., & Egberink, A. (2013). Children's acquisition and use of the control-of-variables strategy: effects of explicit and implicit instructional guidance. *Instructional Science*, 1–14.
- Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development*, 83(1), 173–185.
- Legare, C. H. (2014). The contributions of explanation and exploration to children's scientific reasoning. *Child Development Perspectives*, 8(2), 101–106.
- Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*, 81(3), 929–944.
- Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, 126, 198–212.
- León, J. M. (2015). A Baseline Study of Strategies to Promote Critical Thinking in the Preschool Classroom. *GiST Education and Learning Research Journal*, pp. 113–127.
- Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25(3), 265–288.
- Lieberman, D. A., Bates, C. H., & So, J. (2009). Young Children's Learning With Digital Media. *Computers in the Schools*, 26(4), 271–283.
- Lieberman, D. A., & Linn, M. C. (1991). Learning to learn revisited: Computers and the development of self-directed learning skills. *Journal of Research on Computing in Education*, 23(3), 373–395.
- Linn, M. C., Clement, C., & Pulos, S. (1983). Is it formal if it's not physics?(The influence of content on formal reasoning). *Journal of Research in Science Teaching*, 20(8), 755–770.

- Lorch, R. F., Jr, Lorch, E. P., Calderhead, W. J., Dunlap, E. E., Hodell, E. C., & Freer, B. D. (2010). *Learning the Control of Variables Strategy in Higher and Lower Achieving Classrooms : Contributions of Explicit Instruction and Experimentation*. *102*(1), 90–101.
- Lowe, R. K. (2003). Animation and learning: selective processing of information in dynamic graphics. *Learning and Instruction*, *13*(2), 157–176.
- Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.
- Luo, Y., Kaufman, L., & Baillargeon, R. (2009). Young infants' reasoning about physical events involving inert and self-propelled objects. *Cognitive Psychology*, *58*(4), 441–486.
- Malone, T. (8AD). Lepper (1987). Making Learning Fun: A Taxonomy of Intrinsic Motivations for Learning. *Aptitude, Learning, and Instruction*, *3*.
- Maric, M., & Sakac, M. (2018). Metacognitive components as predictors of preschool children's performance in problem-solving tasks. *Psihologija*, Vol. 51, pp. 1–16.
- Markopoulos, P., Read, J. C., MacFarlane, S., & Hoysniemi, J. (2008). *Evaluating Children's Interactive Products: Principles and Practices for Interaction Designers*. Elsevier.
- Mascalzoni, E., Regolin, L., Vallortigara, G., & Simion, F. (2013). The cradle of causal reasoning: newborns' preference for physical causality. *Developmental Science*, Vol. 16, pp. 327–335.
- Masnick, A. M., & Klahr, D. (2003). Error Matters: An Initial Exploration of Elementary School Children's Understanding of Experimental Error. *Journal of Cognition and Development: Official Journal of the Cognitive Development Society*, *4*(1), 67–98.

- Matlen, B. J., & Klahr, D. (2013). Sequential effects of high and low instructional guidance on children's acquisition of experimentation skills: Is it all in the timing? *Instructional Science*, *41*(3), 621–634.
- Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning and Instruction*, *29*, 43–55.
- Mayer, R. E. (2002). Multimedia learning. In *Psychology of Learning and Motivation* (Vol. 41, pp. 85–139). Academic Press.
- Mayer, R. E. (2005). Introduction to multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press.
- Mayer, R. E. (2014). Principles Based on Social Cues in Multimedia Learning: Personalization, Voice, Image, and Embodiment Principles. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 345–368).
- Mayer, R. E., & DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology. Applied*, *18*(3), 239–252.
- Mayer, R. E., & Pilegard, C. (2005). Principles for managing essential processing in multimedia learning: Segmenting, pretraining, and modality principles. *The Cambridge Handbook of Multimedia Learning*, 169–182.
- McComas, W. F., Clough, M. P., & Almazroa, H. (1998). The role and character of the nature of science in science education. In *The nature of science in science education* (pp. 3-39). Springer, Dordrecht.
- McLeod, S. A. (2019). Controlled Experiment. Retrieved 2019, from Simply Psychology website: <https://www.simplypsychology.org/controlled-experiment.html>

- McManis, L. D., & Gunnewig, S. B. (2012). Finding the education in educational technology with early learners. *YC Young Children*, 67(3), 14–24.
- Mercier, H. (2011). Reasoning serves argumentation in children. *Cognitive Development*, 26(3), 177–191.
- Millar, R., & Driver, R. (1987). Beyond Processes. *Studies in Science Education*, 14(1), 33–62.
- Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.
- Miller, M. R., Giesbrecht, G. F., Müller, U., McInerney, R. J., & Kerns, K. A. (2012). A Latent Variable Approach to Determining the Structure of Executive Function in Preschool Children. *Journal of Cognition and Development: Official Journal of the Cognitive Development Society*, 13(3), 395–423.
- Miller, S. A. (2012). *Theory of mind: Beyond the preschool years*. Psychology Press.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex “Frontal Lobe” Tasks: A Latent Variable Analysis. *Cognitive Psychology*, 41(1), 49–100.
- Monk, A., Davenport, L., Haber, J., & Wright, P. (1993). *Improving your human-computer interface: A practical technique*. New York: Prentice-Hall.
- Morris, B. J., Croker, S., Masnick, A. M., & Zimmerman, C. (2012). The Emergence of Scientific Reasoning. *Current Topics in Children’s Learning and Cognition*, 2, 64.
- Muentener, P., & Bonawitz, E. (2018). The development of causal reasoning. In M. Waldmann (Ed.), *Oxford Handbook of Causal Reasoning*. Oxford, UK: Oxford University Press.

- Muentener, P., & Carey, S. (2010). Infants' causal representations of state change events. *Cognitive Psychology*, *61*(2), 63–86.
- National Committee on Science Education Standards. (1996). *National Science Education Standards*.
- National Research Council (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, DC: The National Academies.
- Nayfeld, I., Fuccillo, J., & Greenfield, D. B. (2013). Executive functions in early learning: Extending the relationship between executive functions and school readiness to science. *Learning and Individual Differences*, *26*, 81–88.
- Nazzi, T., & Gopnik, A. (2003). Sorting and acting with objects in early childhood: an exploration of the use of causal cues. *Cognitive Development*, Vol. 18, pp. 299–317.
- Nenciovici, L., Allaire-Duquette, G., & Masson, S. (2019). Brain activations associated with scientific reasoning: a literature review. *Cognitive Processing*, *20*(2), 139–161.
- Newman, G. E., Choi, H., Wynn, K., & Scholl, B. J. (2008). The origins of causal perception: evidence from postdictive processing in infancy. *Cognitive Psychology*, *57*(3), 262–291.
- NGSS Lead States (2013). Next generation science standards: For states, by states. Washington, DC: The National Academies Press.
- Nielsen, J. (1994). Usability inspection methods. *Conference Companion on Human Factors in Computing Systems*, 413–414. ACM.
- O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science. American Society for Information Science*, *59*(6), 938–955.

- Oerke, B., & Bogner, F. X. (2013). Social Desirability, Environmental Attitudes, and General Ecological Behaviour in Children. *International Journal of Science Education, 35*(5), 713–730.
- Ofcom. (2017). *Children and Parents: Media Use and Attitudes Report*. Ofcom.
- Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning – a review of test instruments. *Educational Research and Evaluation, 23*(3-4), 78–101.
- Osborne, J. (2010). Arguing to learn in science: the role of collaborative, critical discourse. *Science, 328*(5977), 463–466.
- Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10*, 265–279.
- Osterhaus, C., Koerber, S., & Sodian, B. (2017). Scientific thinking in elementary school: Children’s social cognition and their epistemological understanding promote experimentation skills. *Developmental Psychology, 53*(3), 450–462.
- Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist, 38*(1), 1–4.
- Paivio, A. (1990). *Mental Representations: A Dual Coding Approach*. Oxford University Press.
- Palaiologou, I. (2016). Children under five and digital technologies: implications for early years pedagogy. *European Early Childhood Education Research Journal, 24*(1), 5–24.
- Partnership for 21st Century Skills. (2009). *P21 Framework Definitions*. ERIC Clearinghouse.
- Perner, J. (1991). *Understanding the representational mind*. The MIT Press.

- Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, 39(3), 437-471.
- Petermann, F., & Lipsius, M. (2009). *WPPSI-III. Wechsler Preschool and Primary Scale of Intelligence--Third Edition. Deutsche Version*. Frankfurt: Pearson Assessment.
- Piaget, J. (1970). *Science of education and the psychology of the child*. Trans. D. Coltman.
- Piekny, J., Grube, D., & Maehler, C. (2013). The relation between preschool children's false-belief understanding and domain-general experimentation skills. *Metacognition and Learning*, 8(2), 103–119.
- Piekny, J., Grube, D., & Maehler, C. (2014). The Development of Experimentation and Evidence Evaluation Skills at Preschool Age. *International Journal of Science Education*, 36(2), 334–354.
- Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *The British Journal of Developmental Psychology*, 31(2), 153–179.
- Plowman, L., & Stephen, C. (2003). A "benign addition"? Research on ICT and pre-school children. *Journal of Computer Assisted Learning*, 19(2), 149–164.
- Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the Horizon*, 9(5), 1–6.
- Prenzel, M., Rost, J., Senkbeil, M., Häußler, P., & Klopp, A. (2001).
Naturwissenschaftliche Grundbildung: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, ... M. Weiß (Eds.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 191–248). Wiesbaden: VS Verlag für Sozialwissenschaften.

- Rakytianska, V. (2019). *Influence of Design Elements on Preschoolers' Acquisition of the Control of Variables Strategy* (Bachelor; A. M. Bachhuber, Ed.). Ludwig-Maximilians-Universität München.
- Read, J., & Fine, K. (2005). Using survey methods for design and evaluation in child computer interaction. *Workshop on Child Computer Interaction: Methodological Research at Interact*.
- Read, J., MacFarlane, S., & Casey, C. (2002). Endurability, engagement and expectations: Measuring children's fun. *Interaction Design and Children*, 2, 1–23. Shaker Publishing Eindhoven.
- Resnick, M. (1998). Technologies for lifelong kindergarten. *Educational Technology Research and Development: ETR & D*, 46(4), 43–55.
- Rideout, V. (2017). The common sense census: Media use by kids age zero to eight. *San Francisco, CA: Common Sense Media*, 263–283.
- Rieber, L. P., & Kini, A. S. (1991). Theoretical foundations of instructional applications of computer-generated animated visuals. *Journal of Computer Based Instruction*, 18(3), 83-88.
- Risden, K., Hanna, E., & Kanerva, A. (1997). Dimensions of intrinsic motivation in children's favorite computer activities. *Society for Research in Child Development, Washington, DC*, 7.
- Rittle-Johnson, B., Saylor, M., & Swygert, K. E. (2008). Learning from explaining: Does it matter if mom is listening? *Journal of Experimental Child Psychology*, 100(3), 215–224.
- Robin, B. R. (2008). Digital Storytelling: A Powerful Technology Tool for the 21st Century Classroom. *Theory into Practice*, 47(3), 220–228.

- Rohwer, M., Kloos, D., & Perner, J. (2012). Escape From Metaignorance: How Children Develop an Understanding of Their Own Lack of Knowledge. *Child Development*, 83(6), 1869–1883.
- Rosengren, K. S., Gelman, S. A., Kalish, C. W., & McCormick, M. (1991). As time goes by: children's early understanding of growth in animals. *Child Development*, 62(6), 1302–1320.
- Ross, J. A. (1988). Controlling Variables: A Meta-Analysis of Training Studies. *Review of Educational Research*, 58(4), 405–437.
- Ruffman, T., Perner, J., Olson, D. R., & Doherty, M. (1993). Reflecting on Scientific Thinking: Children's Understanding of the Hypothesis-Evidence Relation. *Child Development*, 64(6), 1617–1636.
- Saçkes, M., Trundle, K. C., Bell, R. L., & O'Connell, A. A. (2011). The influence of early science experience in kindergarten on children's immediate and later science achievement: Evidence from the early childhood longitudinal study. *Journal of Research in Science Teaching*, 48(2), 217–235.
- Saffran, A., Barchfeld, P., & Sodian, B. (2015). Die Interpretation von imperfekten Kovariationsdaten im Vorschulalter. *Poster Präsentiert Bei Der 22. Tagung Der Fachgruppe Entwicklungspsychologie Der Deutschen Gesellschaft Für Psychologie*. DGPs, Frankfurt, Deutschland.
- Saffran, A., Barchfeld, P., Sodian, B., & Alibali, M. W. (2016). Children's and adults' interpretation of covariation data: Does symmetry of variables matter? *Developmental Psychology*, 52(10), 1530–1544.
- Samarapungavan, A. (2018). Construing scientific evidence: The role of disciplinary knowledge in reasoning with and about evidence in scientific practice. In F. Fischer,

- C. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific Reasoning and Argumentation* (pp. 56–76). Routledge.
- Sandoval, W., Sodian, B., Koerber, S., & Wong, J. (2014). Developing Children's Early Competencies to Engage With Science. *Educational Psychologist*, *49*(2), 139–152.
- Sao Pedro, M. A., Gobert, J. D., & Raziuddin, J. J. (2010). Comparing pedagogical approaches for the acquisition and long-term robustness of the control of variables strategy. *Journal of the Learning Sciences*, (October 2016), 1024–1031.
- Sao Pedro, M. A., Gobert, J., Heffernan, N., & Beck, J. (2009). Comparing pedagogical approaches for teaching the control of variables strategy. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, *1*, 1294–1299.
- Saxe, R., & Carey, S. (2006). The perception of causality in infancy. *Acta Psychologica*, *123*(1-2), 144–165.
- Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology*, Vol. 49, pp. 31–57.
- Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*(1), 102–119.
- Schauble, L. (2018). In the eye of the beholder: Domain-General and domain-specific reasoning in science. In *Scientific reasoning and argumentation* (pp. 11–33). Routledge.
- Schauble, L., & Glaser, R. (1990). Scientific thinking in children and adults. *Developmental Perspectives on Teaching and Learning Thinking Skills.*, *134*, 9–27.
- Schauble, L., Glaser, R., Duschl, R. A., Schulze, S., & John, J. (1995). Students' Understanding of the Objectives and Procedures of Experimentation in the Science Classroom. *Journal of the Learning Sciences*, Vol. 4, pp. 131–166.

- Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1991). Causal Models and Experimentation Strategies in Scientific Reasoning. *Journal of the Learning Sciences*, *1*(2), 201–238.
- Schneider, W. (2008). The Development of Metacognitive Knowledge in Children and Adolescents: Major Trends and Implications for Education. *Mind, Brain and Education: The Official Journal of the International Mind, Brain, and Education Society*, *2*(3), 114–121.
- Schork, N. (2017). *Direct Instruction & Engagement - Guidelines for designing an educational video tutorial for children* (Bachelor; A. Moeller, Ed.). Ludwig-Maximilians-Universität München.
- Schubert, M. (2017). *The Effect of Animation on Preschoolers' Acquisition of the Control of Variables Strategy* (Bachelor; A. Moeller, Ed.). Ludwig-Maximilians-Universität München.
- Schult, C. A., & Wellman, H. M. (1997). Explaining human movements and actions: Children's understanding of the limits of psychological explanation. *Cognition*, *62*(3), 291–324.
- Schulz, L. E., & Bonawitz, E. B. (2007). Serious Fun: Preschoolers Engage in More Exploratory Play When Evidence Is Confounded. *Developmental Psychology*, *43*(4), 1045–1050.
- Schulz, L. E., & Gopnik, A. (2004). Causal Learning Across Domains. *Developmental Psychology*, *40*(2), 162–176.
- Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*(May), 322–332.

- Schulz, L. E., Standing, H. R., & Bonawitz, E. B. (2008). Word, thought, and deed: the role of object categories in children's inductive inferences and exploratory play. *Developmental Psychology*, *44*(5), 1266–1276.
- Schunn, C., & Anderson, J. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science*, *23*(3), 337–370.
- Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review: DR*, *39*, 37–63.
- Sharples, M., Taylor, J., & Vavoula, G. (2007). A Theory of Learning for the Mobile Age. *Medienbildung in Neuen Kulturräumen*, 87–99.
- Shultz, T. R. (1982). Rules of Causal Attribution. *Monographs of the Society for Research in Child Development*, Vol. 47, p. 1.
- Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*(4), 481–520.
- Siegler, R. S., & Chen, Z. (1998). Developmental Differences in Rule Learning: A Microgenetic Analysis. *Cognitive Psychology*, Vol. 36, pp. 273–310.
- Siler, S. A., & Klahr, D. (2012). Detecting, Classifying, and Remediating. *Psychology of Science*, pp. 137–180.
- Sim, G., MacFarlane, S., & Horton, M. (2005). Evaluating usability, fun and learning in educational software for children. *EdMedia+ Innovate Learning*, 1180–1187. Association for the Advancement of Computing in Education (AACE).
- Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L. W. Gregg (Ed.), *Knowledge and cognition*. Lawrence Erlbaum.
- Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*, *26*(2), 145.

- Sinatra, G. M., & Chinn, C. A. (2012). *Thinking and reasoning in science: Promoting epistemic conceptual change*.
- Smith, A., Toor, S., & van Kessel, P. (2018, November). Many Turn to YouTube for Children's Content, News, How-To Lessons.
- Smith, P. L., & Ragan, T. J. (2004). *Instructional Design*. John Wiley & Sons.
- Sobel, D. M. (2004). Exploring the coherence of young children's explanatory abilities: Evidence from generating counterfactuals. *The British Journal of Developmental Psychology*, 22(1), 37–58.
- Sobel, D. M., Erb, C. D., Tassin, T., & Weisberg, D. S. (2017). The Development of Diagnostic Inference About Uncertain Causes. *Journal of Cognition and Development: Official Journal of the Cognitive Development Society*, 18(5), 556–576.
- Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology*, 42(6), 1103–1115.
- Sobel, D. M., & Sommerville, J. A. (2009). Rationales in children's causal learning from others' actions. *Cognitive Development*, 24(1), 70–79.
- Sobel, D. M., & Sommerville, J. A. (2010). The Importance of Discovery in Children's Causal Learning from Interventions. *Frontiers in Psychology*, 1(November), 176.
- Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, 28(3), 303–333.
- Sobel, D. M., Yoachim, C. M., Gopnik, A., Meltzoff, A. N., & Blumenthal, E. J. (2007). The Blicket Within: Preschoolers' Inferences About Insides and Causes. *Journal of Cognition and Development: Official Journal of the Cognitive Development Society*, 8(2), 159–182.

- Sodian, B. (2018). The Development of Scientific Thinking in Preschool and Elementary School Age: A Conceptual Model. In *Scientific Reasoning and Argumentation* (pp. 227-250). Routledge.
- Sodian, B., & Bullock, M. (2008). Scientific reasoning-Where are we now? *Cognitive Development*, 23(4), 431–434.
- Sodian, B., & Frith, U. (2008). Metacognition, Theory of Mind, and Self-Control: The Relevance of High-Level Cognitive Processes in Development, Neuroscience, and Education. *Mind, Brain and Education: The Official Journal of the International Mind, Brain, and Education Society*, 2(3), 111–113.
- Sodian, B., Jonen, A., Thoermer, C., & Kircher, E. (2006). Die Natur der Naturwissenschaften verstehen. *Implementierung wissenschaftstheoretischen Unterrichts in der Grundschule*. *kr: Prenzel, M*, 147-160.
- Sodian, B., Kristen-Antonow, S., & Koerber, S. (2016). Theory of Mind Predicts Scientific Reasoning. A Longitudinal Study from Preschool to Elementary School Age. *International Journal of Psychology: Journal International de Psychologie*, 360–360.
- Sodian, B., & Mayer, D. (2013). Entwicklung des wissenschaftlichen Denkens im Vor- und Grundschulalter. In M. Stamm & D. Edelmann (Eds.), *Handbuch frühkindliche Bildungsforschung* (pp. 617–631). Wiesbaden: Springer Fachmedien Wiesbaden.
- Sodian, B., Thoermer, C., Kircher, E., Grygier, P., & Günther, J. (2002). Vermittlung von Wissenschaftsverständnis in der Grundschule. In *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen* (pp. 192-206).
- Sodian, B., Zaitchik, D., & Carey, S. (1991). Young Children's Differentiation of Hypothetical Beliefs from Evidence. *Child Development*, 62(4), 753–766.

- Song, J., & Black, P. J. (1992). The effects of concept requirements and task contexts on pupils' performance in control of variables. *International Journal of Science Education, 14*(1), 83–93.
- Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14*(1), 29–56.
- Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review, 99*(4), 605.
- StoryCenter. Retrieved 2019, from StoryCenter website: <https://www.storycenter.org/>
- Strand-Cary, M., & Klahr, D. (2008). Developing elementary science skills: Instructional effectiveness and path independence. *Cognitive Development, 23*(4), 488–511.
- Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science, Vol. 12*, pp. 257–285.
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*.
- Taylor, J. A., & Dana, T. M. (2003). Secondary school physics teachers' conceptions of scientific evidence: An exploratory case study. *Journal of Research in Science Teaching, Vol. 40*, pp. 721–736.
- The 7 Elements of Digital Storytelling. Retrieved 2019, from Educational Uses of Digital Storytelling <https://digitalstorytelling.coe.uh.edu/>
- Tolmie, A. K., Ghazali, Z., & Morris, S. (2016). Children's science learning: A core skills approach. *The British Journal of Educational Psychology, 86*(3), 481–497.
- Toth, E. E., Klahr, D., & Chen, Z. (2000). Bridging Research and Practice: A Cognitively Based Classroom Intervention for Teaching Experimentation Skills to Elementary School Children. *Cognition and Instruction, 18*(4), 423–459.
- Tricot, A., & Sweller, J. (2014). Domain-Specific Knowledge and Why Teaching Generic Skills Does Not Work. *Educational Psychology Review, 26*(2), 265–283.

- Trilling, B., & Fadel, C. (2009). *21st Century Skills: Learning for Life in Our Times*. John Wiley & Sons.
- Triona, L. M., & Klahr, D. (2003). Point and Click or Grab and Heft: Comparing the Influence of Physical and Virtual Instructional Materials on Elementary School Students' Ability to Design Experiments. *Cognition and Instruction*, *21*(2), 149–173.
- Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Development*, *51*(1), 1–10.
- Tullos, A., & Woolley, J. D. (2009). The development of children's ability to use evidence to infer reality status. *Child Development*, *80*(1), 101–114.
- UNESCO. (2014). *EFA Global Monitoring Report. Teaching and learning: Achieving quality for all*. UNESCO Paris.
- Van de Keere, K., Mestdagh, N., Dejonckheere, P., Vervae, S., & Tallir, I. (2014). An ICT simulation program to be used as a support and/or evaluation tool for scientific thinking in primary education. *Inquiry in Primary Science Education*, *1*(2), 4–12.
- van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: dynamic assessment of the control of variables strategy. *Instructional Science*, *43*(3), 381–400.
- van der Graaf, J., Segers, E., & Verhoeven, L. (2016). Scientific reasoning in kindergarten: Cognitive factors in experimentation and evidence evaluation. *Learning and Individual Differences*, *49*(July), 190–200.
- van der Graaf, J., Segers, E., & Verhoeven, L. (2018). Individual differences in the development of scientific thinking in kindergarten. *Learning and Instruction*, *56*(March), 1–9.
- van der Graaf, J., van de Sande, E., Gijssels, M., & Segers, E. (2019). A combined approach to strengthen children's scientific thinking: direct instruction on scientific reasoning

- and training of teacher's verbal support. *International Journal of Science Education*, 41(9), 1119–1138.
- van der Meij, H., & van der Meij, J. (2014). A comparison of paper-based and video tutorials for software learning. *Computers & Education*, 78, 150–159.
- Van Merriënboer, J. J. G., & Kester, L. (2005). The four-component instructional design model: Multimedia principles in environments for complex learning. *The Cambridge Handbook of Multimedia Learning*.
- van Schijndel, T. J. P. P., Visser, I., van Bers, B. M. C. W., & Raijmakers, M. E. J. J. (2015). Preschoolers perform more informative experiments after observing theory-violating evidence. *Journal of Experimental Child Psychology*, 131(March), 104–119.
- Vygotsky, L. S. (1980). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.
- Wagensveld, B., Segers, E., Kleemans, T., & Verhoeven, L. (2015). Child predictors of learning to control variables via instruction or self-discovery. *Instructional Science*, 43(3), 365–379.
- Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2013). Explaining to others prompts children to favor inductively rich properties. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35.
- Walker, C. M., Goel, D., Nyhout, A., & Ganea, P. (2019, March). *Evidence for early recognition of inconclusive data in children's evaluation of evidence*. Symposium presentation at the biennial conference of the Society for Research in Child Development (SRCD), Baltimore, Maryland, USA.
- Wang, Z. (2015). Theory of mind and children's understanding of teaching and learning during early childhood. *Cogent Education*, 2(1), 1011973.

- Wartella, E., O'Keefe, B., & Scantlin, R. (2000). Children and interactive media. *A Compendium of Current Research and Directions for the Future, Markle Foundation.*
- Waters, L. J., Siegal, M., & Slaughter, V. (2000). Development of Reasoning and the Tension between Scientific and Conversational Inference. *Social Development*, 9(3), 383–396.
- Webster, J., & Ho, H. (1997). Audience Engagement in Multimedia Presentations. *SIGMIS Database*, 28(2), 63–77.
- Wechsler, D. (2012). Wechsler Preschool and Primary Scale of Intelligence--Third Edition. *PsycTESTS Dataset.*
- Wellman, H. M. (1985). The child's theory of mind: The development of conceptions of cognition. *The Growth of Reflection in Children*, 169–206.
- Wellman, H. M. (1988). First steps in the child's theorizing about the mind. *Developing Theories of Mind*, 64–92.
- Wellman, H. M. (1992). *The child's theory of mind.* The MIT Press.
- Wellman, H. M. (2011). Reinvigorating Explanations for the Study of Early Cognitive Development. *Child Development Perspectives*, 5(1), 33–38.
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3), 655-684.
- Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: foundational theories of core domains. *Annual Review of Psychology*, 43, 337–375.
- Wellman, H. M., Hickling, A. K., & Schult, C. A. (1997). Young children's psychological, physical, and biological explanations. *New Directions for Child Development*, (75), 7–25.
- Wellman, H. M., & Liu, D. (2004). *Scaling of Theory-of-Mind Tasks.* 75(2), 523–541.

- Welsh, J. A., Nix, R. L., Blair, C., Bierman, K. L., & Nelson, K. E. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of Educational Psychology*, Vol. 102, pp. 43–53.
- Wiebe, S. A., Espy, K. A., & Charak, D. (2008). Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. *Developmental Psychology*, 44(2), 575–587.
- Wiebe, S. A., Sheffield, T., Nelson, J. M., Clark, C. A. C., Chevalier, N., & Espy, K. A. (2011). The structure of executive function in 3-year-olds. *Journal of Experimental Child Psychology*, 108(3), 436–452.
- Wilkening, F., & Sodian, B. (2005). Scientific, reasoning in young children: Introduction. *Swiss Journal of Psychology: Official Publication of the Swiss Psychological Society Schweizerische Zeitschrift Fur Psychologie = Revue Suisse de Psychologie*, 64(3), 137–139.
- Williams, J. J., & Lombrozo, T. (2010). Explanation constrains learning, and prior knowledge constrains explanation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 32.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103–128.
- Wouters, P., Paas, F., & van Merriënboer, J. J. G. (2008). How to Optimize Learning From Animated Models: A Review of Guidelines Based on Cognitive Load. *Review of Educational Research*, 78(3), 645–675.
- Yelland, N. (2005). *Critical Issues In Early Childhood Education*. McGraw-Hill Education (UK).

- YouTube. (2019, August). Retrieved August 2019, from YouTube website:
www.youtube.com
- Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, Vol. 42, p. 121.
- Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44(4), 1049–1060.
- Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): a method of assessing executive function in children. *Nature Protocols*, 1(1), 297–301.
- Zelazo, P. D., Carlson, S. M., & Kesek, A. (2008). *The development of executive function in childhood*.
- Zelazo, P. D., & Müller, U. (2010). Executive functioning in typical and atypical children. In U. Goswami (Ed.), *Blackwell Handbook of Childhood Cognitive Development* (pp. 574-603). Blackwell Publishers Ltd.
- Zimmerman, C. (2000). The Development of Scientific Reasoning Skills. *Developmental Review: DR*, 20(1), 99–149.
- Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review: DR*, 27(2), 172–223.
- Zimmerman, C., & Glaser, R. (2001). Testing Positive Versus Negative Claims: A Preliminary Investigation of the Role of Cover Story on the Assessment of Experimental Design Skills. *PsycEXTRA Dataset*.
- Zimmerman, C., & Klahr, D. (2018). Development of Scientific Thinking. In J. T. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (Vol. 27, pp. 1–25). Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Zohar, A. (1994). Teaching a thinking strategy: Transfer across domains and self learning versus class-like setting. *Applied Cognitive Psychology*, 8(6), 549–563.

- Zohar, A., & Aharon-Kravetsky, S. (2005). Exploring the effects of cognitive conflict and direct teaching for students of different academic levels. *Journal of Research in Science Teaching*, 42(7), 829–855.
- Zohar, A., & David, A. B. (2008). Explicit teaching of meta-strategic knowledge in authentic classroom situations. *Metacognition and Learning*, 3(1), 59–82.
- Zohar, A., & Peled, B. (2008). The effects of explicit teaching of metastrategic knowledge on low- and high-achieving students. *Learning and Instruction*, 18(4), 337–353.
- Zosh, J. M., Lytle, S. R., Golinkoff, R. M., & Hirsh-Pasek, K. (2016). Media Exposure During Infancy and Early Childhood. *Media Exposure During Infancy and Early Childhood*, 259–282.

Appendix

Appendix A: Study materials for the Lego scientific reasoning tasks (Studies 1, 2a, 2b, 3, 4, & 6)

Appendix B: Protocol of the Lego scientific reasoning tasks for Study 1

Appendix C: Summary of results of Studies 1, 2a, 2b, & 4

Appendix D: Protocol of the Lego scientific reasoning tasks for Study 2a

Appendix E: Protocol of the Lego scientific reasoning tasks for Studies 2b & 3

Appendix F: Protocol of the Lego scientific reasoning tasks for Study 4

Appendix G: Simple bivariate correlation table for the relevant variables, split by trials and tasks, in Study 4

Appendix H: Cross tabulation of mastery of Theory of Mind and the three scientific reasoning tasks (Study 4)

Appendix I: Protocol for Study 5a

Appendix J: Protocol for Study 5b

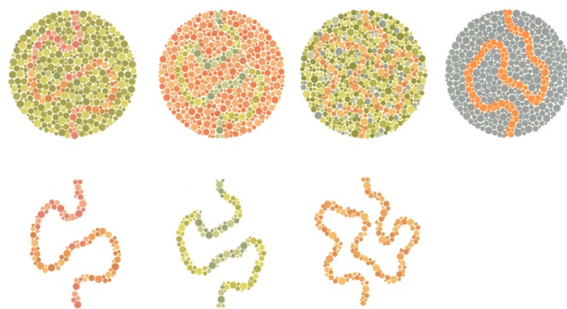
Appendix K: Protocol for Study 6

Appendix A: Materials for the Lego Scientific Reasoning Tasks (Studies 1, 2a, 2b, 3, 4, & 6)

The warm-up puzzle



The color vision test

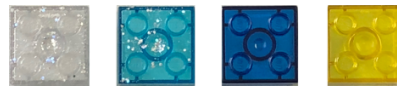


The blicket detector (not activated) with foot pedal; the blicket detector (activated)



The testing materials for Study 1

Familiarization



SET 1

SET 2

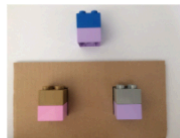
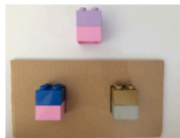
SET 3

SET 4

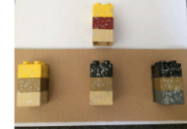
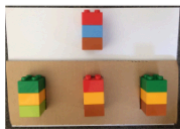
ICE Task



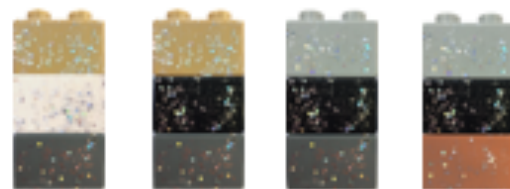
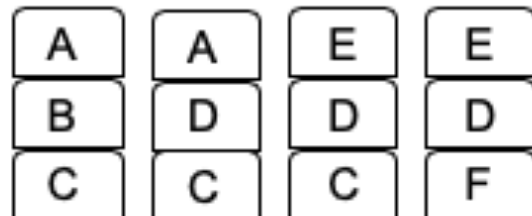
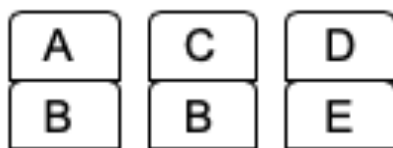
2-Variable Task



3-Variable Task



The testing materials for Study 2a, 2.3, 2.4, 3, & 4.3



The counterbalancing pattern for Study 2a, 2.3, 2.4, & 4.3

Set number	ICE 1	CVS Task 1	CVS Task 2	CVS Task 3	CVS Task 4	ICE 2
1		 Left	 Middle	 Right	 Left	
2		 Left	 Left	 Right	 Right	
3		 Right	 Left	 Middle	 Right	
4		 Right	 Right	 Middle	 Left	

Set number	ICE 1	CVS Task 1	CVS Task 2	CVS Task 3	CVS Task 4	ICE 2
5		 Left	 Left	 Right	 Middle	
6		 Left	 Right	 Right	 Left	
7		 Right	 Right	 Middle	 Left	
8		 Middle	 Left	 Left	 Right	

Appendix B: Protocol for Study 1 (original in German; *translated to English*)

PROTOCOL

Getting ready for the study

- Arrange the testing room so that you are sitting opposite the child
- Use an extra chair to place the materials
- Arrange the camera so that the child is fully in the frame and the materials they interact with are also visible
- If a teacher wants to stay in the room during the experiment, they should be out of view of the child (i.e., behind them)

Warm-up	
Introduce yourself, play warm-up game. This is an opportunity to make the child feel comfortable and at ease. Engage in small talk, encourage their performance, generally be friendly and enthusiastic.	
Schau, hier kannst du dich hinsetzen und dann können wir starten. <i>Here, have a seat and then we'll get started.</i>	
Ich heiße (Name) und ich habe ein paar Spiele dabei, die wir heute spielen können. <i>My name is (Name) and I brought a couple games with me that we can play.</i>	
So, lass uns doch mit dem Puzzle anfangen. <i>So, let's get started with a puzzle.</i>	
Ich glaube wir müssen die Tiere zu ihren Kindern zuordnen. <i>I think we have to match the animals to their babies.</i>	
Siehst du welche, die zusammenpassen? <i>Do you see any that go together?</i>	Let the child do the matching, simply encourage or comment on actions.
	Exactly, ah you found them, good job, hm what do we have left, etc.

<p>Jetzt sind wir mit diesem Spiel schon fertig, aber ich habe auch noch was anderes mit dabei. <i>Now we're done with this game, but I have something else with me.</i></p>	
<p>Aber ich würde sagen, wir räumen noch mal kurz dieses Puzzle hier auf. Dann haben wir ganz viel Platz. <i>But first let's clear away the puzzle so we have some room.</i></p>	

Vision test	
Give the child the color vision test	
<p>Jetzt habe ich ein kurzes Spiel auf dem Tablet. <i>I have a short game we can play on the tablet</i></p>	Present tablet to child in landscape orientation
<p>Schau, hier gibt's ein Kreis und da drin ist eine Linie. <i>Look, here is a circle with a line inside it.</i></p>	
<p>Du musst diese Linie mit deinem Finger folgen. <i>You have to trace the line with your finger.</i></p>	
<p>Genau! Jetzt kommen noch sieben Kreisen. <i>Exactly, now there will be seven more circles like this.</i></p>	Move to the next circle when the child has finished tracing the line.
<p>Gut gemacht! Das war es schon für diese Spiel. <i>Good job! We're already finished with this game.</i></p>	Record color vision score and any notes about difficulties completing the task.

Familiarization	
<p>In the familiarization phase there are bricks of four different colors. You will introduce the bricks, point out the different colors, and show that the bricks can be stuck together and taken apart. You will then allow the child to play independently with the bricks for up to 60 seconds.</p>	
<p>Schau mal, ich habe ein paar Lego Steine mit denen wir spielen können <i>Look, I have a couple Lego bricks we can play with.</i></p>	<p>Show child the 4 individual Lego bricks</p>
<p>Was für Farben sind sie? <i>What colors are they?</i></p>	<p>Let the child respond, if they don't then say the colors yourself</p>
<p>(Es gibt dunkel Blau, Gelb, hell Blau Glitter, und Weiß Glitter) <i>There's dark blue, yellow, light blue with glitter, and white with glitter.</i></p>	<p>Point out the different colors</p>
<p>Und schau mal, wir können die Steine zusammensetzen und wieder auseinanderziehen. <i>And look, we can stick them together and pull them apart again.</i></p>	<p>Demonstrate putting them together, pulling them apart</p>
<p>Hier, du kannst mit den Legos spielen <i>Here, you can play with the Legos.</i></p>	<p>Give the Legos to the child</p>
	<p>Allow child to play as long as they like up to 60s</p>

Training	
<p>In the base-rate training phase, the goal is to introduce the lightbox to the child and to show them that some things make the box light up and some things don't. (And to get used to placing the Legos on the box themselves)</p>	
<p>Jetzt habe ich diese Kiste mit dabei. Es ist eine ganz besondere Kiste. Es ist eine Leuchtkiste. Außen an der Kiste sind diese Streifen, die leuchten können, wenn man bestimmte Dinge auf die Kiste legt.</p>	<p>Introduce the box</p>

<p>Manche Dinge bringen die Kiste zum Leuchten und manche nicht. Wollen wir damit spielen? Kuch mal. Vielleicht können wir die Legos oben draufstellen. Sollen wir das mal ausprobieren? Dann schau mal nach.</p> <p><i>I also have this box with me. It is a special box, a lightbox. On the top of this box there's this stripe that can light up when certain things are placed on the box. Some things make the box light up and some things don't make the box light up. Shall we play with it? Let's see, maybe we can place the Legos on top. Shall we try it?</i></p>	
<p>Ah! Da leuchtet die Kiste! Der (Farbe) Stein bringt die Kiste zum Leuchten! Magst du das selber auch machen? Schau mal, dann nehme ich die weg und du kannst selber weiter ausprobieren.</p> <p><i>Ah! The box lights up. The (color) brick makes the box light up. Do you want to try? Watch, I'll take this one away and you can try the rest.</i></p>	<p>Place the individual Lego bricks on the machine one at a time First and Third or First and Fourth make it light up</p>
<p>Ah! Da leuchtet die Kiste ja gar nicht! Der (Farbe) Stein bringt die Kiste nicht zum Leuchten! Magst du noch ein anderes Lego probieren?</p> <p><i>Ah! The box doesn't light up. The (color) brick does not make the box light up. Do you want to try another one?</i></p>	
<p>Ah! Da leuchtet die Kiste! Der (Farbe) Stein bringt die Kiste zum Leuchten!</p> <p><i>Ah! The box lights up. The (color) brick makes the box light up.</i></p> <p>OR</p> <p>Hm! Da leuchtet die Kiste ja gar nicht! Der (Farbe) Stein bringt die Kiste nicht zum Leuchten!</p> <p><i>Ah! The box doesn't light up. The (color) brick does not make the box light up.</i></p>	
<p>Ah! Da leuchtet die Kiste! Der (Farbe) Stein bringt die Kiste zum Leuchten!</p> <p><i>Ah! The box lights up. The (color) brick makes the box light up.</i></p> <p>OR</p>	

<p>Hm! Da leuchtet die Kiste ja gar nicht! Der (Farbe) Stein bringt die Kiste nicht zum Leuchten! <i>Ah! The box doesn't light up. The (color) brick does not make the box light up.</i></p>	
<p>Also, erinnerst du welche Farben von Legos die Kiste zum Leuchten bringen? <i>So, do you remember which bricks made the box light up?</i></p>	<p>Allow child to point out the lighters/nonlighters If they don't remember, have them test again Make sure you also remember which ones work and which don't!</p>

Combined Legos	
<p>In the combined Legos phase, the goal is to show children how the bricks behave when they are combined.</p>	
<p>Hm aber was könnte passieren wenn wir einen von denen zusammen mit einen von denen machen? <i>Hm, but what happens when we stick one of those bricks together with one of those bricks?</i></p>	<p>Point at a lighter and a non-lighter. Allow the child to stick them together, But you need to place it on the box *ALWAYS place the Lego sticks horizontally*</p>
<p>Sollen wir das dann probieren? Schau mal was passiert. <i>Let's try it and see what happens.</i></p>	<p>Box lights up</p>
<p>Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i></p>	
<p>Lassen uns noch eins probieren. Was passiert, wenn wir zwei von denen zusammenmachen? <i>Let's try another one. What happens when we stick two of those together?</i></p>	<p>Point at a lighter and a lighter. Allow the child to stick them together, But you need to place it on the box</p>
<p>Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i></p>	<p>Box lights up</p>
<p>Und probieren wir ein letztes Mal. Was passiert, wenn wir zwei von denen zusammenmachen?</p>	<p>Point at a non-lighter and a non-lighter. Allow the child to stick</p>

<i>And let's try one more. What happens when we stick two of those together?</i>	them together, But you need to place it on the box
Ah! Da leuchtet die Kiste ja gar nicht! <i>Ah! The box doesn't light up.</i>	Box doesn't light up
Also, erinnerst du welche Farben von Legos die Kiste zum Leuchten bringen? <i>So, do you remember which bricks made the box light up?</i>	Check that they still understand which ones work
Also jetzt wissen wir, dass manche Legos die Kiste zum Leuchten bringen und manche nicht. <i>So, now we know that some Legos make the box light up and some don't.</i>	
Okay, jetzt räumen wir diese Legos zur Seite <i>Okay, let's clear these away now.</i>	Remove the training bricks

Understanding of confounded evidence	
<p>In this phase, we want to determine the child's understanding of confounded evidence. You will place a stick of four bricks on the box and the box will light up. You will ask the child if they know what color Legos make the box light up.</p>	
Schau mal, hier habe ich ein Lego Stange mit vier Steinen <i>Look, here I have a stick of four bricks.</i>	Show child the new object (with four bricks)
Aber diese Legos kann man nicht auseinandernehmen. Sie sind fest zusammengesteckt. <i>But you cannot take these bricks apart. They are stuck together.</i>	
Sollen wir das mal auf die Kiste ausprobieren? <i>Should we try this on the box?</i>	Place on the box; Box lights up
Wow! Die Kiste leuchtet! <i>Wow, the box lights up.</i>	
Weißt du welche Farbe von Legos die Kiste zum Leuchten bringt? <i>Do you know which bricks make the box light up?</i>	Allow child to answer (Yes or No)

Ja, Ich weiß (<i>Yes I know</i>)	Nein, Ich weiß nicht / Ich rate (<i>No, I don't know / I guess</i>)	
Welche Legos bringen die Kisten zum Leuchten? <i>Which make the box light up?</i>	Warum weißt du das nicht? <i>Why don't you know that?</i>	Wait for response
Weißt du es sicher oder rätst du nur? <i>Do you know that for sure or are you guessing?</i>		Wait for response
Woher weißt du das? <i>How do you know?</i>		Wait for response
Okay, dann lassen wir den kurz zur Seite <i>Okay, let's clear these away now.</i>		Remove the object, place out of sight

CVS: 2-variable Task	
In the CVS choice task, children will be asked to find out if one of the Legos makes the box light up. They will be shown that one stick lights up and then asked to choose one of the two sticks to test to determine if the X Lego makes the box light up. The correct choice controls variables, i.e., varies the color in question and keeps the other color the same.	
Jetzt können wir ein Spiel spielen. <i>Now we can play a game.</i>	
Hier ist eine Lego Stange mit zwei Steinen <i>Here is a stick with two bricks</i>	Just show the one you will test The choices should be prepared on the tray
Sie sind auch fest zusammengesteckt. <i>These are also stuck together.</i>	
Jetzt probiere ich die mal auf die Kiste <i>Now I will try it on the box.</i>	Place the first stick on the box
Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i>	

Also jetzt kommt das Ziel des Spiels <i>So, now, the goal of the game:</i>		
Wir wollen herausfinden ob der X Lego die Kiste zum Leuchten bringen <i>We want to find out if the X brick makes the box light up.</i>		Top Lego
Hier sind zwei Stangen. Die Regel dabei ist, dass du nur eine Stange wählen und auf der Kiste legen darfst, um herauszufinden ob der X Lego die Kiste zum Leuchten bringt. <i>Here are two more sticks. The rule is that you can only pick one stick to try on the box to find out if the X brick makes the box light up.</i>		Pull the box out of reach of the child Place the tray with the two sticks in front of the child
Welche Stange willst du dann wählen? <i>Which stick do you want to pick?</i>		Let the child pick one of the Lego sticks
Warum hast du den ausgewählt? <i>Why did you pick this stick?</i>		After they answer: Remove the tray with the second stick
Okay, dann probiere ich den auf die Kiste <i>Okay, then I'll place it on the box.</i>		Place the stick on the box
Ah da leuchtet die Kiste gar nicht! <i>Ah, the box does not light up.</i>		Wait for any explanation from the child
Weißt du jetzt, ob der X Legos die Kiste zum Leuchten bringen oder weißt du nicht? <i>Do you know if the X brick makes the box light up, or do you not know?</i>		Wait for response
Ja/ Ich weiß (<i>Yes / I know</i>)	Nein/ Ich weiß nicht (<i>No / I don't know</i>)	
Weißt du es sicher oder rätst du nur? <i>Do you know for sure or are you guessing?</i>	Warum weißt du das nicht? <i>Why don't you know that?</i>	Wait for response
Woher weißt du das?		Wait for response

<i>How do you know that?</i>	
Okay, super. Jetzt können wir diese Legos weglegen. <i>Okay, great, then we'll clear these Legos away.</i>	Remove Legos

CVS: 3-variable Task	
Schau mal, Hier ist eine Lego Stange mit drei Steinen <i>Look, here is a stick with three bricks.</i>	
Jetzt probiere ich die mal auf die Kiste <i>Now I will try it on the box.</i>	Place the first stick on the box
Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i>	
Also jetzt kommt das Ziel des Spiels <i>So, now, the goal of the game:</i>	
Wir wollen herausfinden ob der X Lego die Kiste zum Leuchten bringen <i>We want to find out if the X brick makes the box light up.</i>	Middle Lego
Diesmal gibt es drei Stangen. Wie vorher, die Regeln dabei lauten, dass du nur eine Stange wählen und auf der Kiste testen darfst, um herauszufinden ob der X Lego die Kisten zum Leuchten bringt. <i>This time, there are three more sticks. Like before, the rule is that you can only pick one stick to try on the box to find out if the X brick makes the box light up.</i>	Pull the box out of reach of the child Place the tray with the three sticks in front of the child Place the test Lego between you and the tray
Welche Stange willst du dann wählen? <i>Which stick do you want to pick?</i>	Let the child pick one of the Lego sticks
Warum hast du den ausgewählt? <i>Why did you pick this stick?</i>	After they answer: Remove the tray with the other sticks

Okay, dann probiere ich den auf die Kiste <i>Okay, then I'll place it on the box.</i>		Place the stick on the box
Ah da leuchtet die Kiste gar nicht! <i>Ah, the box does not light up.</i>		Wait for any explanation from the child
Weißt du jetzt, ob der X Legos die Kiste zum Leuchten bringen oder weißt du nicht? <i>Do you know if the X brick makes the box light up, or do you not know?</i>		Wait for response
Ja/ Ich weiß (<i>Yes / I know</i>)	Nein/ Ich weiß nicht (<i>No / I don't know</i>)	
Weißt du es sicher oder rätst du nur? <i>Do you know for sure or are you guessing?</i>	Warum weißt du das nicht? <i>Why don't you know that?</i>	Wait for response
Woher weißt du das? <i>How do you know that?</i>		Wait for response
Okay, super. Jetzt sind wir mit den Spielen schon fertig <i>Okay, great, we are finished with all the games!</i>		Remove Legos

Appendix C: Summary of results from Studies 1, 2a, 2b, & 4

Task	Study 1			Study 2a		Study 2b		Study 4	
	M = 60 months N = 60			M = 65 months N = 51		M = 65 months N = 57		M = 48 months N = 187	
	Session 1	Session 2		Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2
ICE	58%	49%		63%	46%	47%	32%	55%	54%
Knowledge Claims	% correct								
	Across 2 trials	41-25-34		38-32-30		23-32-46		42-26-31	
	Effects	-		Trial		Age		-	
Robust	% correct	15%	18%	37%	14%	26%	18%	19%	8%
	Across 2 trials	7-19-74		10-30-60		16-12-72		6-14-80	
	Effects	-		Trial		Age Trial		Trial	
CVS	% correct	52%	72%*	63%†	55%	70%*	61%†	67%*	57%†
2-variable	Across 2 trials	40-43-17*		35-47-18		46-40-14*		36-52-12*	
	% correct	39%	45%†	51%*	45%†	47%*	39%	39%	37%
3-variable	Across 2 trials	23-39-38*		24-49-27*		17-51-32†		17-41-42*	
All CVS tasks	Effects	Age Session		-		Task		Task Gender	
Justifications	Effects	23%		26%		38%		11%	
		Age Choice		Age Choice MLUw		Age Choice MLUw Task		Choice Task Robust ICE	

Notes. † $p < .10$, * $p < .05$; x-y-z: x = 2 correct, y = 1 correct, z = 0

Appendix D: Changes to Protocol for Study 2a & 3 (original in German; translated to English; changes from previous protocol highlighted)

Training	
<p>In the base-rate training phase, the goal is to introduce the lightbox to the child and to show them that some things make the box light up and some things don't.</p>	
<p>Jetzt habe ich diese Kiste mit dabei. Es ist eine ganz besondere Kiste. Es ist eine Leuchtkiste. Außen an der Kiste sind diese Streifen, die leuchten können, wenn man bestimmte Dinge auf die Kiste legt. Manche Dinge bringen die Kiste zum Leuchten und manche nicht. Wollen wir damit spielen? Kuch ma mal. Vielleicht können wir die Legos oben draufstellen. Sollen wir das mal ausprobieren? Dann schau mal nach.</p> <p><i>I also have this box with me. It is a special box, a lightbox. On the top of this box there's this stripe that can light up when certain things are placed on the box. Some things make the box light up and some things don't make the box light up. Shall we play with it? Let's see, maybe we can place the Legos on top. Shall we try it?</i></p>	<p>Introduce the box</p> <p>Arrange the bricks in following order: White, light blue, dark blue, yellow</p>
<p>Ah! Da leuchtet die Kiste! Der Weiß Glitter Stein bringt die Kiste zum Leuchten! Das ist ein Toma. Tomas bringen die Kiste zum Leuchten. Schau mal, dann nehme ich die weg und du kannst weiter ausprobieren.</p> <p><i>Ah! The box lights up. The white glitter brick makes the box light up. This is a Toma. Tomas make the box light up. I'll take this one away and you can try the rest.</i></p>	<p>Let child place the individual Lego bricks on the machine one at a time</p> <p>White glitter lights up</p>
<p>Ah! Da leuchtet die Kiste ja gar nicht! Der hell Blaue Stein bringt die Kiste nicht zum Leuchten! Das ist nicht ein Toma. Es bringet die Kiste nicht zum Leuchten.</p> <p>Magst du noch ein anderes Lego probieren?</p> <p><i>Ah! The box doesn't light up. The light blue brick does not make the box light up. This is not a Toma, it does not make the box light up. Do you want to try another one?</i></p>	<p>Blue glitter does not light up</p>

<p>Ah! Da leuchtet die Kiste! Der dunkel blaue Stein bringt die Kiste zum Leuchten! Das ist ein Toma. <i>Ah! The box lights up. The dark blue brick makes the box light up. This is a Toma.</i></p>	<p>Dark blue lights up</p>
<p>Ah! Da leuchtet die Kiste ja gar nicht! Der gelbe Stein bringt die Kiste nicht zum Leuchten! Das ist dann nicht ein Toma. <i>Ah! The box doesn't light up. The yellow brick does not make the box light up. This is not a Toma.</i></p>	<p>Yellow does not light up</p>
<p>Also, erinnerst du welche Farben von Legos Tomas sind? <i>So, do you remember which are Tomas?</i></p>	<p>Allow child to point out the lighters/nonlighters If they don't remember, have them test again Make sure you also remember which ones work and which don't!</p>

<p>Combined Legos</p>	
<p>In the combined Legos phase, the goal is to show children how the bricks behave when they are combined.</p>	
<p>Hm aber was könnte passieren wenn wir ein Toma zusammen mit ein nicht-Toma machen? <i>Hm, but what happens when we stick a Toma together with a not-Toma?</i></p>	<p>Point at a lighter and a non-lighter. Allow the child to stick them together, But you need to place it on the box *ALWAYS place the Lego sticks horizontally*</p>
<p>Sollen wir das dann probieren? Schau mal was passiert. <i>Let's try it and see what happens.</i></p>	<p>Box lights up</p>
<p>Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i></p>	
<p>Lassen uns noch eins probieren. Was passiert, wenn wir den anderen Toma zusammen mit den anderen nicht-Toma machen? <i>Let's try another one. What happens when we stick the other Toma together with the other not-Toma?</i></p>	<p>Point at the other lighter and the other non-lighter. Allow the child to stick them together, But you need to place it on the box</p>

Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i>	
Lassen uns noch eins probieren. Was passiert, wenn wir zwei Tomas zusammenmachen? <i>Let's try another one. What happens when we stick two Tomas together?</i>	Point at a lighter and a lighter. Allow the child to stick them together, But you need to place it on the box
Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i>	Box lights up
Und probieren wir ein letztes Mal. Was passiert, wenn wir zwei nicht-Tomas zusammenmachen? <i>And let's try one more. What happens when we stick two not-Tomas together?</i>	Point at a non-lighter and a non-lighter. Allow the child to stick them together, But you need to place it on the box
Ah! Da leuchtet die Kiste ja gar nicht! <i>Ah! The box doesn't light up.</i>	Box doesn't light up
Also, erinnerst du welche Farben von Legos Tomas sind? <i>So, do you remember which bricks are Tomas?</i>	Check that they still understand which ones work
Also jetzt wissen wir, dass manche Legos sind Tomas und manche nicht. <i>So, now we know that some are Tomas and some are not.</i>	
Okay, jetzt räumen wir diese Legos zur Seite <i>Okay, let's clear these away now.</i>	Remove the training bricks

Understanding of confounded evidence	
In this phase, we want to determine the child's understanding of confounded evidence. You will place a stick of four bricks on the box and the box will light up. You will ask the child if they know what color Legos make the box light up.	
Schau mal, hier habe ich ein Lego Stange mit vier Steinen <i>Look, here I have a stick of four bricks.</i>	Show child the new object (with four bricks)
Aber diese Legos kann man nicht auseinandernehmen. Sie sind fest zusammengesteckt. <i>But you cannot take these bricks apart. They are stuck together.</i>	

Sollen wir das mal auf die Kiste ausprobieren? <i>Should we try this on the box?</i>		Place on the box; Box lights up
Wow! Die Kiste leuchtet! <i>Wow, the box lights up.</i>		
Kannst du sicher wissen welche Legos sind Tomas, oder kannst du das nicht sicher wissen? <i>Can you know for sure which bricks are Tomas or can you not know for sure?</i>		Allow child to answer (Yes or No)
Ja, Ich weiß (<i>Yes I know</i>)	Nein, Ich weiß nicht / Ich rate (<i>No, I don't know / I guess</i>)	
Weißt du es sicher oder rätst du nur? <i>Do you know that for sure or are you guessing?</i>	Warum weißt du das nicht? <i>Why don't you know that?</i>	Wait for response
Woher weißt du das? <i>How do you know?</i>		Wait for response
Okay, dann lassen wir den kurz zur Seite <i>Okay, let's clear these away now.</i>		Remove the object, place out of sight

CVS: 2-variable Task	
In the CVS choice task, children will be asked to find out if one of the Legos makes the box light up. They will be shown that one stick lights up and then asked to choose one of the two sticks to test to determine if the X Lego makes the box light up. The correct choice controls variables, i.e., varies the color in question and keeps the other color the same.	
Jetzt können wir ein Spiel spielen. <i>Now we can play a game.</i>	
Hier ist eine Lego Stange mit zwei Steinen <i>Here is a stick with two bricks</i>	Just show the one you will test The choices should be prepared on the tray

<p>Sie sind auch fest zusammengesteckt. <i>These are also stuck together.</i></p>		
<p>Jetzt probiere ich die mal auf die Kiste <i>Now I will try it on the box.</i></p>		Place the first stick on the box
<p>Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i></p>		
<p>Also jetzt kommt das Ziel des Spiels <i>So, now, the goal of the game:</i></p>		
<p>Wir wollen herausfinden ob der X Lego ein Toma ist. <i>We want to find out if the X brick is a Toma..</i></p>		Top Lego
<p>Hier sind zwei Stangen. Die Regel dabei ist, dass du nur eine Stange wählen und auf der Kiste legen darfst, um herauszufinden ob der X Lego ein Toma ist. <i>Here are two more sticks. The rule is that you can only pick one stick to try on the box to find out if the X is a Toma.</i></p>		<p>Pull the box out of reach of the child</p> <p>Place the tray with the two sticks in front of the child</p>
<p>Welche Stange ist die beste, um herauszufinden ob der X Lego ein Toma ist? <i>Which stick is the best to find out if the X brick is a Toma?</i></p>		Let the child pick one of the Lego sticks
<p>Warum ist diese Stange die beste, um das herauszufinden? <i>Why is this the best stick to find that out?</i></p>		After they answer: Remove the tray with the second stick
<p>Okay, dann probiere ich den auf die Kiste <i>Okay, then I'll place it on the box.</i></p>		Place the stick on the box
<p>Ah da leuchtet die Kiste gar nicht! <i>Ah, the box does not light up.</i></p>		Wait for any explanation from the child
<p>Weißt du jetzt, ob der X Lego ein Toma ist oder weißt du nicht? <i>Do you know if the X brick is a Toma, or do you not know?</i></p>		Wait for response
Ja/ Ich weiß (<i>Yes / I know</i>)	Nein/ Ich weiß nicht (<i>No / I don't know</i>)	

<p>Weißt du es sicher oder rätst du nur? <i>Do you know for sure or are you guessing?</i></p>	<p>Warum weißt du das nicht? <i>Why don't you know that?</i></p>	<p>Wait for response</p>
<p>Woher weißt du das? <i>How do you know that?</i></p>		<p>Wait for response</p>
<p>Okay, super. Jetzt können wir diese Legos weglegen. <i>Okay, great, then we'll clear these Legos away.</i></p>		<p>Remove Legos</p>

CVS: 3-variable Task	
<p>Schau mal, Hier ist eine Lego Stange mit drei Steinen <i>Look, here is a stick with three bricks.</i></p>	
<p>Jetzt probiere ich die mal auf die Kiste <i>Now I will try it on the box.</i></p>	<p>Place the first stick on the box</p>
<p>Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i></p>	
<p>Also jetzt kommt das Ziel des Spiels <i>So, now, the goal of the game:</i></p>	
<p>Wir wollen herausfinden ob der X Lego ein Toma ist. <i>We want to find out if the X brick is a Toma.</i></p>	<p>Middle Lego</p>
<p>Diesmal gibt es drei Stangen. Wie vorher, die Regeln dabei lauten, dass du nur eine Stange wählen und auf der Kiste testen darfst, um herauszufinden ob der X Lego ein Toma ist. <i>This time, there are three more sticks. Like before, the rule is that you can only pick one stick to try on the box to find out if the X brick is a Toma.</i></p>	<p>Pull the box out of reach of the child Place the tray with the three sticks in front of the child Place the test Lego between you and the tray</p>
<p>Welche Stange ist die beste, um herauszufinden ob der X Lego ein Toma ist? <i>Which stick is the best to find out if the X brick is a Toma?</i></p>	<p>Let the child pick one of the Lego sticks</p>

Warum ist diese Stange die beste, um das herauszufinden? <i>Why is this the best stick to find that out?</i>		After they answer: Remove the tray with the other sticks
Okay, dann probiere ich den auf die Kiste <i>Okay, then I'll place it on the box.</i>		Place the stick on the box
Ah da leuchtet die Kiste gar nicht! <i>Ah, the box does not light up.</i>		Wait for response
Weißt du jetzt, ob der X Lego ein Toma ist oder weißt du nicht? <i>Do you know if the X brick is a Toma, or do you not know?</i>		Wait for response
Ja/ Ich weiß (<i>Yes / I know</i>)	Nein/ Ich weiß nicht (<i>No / I don't know</i>)	
Weißt du es sicher oder rätst du nur? <i>Do you know for sure or are you guessing?</i>	Warum weißt du das nicht? <i>Why don't you know that?</i>	Wait for response
Woher weißt du das? <i>How do you know that?</i>		Wait for response
Okay, super. Jetzt sind wir mit den Spielen schon fertig <i>Okay, great, we are finished with all the games!</i>		Remove Legos

Appendix E: Changes to Protocol for Study 2b (original in German; translated to English; changes from previous protocol highlighted)

Combined Legos	
In the combined Legos phase, the goal is to show children how the bricks behave when they are combined.	
Hm aber was könnte passieren wenn wir ein Toma zusammen mit ein nicht-Toma machen? <i>Hm, but what happens when we stick a Toma together with a not-Toma?</i>	Point at a lighter and a non-lighter. Allow the child to stick them together, But you need to place it on the box *ALWAYS place the Lego sticks horizontally*
Sollen wir das dann probieren? Schau mal was passiert. <i>Let's try it and see what happens.</i>	Box lights up
Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i>	
Lassen uns noch eins probieren. Was passiert, wenn wir den anderen Toma zusammen mit den anderen nicht-Toma machen? <i>Let's try another one. What happens when we stick the other Toma together with the other not-Toma?</i>	Point at the other lighter and the other non-lighter. Allow the child to stick them together, But you need to place it on the box
Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i>	
Lassen uns noch eins probieren. Was passiert, wenn wir zwei Tomas zusammenmachen? <i>Let's try another one. What happens when we stick two Tomas together?</i>	Point at a lighter and a lighter. Allow the child to stick them together, But you need to place it on the box
Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i>	Box lights up
Was passiert, wenn wir zwei nicht-Tomas zusammenmachen? <i>And let's try one more. What happens when we stick two not-Tomas together?</i>	Point at a non-lighter and a non-lighter. Allow the child to stick them together, But you need to place it on the box
Ah! Da leuchtet die Kiste ja gar nicht!	Box doesn't light up

<i>Ah! The box doesn't light up.</i>	
Und probieren wir ein letztes Mal. Was passiert, wenn wir alle vier zusammenmachen? <i>And let's try one more. What happens when we stick all four together?</i>	
Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i>	Box lights up
Also, erinnerst du welche Farben von Legos Tomas sind? <i>So, do you remember which bricks are Tomas?</i>	Check that they still understand which ones work
Also jetzt wissen wir, dass manche Legos sind Tomas und manche nicht. <i>So, now we know that some are Tomas and some are not.</i>	
Okay, jetzt räumen wir diese Legos zur Seite <i>Okay, let's clear these away now.</i>	Remove the training bricks

CVS: 2-variable Task	
In the CVS choice task, children will be asked to find out if one of the Legos makes the box light up. They will be shown that one stick lights up and then asked to choose one of the two sticks to test to determine if the X Lego makes the box light up. The correct choice controls variables, i.e., varies the color in question and keeps the other color the same.	
Jetzt können wir ein Spiel spielen. <i>Now we can play a game.</i>	
Hier ist eine Lego Stange mit zwei Steinen <i>Here is a stick with two bricks</i>	Just show the one you will test The choices should be prepared on the tray
Sie sind auch fest zusammengesteckt. <i>These are also stuck together.</i>	
Jetzt probiere ich die mal auf die Kiste <i>Now I will try it on the box.</i>	Place the first stick on the box

Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i>		
Also jetzt kommt das Ziel des Spiels <i>So, now, the goal of the game:</i>		
Wir wollen herausfinden ob der X Lego ein Toma ist. <i>We want to find out if the X brick is a Toma..</i>		Top Lego
Hier sind zwei Stangen. Die Regel dabei ist, dass du nur eine Stange wählen und auf der Kiste legen darfst, um herauszufinden ob der X Lego ein Toma ist. <i>Here are two more sticks. The rule is that you can only pick one stick to try on the box to find out if the X is a Toma.</i>		Pull the box out of reach of the child Place the tray with the two sticks in front of the child
Welche Stange ist die beste, um herauszufinden ob der X Lego ein Toma ist? <i>Which stick is the best to find out if the X brick is a Toma?</i>		Let the child pick one of the Lego sticks
Warum ist diese Stange die beste, um das herauszufinden? <i>Why is this the best stick to find that out?</i>		After they answer: Remove the tray with the second stick
Okay, dann probiere ich den auf die Kiste <i>Okay, then I'll place it on the box.</i>		Place the stick on the box
Ah da leuchtet die Kiste gar nicht! <i>Ah, the box does not light up.</i>		Wait for any explanation from the child
Also, ist der X Lego ein Toma, ein nicht-Toma, oder kannst du das nicht wissen? <i>So, is the X brick a Toma, a not-Toma, or can you not know?</i>		Wait for response
Ja/ Ich weiß (<i>Yes / I know</i>)	Nein/ Ich weiß nicht (<i>No / I don't know</i>)	
Weißt du es sicher oder rätst du nur? <i>Do you know for sure or are you guessing?</i>	Warum weißt du das nicht? <i>Why don't you know that?</i>	Wait for response

Woher weißt du das? <i>How do you know that?</i>		Wait for response
Okay, super. Jetzt können wir diese Legos weglegen. <i>Okay, great, then we'll clear these Legos away.</i>		Remove Legos

CVS: 3-variable Task		
Schau mal, Hier ist eine Lego Stange mit drei Steinen <i>Look, here is a stick with three bricks.</i>		
Jetzt probiere ich die mal auf die Kiste <i>Now I will try it on the box.</i>		Place the first stick on the box
Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i>		
Also jetzt kommt das Ziel des Spiels <i>So, now, the goal of the game:</i>		
Wir wollen herausfinden ob der X Lego ein Toma ist. <i>We want to find out if the X brick is a Toma.</i>		Middle Lego
Diesmal gibt es drei Stangen. Wie vorher, die Regeln dabei lauten, dass du nur eine Stange wählen und auf der Kiste testen darfst, um herauszufinden ob der X Lego ein Toma ist. <i>This time, there are three more sticks.</i> <i>Like before, the rule is that you can only pick one stick to try on the box to find out if the X brick is a Toma.</i>		Pull the box out of reach of the child Place the tray with the three sticks in front of the child Place the test Lego between you and the tray
Welche Stange ist die beste, um herauszufinden ob der X Lego ein Toma ist? <i>Which stick is the best to find out if the X brick is a Toma?</i>		Let the child pick one of the Lego sticks
Warum ist diese Stange die beste, um das herauszufinden? <i>Why is this the best stick to find that out?</i>		After they answer: Remove the tray with the other sticks

Okay, dann probiere ich den auf die Kiste <i>Okay, then I'll place it on the box.</i>		Place the stick on the box
Ah da leuchtet die Kiste gar nicht! <i>Ah, the box does not light up.</i>		Wait for any explanation from the child
Also, ist der X Lego ein Toma, ein nicht-Toma, oder kannst du das nicht wissen? <i>So, is the X brick a Toma, a not-Toma, or can you not know?</i>		Wait for response
Ja/ Ich weiß (<i>Yes / I know</i>)	Nein/ Ich weiß nicht (<i>No / I don't know</i>)	
Weißt du es sicher oder rätst du nur? <i>Do you know for sure or are you guessing?</i>	Warum weißt du das nicht? <i>Why don't you know that?</i>	Wait for response
Woher weißt du das? <i>How do you know that?</i>		Wait for response
Okay, super. Jetzt sind wir mit den Spielen schon fertig <i>Okay, great, we are finished with all the games!</i>		Remove Legos

Appendix F: Full protocol for Study 4 (original in German; translated to English)

PROTOCOL

Getting ready for the study

- Arrange the testing room so that you are sitting opposite the child
- Use an extra chair to place the materials
- Arrange the camera so that the child is fully in the frame and the materials they interact with are also visible
- If a teacher wants to stay in the room during the experiment, they should be out of view of the child (i.e., behind them)

Warm-up	
Introduce yourself, play warm-up game. This is an opportunity to make the child feel comfortable and at ease. Engage in small talk, encourage their performance, generally be friendly and enthusiastic.	
Schau, hier kannst du dich hinsetzen und dann können wir starten. <i>Here, have a seat and then we'll get started.</i>	
Ich heiße (Name) und ich habe ein paar Spiele dabei, die wir heute spielen können. <i>My name is (Name) and I brought a couple games with me that we can play.</i>	
So, lass uns doch mit dem Puzzle anfangen. <i>So, let's get started with a puzzle.</i>	
Ich glaube wir müssen die Tiere zu ihren Kindern zuordnen. <i>I think we have to match the animals to their babies.</i>	
Siehst du welche, die zusammenpassen? <i>Do you see any that go together?</i>	Let the child do the matching, simply encourage or comment on actions.
	Exactly, ah you found them, good job, hm what do we have left, etc.
Jetzt sind wir mit diesem Spiel schon fertig, aber ich habe auch noch was anderes mit dabei.	

<i>Now we're done with this game, but I have something else with me.</i>	
Aber ich würde sagen, wir räumen noch mal kurz dieses Puzzle hier auf. Dann haben wir ganz viel Platz. <i>But first let's clear away the puzzle so we have some room.</i>	

Vision test	
Give the child the color vision test	
Jetzt habe ich ein kurzes Spiel auf dem Tablet. <i>I have a short game we can play on the tablet</i>	Present tablet to child in landscape orientation
Schau, hier gibt's ein Kreis und da drin ist eine Linie. <i>Look, here is a circle with a line inside it.</i>	
Du must diese Linie mit deinem Finger folgen. <i>You have to trace the line with your finger.</i>	
Genau! Jetzt kommen noch sieben Kreisen. <i>Exactly, now there will be seven more circles like this.</i>	Move to the next circle when the child has finished tracing the line.
Gut gemacht! Das war es schon für diese Spiel. <i>Good job! We're already finished with this game.</i>	Record color vision score and any notes about difficulties completing the task.

Familiarization
In the familiarization phase there are bricks of four different colors. You will introduce the bricks, point out the different colors, and show that the bricks can be stuck together and taken apart. You will then allow the child to play independently with the bricks for up to 60 seconds.

Schau mal, ich habe ein paar Lego Steine mit denen wir spielen können <i>Look, I have a couple Lego bricks we can play with.</i>	Show child the 4 individual Lego bricks
Was für Farben sind sie? <i>What colors are they?</i>	Let the child respond, if they don't then say the colors yourself
(Es gibt dunkel Blau, Gelb, hell Blau Glitter, und Weiß Glitter) <i>There's dark blue, yellow, light blue with glitter, and white with glitter.</i>	Point out the different colors
Und schau mal, wir können die Steine zusammensetzen und wieder auseinanderziehen. <i>And look, we can stick them together and pull them apart again.</i>	Demonstrate putting them together, pulling them apart
Hier, du kannst mit den Legos spielen <i>Here, you can play with the Legos.</i>	Give the Legos to the child
	Allow child to play as long as they like up to 60s

Training	
In the base-rate training phase, the goal is to introduce the lightbox to the child and to show them that some things make the box light up and somethings don't. (And to get used to placing the Legos on the box themselves)	
Jetzt habe ich diese Kiste mit dabei. Es ist eine ganz besondere Kiste. Es ist eine Leuchtkiste. Außen an der Kiste sind diese Streifen, die leuchten können, wenn man bestimmte Dinge auf die Kiste legt. Manche Dinge bringen die Kiste zum Leuchten und manche nicht. Wollen wir damit spielen? Kuch mal. Vielleicht können wir die Legos oben draufstellen. Sollen wir das mal ausprobieren? Dann schau mal nach. <i>I also have this box with me. It is a special box, a lightbox. On the top of this box there's this stripe that can light up when certain things are placed on the box. Some things make the box light up and</i>	Introduce the box

<p><i>some things don't make the box light up. Shall we play with it? Let's see, maybe we can place the Legos on top. Shall we try it?</i></p>	
<p>Ah! Da leuchtet die Kiste! Der Weiß Glitter Stein bringt die Kiste zum Leuchten! Das ist ein Toma. Tomas bringen die Kiste zum Leuchten. Schau mal, dann nehme ich die weg und du kannst weiter ausprobieren. <i>Ah! The box lights up. The white glitter brick makes the box light up. This is a Toma. Tomas make the box light up. I'll take this one away and you can try the rest.</i></p>	<p>Let child place the individual Lego bricks on the machine one at a time White glitter lights up</p>
<p>Ah! Da leuchtet die Kiste ja gar nicht! Der hell Blaue Stein bringt die Kiste nicht zum Leuchten! Das ist nicht ein Toma. Es bringet die Kiste nicht zum Leuchten. Magst du noch ein anderes Lego probieren? <i>Ah! The box doesn't light up. The light blue brick does not make the box light up. This is not a Toma, it does not make the box light up. Do you want to try another one?</i></p>	<p>Blue glitter does not light up</p>
<p>Ah! Da leuchtet die Kiste! Der dunkel blaue Stein bringt die Kiste zum Leuchten! Das ist ein Toma. <i>Ah! The box lights up. The dark blue brick makes the box light up. This is a Toma.</i></p>	<p>Dark blue lights up</p>
<p>Ah! Da leuchtet die Kiste ja gar nicht! Der gelbe Stein bringt die Kiste nicht zum Leuchten! Das ist dann nicht ein Toma. <i>Ah! The box doesn't light up. The yellow brick does not make the box light up. This is not a Toma.</i></p>	<p>Yellow does not light up</p>
<p>Also, erinnerst du welche Farben von Legos Tomas sind? <i>So, do you remember which are Tomas?</i></p>	<p>Allow child to point out the lighters/nonlighters If they don't remember, have them test again Make sure you also remember which ones work and which don't!</p>

Combined Legos	
In the combined Legos phase, the goal is to show children how the bricks behave when they are combined.	
<p>Hm aber was könnte passieren wenn wir einen von denen zusammen mit einen von denen machen? <i>Hm, but what happens when we stick one of those bricks together with one of those bricks?</i></p>	<p>Point at a lighter and a non-lighter. Allow the child to stick them together, But you need to place it on the box *ALWAYS place the Lego sticks horizontally*</p>
<p>Sollen wir das dann probieren? Schau mal was passiert. <i>Let's try it and see what happens.</i></p>	Box lights up
<p>Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i></p>	
<p>Lassen uns noch eins probieren. Was passiert, wenn wir den anderen Toma zusammen mit den anderen nicht-Toma machen? <i>Let's try another one. What happens when we stick the other Toma together with the other not-Toma?</i></p>	<p>Point at the other lighter and the other non-lighter. Allow the child to stick them together, But you need to place it on the box</p>
<p>Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i></p>	
<p>Lassen uns noch eins probieren. Was passiert, wenn wir zwei Tomas zusammenmachen? <i>Let's try another one. What happens when we stick two Tomas together?</i></p>	<p>Point at a lighter and a lighter. Allow the child to stick them together, But you need to place it on the box</p>
<p>Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i></p>	Box lights up
<p>Was passiert, wenn wir zwei nicht-Tomas zusammenmachen? <i>And let's try one more. What happens when we stick two not-Tomas together?</i></p>	<p>Point at a non-lighter and a non-lighter. Allow the child to stick them together, But you need to place it on the box</p>
<p>Ah! Da leuchtet die Kiste ja gar nicht! <i>Ah! The box doesn't light up.</i></p>	Box doesn't light up

<p>Und probieren wir ein letztes Mal. Was passiert, wenn wir alle vier zusammenmachen? <i>And let's try one more. What happens when we stick all four together?</i></p>	
<p>Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i></p>	Box lights up
<p>Also, erinnerst du welche Farben von Legos Tomas sind? <i>So, do you remember which bricks are Tomas?</i></p>	Check that they still understand which ones work
<p>Also jetzt wissen wir, dass manche Legos sind Tomas und manche nicht. <i>So, now we know that some are Tomas and some are not.</i></p>	

Understanding of confounded evidence	
<p>In this phase, we want to determine the child's understanding of confounded evidence. You will place a stick of four bricks on the box and the box will light up. You will ask the child if they know what color Legos make the box light up.</p>	
<p>Schau mal, hier habe ich ein Lego Stange mit vier Steinen <i>Look, here I have a stick of four bricks.</i></p>	Show child the new object (with four bricks)
<p>Aber diese Legos kann man nicht auseinandernehmen. Sie sind fest zusammengesteckt. <i>But you cannot take these bricks apart. They are stuck together.</i></p>	
<p>Sollen wir das mal auf die Kiste ausprobieren? <i>Should we try this on the box?</i></p>	Place on the box; Box lights up
<p>Wow! Die Kiste leuchtet! <i>Wow, the box lights up.</i></p>	
<p>Weißt du welche Farbe von Legos die Kiste zum Leuchten bring oder weißt du nicht? <i>Do you know which bricks make the box light up?</i></p>	Allow child to answer (Yes or No)
<p>Ja, Ich weiß (<i>Yes I know</i>)</p>	Nein, Ich weiß nicht / Ich rate

	(No, I don't know / I guess)	
Welche Legos bringen die Kisten zum Leuchten? <i>Which make the box light up?</i>	Warum weißt du das nicht? <i>Why don't you know that?</i>	Wait for response
Weißt du es sicher oder rätst du nur? <i>Do you know that for sure or are you guessing?</i>		Wait for response
Woher weißt du das? <i>How do you know?</i>		Wait for response
Okay, dann lassen wir den kurz zur Seite <i>Okay, let's clear these away now.</i>		Remove the object, place out of sight

CVS: 2-variable Task	
In the CVS choice task, children will be asked to find out if one of the Legos makes the box light up. They will be shown that one stick lights up and then asked to choose one of the two sticks to test to determine if the X Lego makes the box light up. The correct choice controls variables, i.e., varies the color in question and keeps the other color the same.	
Jetzt können wir ein Spiel spielen. <i>Now we can play a game.</i>	
Hier ist eine Lego Stange mit zwei Steinen <i>Here is a stick with two bricks</i>	Just show the one you will test The choices should be prepared on the tray
Sie sind auch fest zusammengesteckt. <i>These are also stuck together.</i>	
Jetzt probiere ich die mal auf die Kiste <i>Now I will try it on the box.</i>	Place the first stick on the box
Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i>	

Also jetzt kommt das Ziel des Spiels <i>So, now, the goal of the game:</i>	
Wir wollen herausfinden ob der X Lego die Kiste zum Leuchten bringen <i>We want to find out if the X brick makes the box light up.</i>	Top Lego
Hier sind zwei Stangen. Die Regel dabei ist, dass du nur eine Stange wählen und auf der Kiste legen darfst, um herauszufinden ob der X Lego die Kiste zum Leuchten bringt. <i>Here are two more sticks. The rule is that you can only pick one stick to try on the box to find out if the X brick makes the box light up.</i>	Pull the box out of reach of the child Place the tray with the two sticks in front of the child
Welche Stange ist die beste, um herauszufinden ob der X Lego die Kiste zum Leuchten bringt? <i>Which stick is the best to find out if the X brick is a Toma?</i>	Let the child pick one of the Lego sticks
Warum ist diese Stange die beste, um herauszufinden, ob der X Lego die Kiste zum Leuchten bringt? <i>Why is this the best stick to find that out?</i>	After they answer: Remove the tray with the second stick
Okay, dann probiere ich den auf die Kiste <i>Okay, then I'll place it on the box.</i>	Place the stick on the box
Ah da leuchtet die Kiste gar nicht! <i>Ah, the box does not light up.</i>	Wait for any explanation from the child
Lass uns die nächste Stange anschauen/ausprobieren! <i>Let's try the next stick!</i>	

CVS: 3-variable Task	
Schau mal, Hier ist eine Lego Stange mit drei Steinen <i>Look, here is a stick with three bricks.</i>	
Jetzt probiere ich die mal auf die Kiste <i>Now I will try it on the box.</i>	Place the first stick on the box

<p>Ah! Da leuchtet die Kiste! <i>Ah! The box lights up.</i></p>	
<p>Also jetzt kommt das Ziel des Spiels <i>So, now, the goal of the game:</i></p>	
<p>Wir wollen herausfinden ob der X Lego die Kiste zum Leuchten bringen <i>We want to find out if the X brick makes the box light up.</i></p>	Middle Lego
<p>Diesmal gibt es drei Stangen. Wie vorher, die Regeln dabei lauten, dass du nur eine Stange wählen und auf der Kiste testen darfst, um herauszufinden ob der X Lego die Kisten zum Leuchten bringt. <i>This time, there are three more sticks. Like before, the rule is that you can only pick one stick to try on the box to find out if the X brick makes the box light up.</i></p>	<p>Pull the box out of reach of the child Place the tray with the three sticks in front of the child Place the test Lego between you and the tray</p>
<p>Welche Stange ist die beste, um herauszufinden ob der X Lego die Kiste zum Leuchten bringt? <i>Which stick is the best to find out if the X brick is a Toma?</i></p>	Let the child pick one of the Lego sticks
<p>Warum ist diese Stange die beste, um herauszufinden, ob der X Lego die Kiste zum Leuchten bringt? <i>Why is this the best stick to find that out?</i></p>	After they answer: Remove the tray with the second stick
<p>Okay, dann probiere ich den auf die Kiste <i>Okay, then I'll place it on the box.</i></p>	Place the stick on the box
<p>Ah da leuchtet die Kiste gar nicht! <i>Ah, the box does not light up.</i></p>	Wait for any explanation from the child
<p>Lass uns die nächste Stange anschauen/ausprobieren! <i>Let's try the next stick!</i></p>	

Appendix G: Bivariate correlations between variables of interest in Study 4

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Age	-																
2. Gender	-.01 (185)	-															
3. Intelligence	.12 (177)	-.07 (177)	-														
4. Language	.10 (168)	.03 (170)	.62*** (163)	-													
5. EF - Working Memory	.20* (146)	.07 (148)	.35*** (140)	.38*** (140)	-												
6. EF - Inhibition	-.07 (177)	-.06 (179)	.11 (169)	.16* (165)	.15† (143)	-											
7. EF - Planning	.03 (163)	.03 (165)	.26*** (158)	.21* (152)	.41*** (132)	.20* (158)	-										
8. EF - Cognitive Flex.	.18* (130)	.03 (130)	.32*** (128)	.33*** (119)	.40*** (105)	.09 (125)	.11 (116)	-									
9. ToM - Knowledge Acc.	.05 (133)	-.16† (133)	.13 (132)	.21* (124)	.09 (111)	.14 (129)	.18† (121)	.03 (100)	-								
10. ToM - Content FB	.00 (147)	-.02 (147)	.23** (147)	.11 (135)	.25** (119)	.07 (141)	.30*** (131)	.13 (107)	.24** (11)	-							
11. ToM - Explicit FB	.28** (133)	-.05 (133)	.33*** (132)	.34*** (121)	.11 (106)	.27** (128)	.22* (118)	.03 (99)	.19† (102)	.35*** (113)	-						
12. ICE - Trial 1	.12 (185)	-.08 (187)	.12 (177)	.08 (170)	-.06 (148)	.13† (179)	.13 (165)	.12 (130)	.07 (133)	.12 (147)	.27** (133)	-					
13. ICE - Trial 2	.06 (180)	-.10 (182)	.12 (172)	.04 (165)	-.02 (145)	.05 (175)	.17* (161)	.01 (127)	.22* (133)	.21* (145)	.32*** (131)	.47*** (182)	-				
14. 2-Var CVS - Trial 1	.04 (185)	-.25*** (187)	.001 (177)	.06 (170)	.08 (148)	.10 (179)	.01 (165)	-.04 (130)	-.02 (133)	-.02 (147)	.09 (133)	.03 (187)	-.07 (182)	-			
15. 2-Var CVS - Trial 2	-.02 (185)	-.11 (187)	.08 (173)	-.04 (166)	.08 (145)	.15* (175)	.05 (161)	.02 (126)	.10 (131)	.04 (145)	.10 (130)	.11 (183)	.02 (180)	-.02 (183)	-		
16. 3-Var CVS - Trial 1	-.07 (185)	-.05 (187)	.09 (177)	.02 (170)	.08 (148)	.17* (179)	-.02 (165)	.06 (130)	.08 (133)	.23** (147)	.21* (133)	.09 (187)	.03 (182)	.09 (187)	.20** (183)	-	
17. 3-Var CVS - Trial 2	-.01 (180)	.08 (182)	.06 (173)	.004 (166)	.07 (145)	.12 (174)	.09 (160)	.10 (127)	.08 (131)	.02 (145)	.06 (131)	.13† (182)	.13† (180)	.06 (182)	.21** (180)	.17* (182)	-

† $p < .10$ * $p < .05$. ** $p < .01$. *** $p < .001$

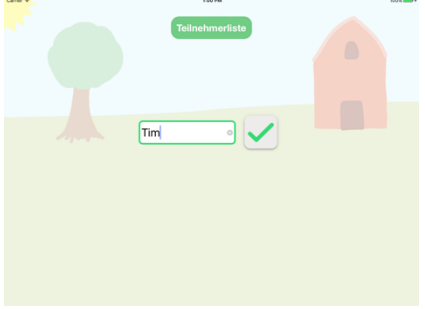


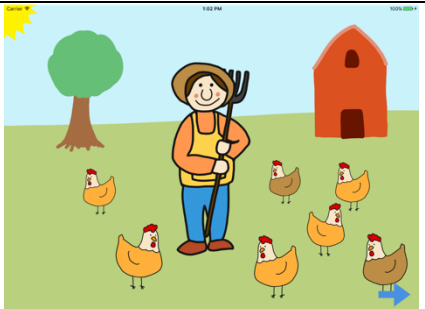
Appendix H: Cross tabulation for mastery of Theory of Mind and the ICE and CVS tasks

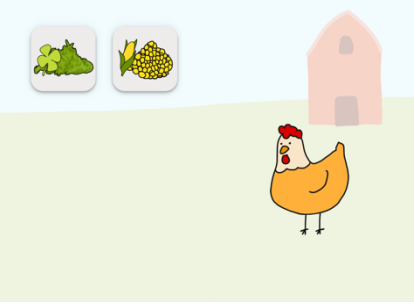
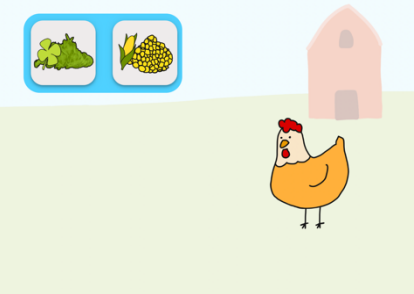
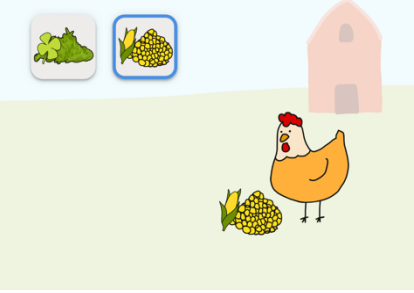
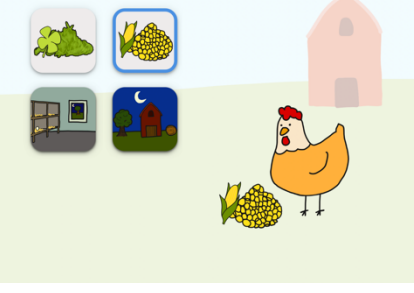
		ICE		
Theory of Mind	No mastery	Mastery	Total	
No mastery	34 (19.9%)	20 (11.7%)	54 (31.6%)	
Mastery	54 (31.6%)	63 (36.8%)	117 (68.4%)	
Total	88 (51.5%)	83 (48.5%)	171 (100%)	

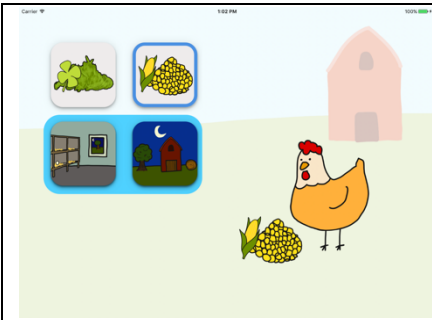
		2-Variable CVS		
Theory of Mind	No mastery	Mastery	Total	
No mastery	34 (20.0%)	19 (11.2%)	53 (31.2%)	
Mastery	65 (38.2%)	52 (30.6%)	117 (68.8%)	
Total	99 (58.2%)	71 (41.8%)	170 (100%)	

		3-Variable CVS		
Theory of Mind	No mastery	Mastery	Total	
No mastery	38 (22.2%)	16 (9.4%)	54 (31.6%)	
Mastery	70 (40.9%)	47 (27.5%)	117 (68.4%)	
Total	108 (63.2%)	63 (36.8%)	171 (100%)	

Appendix I: Protocol for Study 5a (original in German; *translated to English*)

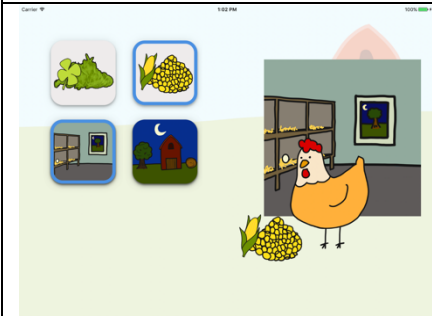
Introduction story and training with choosing variables	
	
	<p>Auf diesem Bild siehst du Bauer Meyer.</p> <p><i>This is Farmer Meyer.</i></p>
	<p>Bauer Meyer hat ganz viele verschiedene Tiere, zum Beispiel Kühe, Esel oder Schweine..</p> <p><i>Farmer Meyer has lots of different animals, for example, cows, donkeys, and pigs...</i></p>
	<p>Bauer Meyer hat außerdem ganz viele verschiedene Hühner!</p> <p><i>Farmer Meyer also has lots of different chickens!</i></p>

	<p>Bauer Meyer kann entscheiden, welches Futter seine Hühner bekommen sollen. Sie können entweder Kräuter oder Mais fressen.</p> <p><i>Farmer Meyer can decide, which food to feed his chickens. He can feed them herbs or corn.</i></p>
	<p>Schau mal, du kannst auf eine Art von Futter tippen, um die Hühner damit zu füttern! Probiere es mal aus.</p> <p><i>Look, you can click on a type of food to feed it to the chicken. Try it!</i></p>
	<p>Du hast Mais ausgewählt und es diesem Huhn gefüttert.</p> <p><i>You picked corn to feed to the chicken.</i></p>
	<p>Bauer Meyer kann außerdem auswählen, wo seine Hühner schlafen sollen, sie können draußen oder drinnen im Stall schlafen.</p> <p><i>Farmer Meyer can also decide where his chickens sleep. They can sleep outside or inside the stall.</i></p>



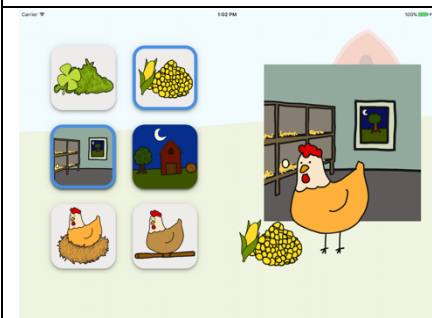
Du kannst auswählen wo die Hühner schlafen sollen, entweder drinnen oder draußen.

You can pick where the chickens should sleep inside or outside.



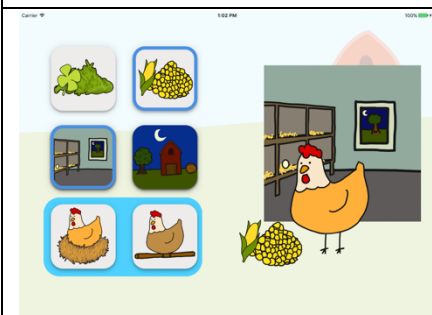
Du hast ausgewählt, dass dieses Huhn drinnen im Stall schlafen sollen.

You chose that this chicken sleeps inside.



Bauer Meyer kann zuletzt noch auswählen, welche Art von Schlafplatz seine Hühner haben sollen. Sie können in einem Nest oder auf einer Sitzstange schlafen.

Farmer Meyer can also decide which type of sleeping place his chickens have. They can sleep in a nest or on a perch.



Du kannst auswählen, ob die Hühner in einem Nest oder auf einer Sitzstange schlafen sollen.

You can choose if the chickens should sleep in a nest or on a perch.

	<p>Du hast ausgewählt, dass dieses Huhn in einem Nest schlafen soll.</p> <p><i>You chose that this chicken sleeps in a nest.</i></p>
--	--

Critical event and hypothesis statement

	<p>Bauer Meyer ist aufgefallen, dass einige seiner Hühner gepunktete Eier und einige weiße Eier legen. Er fragt sich, warum das passiert.</p> <p><i>Farmer Meyer has noticed that some of his chickens are laying spotted eggs. He wonders why this is happening.</i></p>
--	---

	<p>Bauer Meyer denkt, dass es an der Art von Futter, das die Hühner fressen, liegen könnte. Es könnte auch daran liegen, ob die Hühner drinnen oder draußen schlafen. Es könnte auch an der Art von Schlafplatz liegen, den die Hühner bekommen.</p> <p><i>Farmer Meyer thinks that it could be because of the type of food the chickens eat. It could also be because of where the chickens sleep or the type of sleeping place the chickens have.</i></p>
--	---



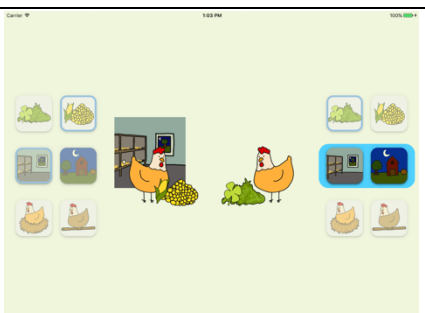
	<p>Bauer Meyer denkt, dass es an der Art von Futter, das die Hühner fressen, liegen könnte, dass die Hühner gepunktete oder weiße Eier legen.</p> <p><i>Farmer Meyer thinks that the type of food makes a difference in whether the chickens lay spotted or plain eggs.</i></p>
--	---

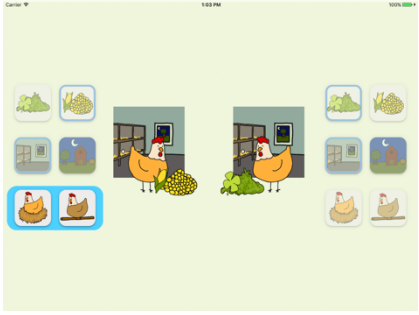
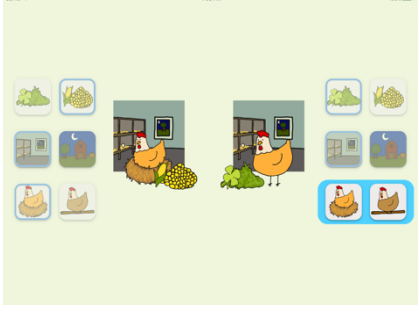
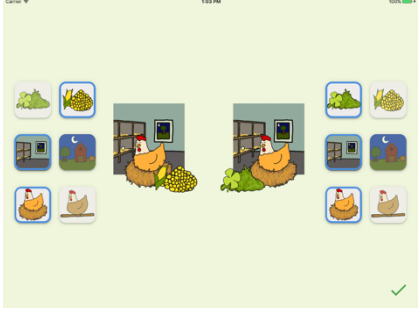
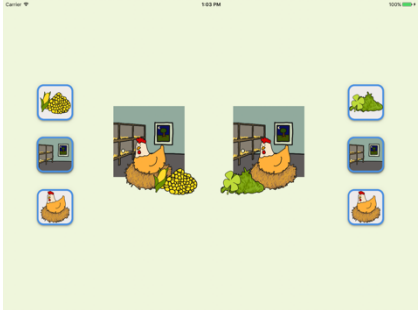
Was glaubst du wie er testen könnte, ob das Futter einen Unterschied macht?

How do you think he can test if the food makes a difference?

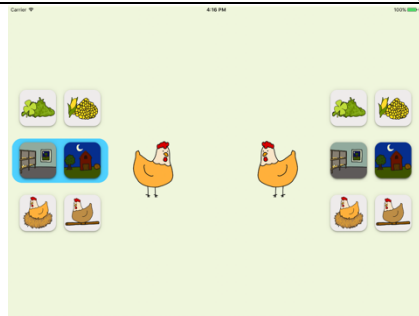
	<p>Wähle nacheinander auf jeder Seite eine Art von Futter, einen Ort zum Schlafen, und eine Art von Schlafplatz aus, um herauszufinden, ob das Futter einen Unterschied macht. Wähle als erstes für das linke Huhn eine Art von Futter, Kräuter oder Mais, aus.</p> <p><i>Choose from each side a type of food to feed the chickens, a sleeping location, and a type of sleeping place to find out if the type of food makes a difference. First, you can pick a type of food for the chicken on the left, herbs or corn.</i></p>
--	---

Task 1: Experimental Design

	<p>Wähle für dieses (rechte) Huhn eine Art von Futter aus.</p> <p><i>Pick a type of food for the chicken on the right.</i></p>
	<p>Wähle für dieses (linke) Huhn einen Ort zum Schlafen, drinnen oder draußen, aus.</p> <p><i>Choose where this chicken should sleep, inside or outside.</i></p>
	<p>Wähle für dieses (rechte) Huhn einen Ort zum Schlafen aus.</p> <p><i>Choose where this chicken, on the right, should sleep.</i></p>

	<p>Wähle für dieses (linke) Huhn eine Art von Schlafplatz, Nest oder Sitzstange, aus.</p> <p><i>Choose a type of sleeping place for this chicken, a nest or a perch.</i></p>
	<p>Wähle für dieses (rechte) Huhn eine Art von Schlafplatz aus.</p> <p><i>Choose a type of sleeping place for this chicken on the right.</i></p>
	<p>Bist du mit deiner Auswahl fertig, oder möchtest du noch etwas verändern?</p> <p><i>Are you finished picking everything or do you want to change something?</i></p>
	<p>Denkst du dass dies ein guter Test ist, um herauszufinden ob das Futter einen Unterschied macht welche Art von Eier die Hühner legen, weiße oder gepunktete? Warum? Warum nicht?</p> <p><i>Do you think this is a good test to find out if the food makes a difference in whether chickens lay spotted or plain eggs? Why do you think that?</i></p>

Task 2: Experimental Design with Feedback



Sehr gut! Jetzt werden wir die Aufgabe nochmal machen und wir können dabei genauer besprechen, warum du eine bestimmte Auswahl getroffen hast.

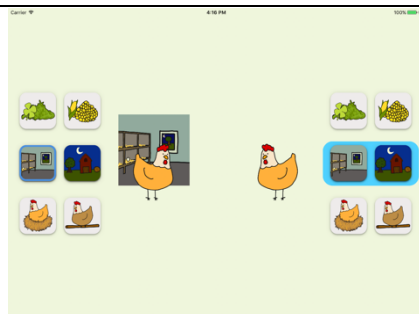
Good job! Now we will do the task again and this time we can talk a little about which things you choose.

Diesmal denkt Bauer Meyer nämlich, dass es an dem Ort zum Schlafen, drinnen oder draußen, liegen könnte, dass die Hühner gepunktete oder weiße Eier legen. Was glaubst du wie er testen könnte, ob der Ort zum Schlafen einen Unterschied macht?

This time, Farmer Meyer thinks that it's the sleeping location that makes a difference in whether chickens lay spotten or plain eggs. How do you think he can test if the sleeping location makes a difference?

Wähle nacheinander wieder auf jeder Seite einen Ort zum Schlafen, eine Art von Schlafplatz und eine Art von Futter aus, um herauszufinden, ob der Ort zum Schlafen einen Unterschied macht. Wähle als erstes für das linke Huhn einen Ort zum Schlafen, drinnen oder draußen, aus.

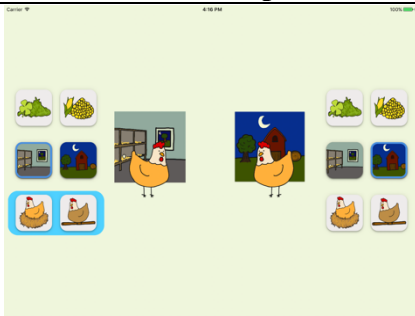
Choose from each side a type of food to feed the chickens, a sleeping location, and a type of sleeping place to find out if the sleeping location makes a difference. First, you can pick a sleeping location for the chicken on the left, inside or outside.



Wähle für dieses (rechte) Huhn einen Ort zum Schlafen aus.

Choose a sleeping location for the second chicken, on the right.

Example: Feedback for the selection of the focal variable

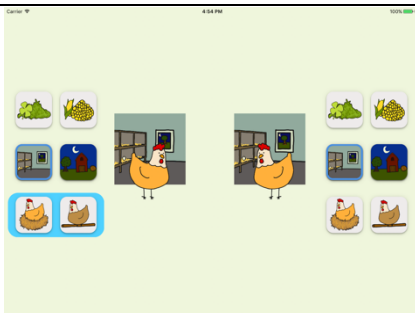


(korrekte Auswahl) Sehr gut! Wir wollen herausfinden, ob der Ort zum Schlafen einen Unterschied macht, deswegen ist es richtig, dass du zwei verschiedene Orte ausgewählt hast, damit wir sie vergleichen können.

(correctly varied focal variable) Very good! We want to find out if the sleeping location makes a difference. So, you're correct that you need to pick two different sleeping locations so that we can compare them.

Als nächstes kannst du jetzt für dieses (linke) Huhn eine Art von Schlafplatz, Nest oder Sitzstange, auswählen.

Next, you can choose a type of sleeping place for the chicken on the left, a nest or a perch.

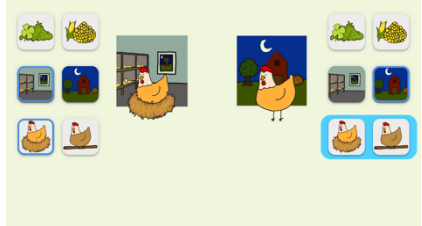


(falsche Auswahl) Du hast zwei gleiche Orte zum Schlafen ausgewählt, allerdings wollen wir herausfinden, ob der Ort zum Schlafen einen Unterschied macht. Deswegen müssten wir eigentlich zwei verschiedene Ort auswählen, damit wir sie vergleichen können. Lass uns einen der Orte verändern, damit sie unterschiedlich sind.

(did not vary focal variable) You picked the same sleeping location for both chickens. But actually, we want to find out if the sleeping location makes a difference, so we have to pick two different sleeping locations so that we can compare them. Let's change one of the sleeping locations so that they are different.

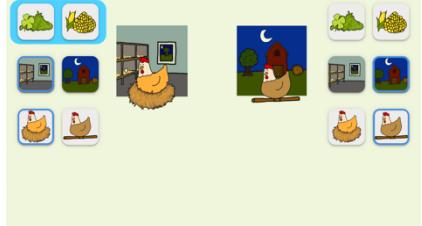
Als nächstes kannst du jetzt für dieses (linke) Huhn eine Art von Schlafplatz, Nest oder Sitzstange, auswählen.


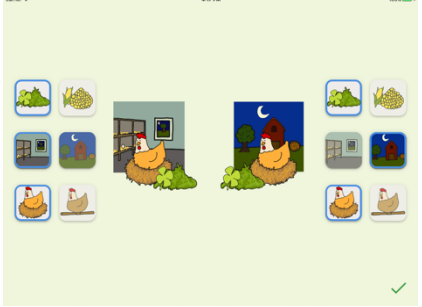
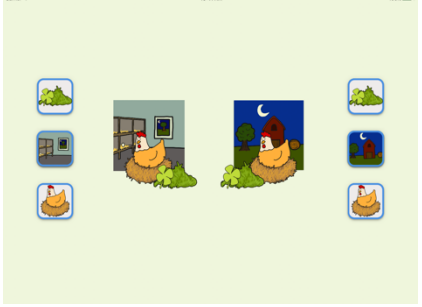
Next, you can choose a type of sleeping place for the chicken on the left, a nest or a perch.

	<p>Wähle für dieses (rechte) Huhn eine Art von Schlafplatz aus.</p> <p><i>Choose a type of sleeping place for the second chicken, on the right.</i></p>
---	---

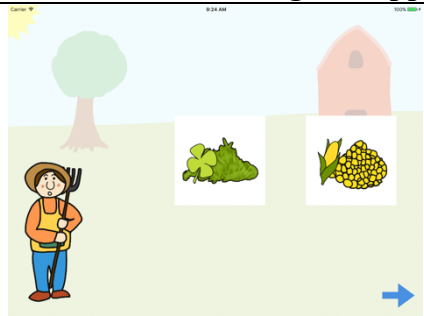
Example: Feedback for the selection of the control variables

	<p>(korrekte Auswahl) Sehr gut! Wir wollen herausfinden, ob der Ort zum Schlafen einen Unterschied macht, deswegen ist es richtig, dass du zwei gleiche Schlafplätze ausgewählt hast. So können wir sicher sein, dass wir wirklich nur vergleichen, ob der Ort zum Schlafen einen Unterschied macht</p> <p><i>(correctly controlled control variable) Very good! We want to find out if the sleeping location makes a difference. So, you're correct that you need to pick the same type of sleeping place for both chickens. Only this way can we compare if the type of sleeping place makes a difference.</i></p> <p>Als nächstes kannst du jetzt für dieses (linke) Huhn eine Art von Futter, Kräuter oder Mais, auswählen.</p> <p><i>Next, you can choose a type of food for the chicken on the left, corn or herbs.</i></p>
---	--

	<p>(falsche Auswahl) Du hast zwei verschiedene Schlafplätze ausgewählt, allerdings wollen wir herausfinden, ob der Ort zum Schlafen einen Unterschied macht dass die Hühner verschiedene Eier legen. Deswegen müssten wir eigentlich zwei gleiche Schlafplätze auswählen, damit der Schlafplatz nicht unseren Vergleich beeinflusst. Lass uns zwei gleiche Schlafplätze auswählen, damit wir nur den Ort zum Schlafen vergleichen.</p> <p><i>(did not control control variable) You picked two difference types of sleeping place for the chickens. But actually, we want to find out if the sleeping location makes a difference in whether the chickens lay spotted or plain eggs. So, we have to pick the same type of sleeping place for both chickens so that they sleeping place doesn't influence our comparison. Let's change one of the sleeping places so that they</i></p>
---	--

	<p><i>are the same for both chickens and so we only compare the sleeping location.</i></p> <p>Als nächstes kannst du jetzt für dieses (linke) Huhn eine Art von Futter, Kräuter oder Mais, auswählen.</p> <p><i>Next, you can choose a type of food for the chicken on the left, corn or herbs.</i></p>
	<p>Wähle noch für dieses (rechte) Huhn eine Art von Futter aus.</p> <p><i>Choose a type of food for the second chicken, on the right.</i></p>
	<p>Bist du mit deiner Auswahl fertig, oder möchtest du noch etwas verändern?</p> <p><i>Are you finished picking everything or do you want to change something?</i></p>
	<p>Denkst du dass dies ein guter Test ist, um herauszufinden ob der Ort zum Schlafen einen Unterschied macht welche Art von Eier die Hühner legen, weiße oder gepunktete? Warum? Warum nicht?</p> <p><i>Do you think this is a good test to find out if the sleeping locations makes a difference in whether chickens lay spotted or plain eggs? Why do you think that?</i></p>
<p>Task 3: Same procedure as Task 1</p>	

Changes to application and script in second iteration



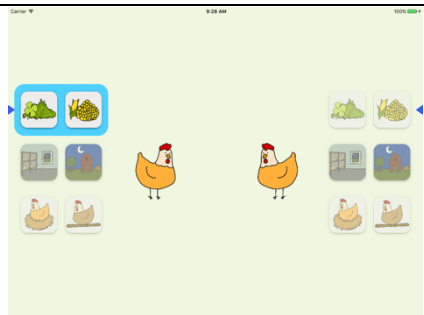
[Addition of screen highlighting the current focal variable]

Bauer Meyer möchte jetzt vergleichen, ob es an der Art von Futter, das die Hühner fressen, liegen könnte, dass manche Hühner gepunktete und manche weiße Eier legen.

Farmer Meyer wants to compare if the type of food that the chickens eat makes a difference in whether the chickens lay spotted or plain eggs.

Was glaubst du wie er testen könnte, ob das Futter einen Unterschied macht?

How do you think he could test if the type of food makes a difference?



[Addition of triangle/arrow to indicate current focal variable]

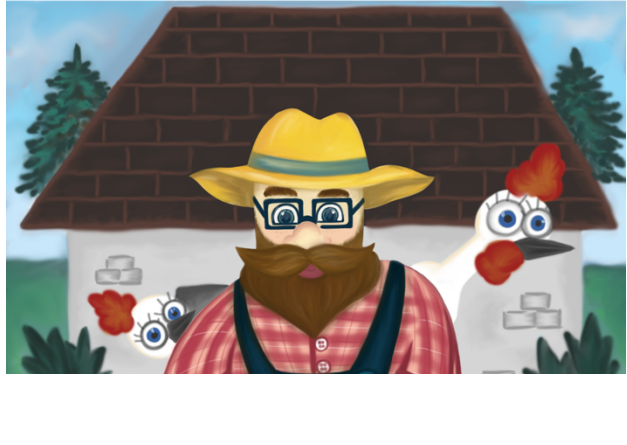

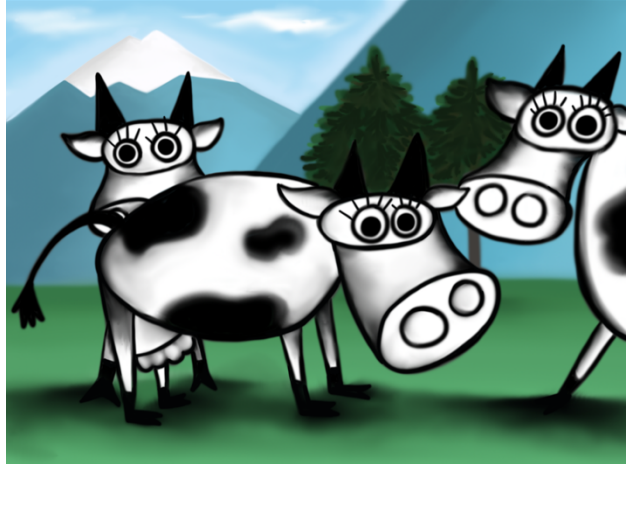
Wähle nacheinander auf jeder Seite eine Art von Futter, einen Ort zum Schlafen, und eine Art von Schlafplatz aus, um herauszufinden, ob das Futter einen Unterschied macht. Denke dran, du möchtest die Art von Futter vergleichen. Damit du dich immer daran erinnern kannst dass wir das Futter vergleichen wollen, siehst du hier zwei kleine Dreiecke.

Choose from each side a type of food to feed the chickens, a sleeping location, and a type of sleeping place to find out if the sleeping location makes a difference. Remember, you want to compare the type of food. So that you can remember that we want to compare the food, you can see this little triangle here.

Wähle als erstes für das erste linke Huhn eine Art von Futter, Kräuter oder Mais, aus.

First, you can pick a type of food for the chicken on the left, herbs or corn.

Appendix J: Protocol for Study 5b

	<p>[Introduction Screen]</p>
	<p>Once upon a time there was a farm on the mountain near the lake. This farm has been known for many years.</p>
	<p>The cows gave the best milk to make the King's favorite cheese.</p>



The sheep gave the best wool to knit the warmest sweaters.

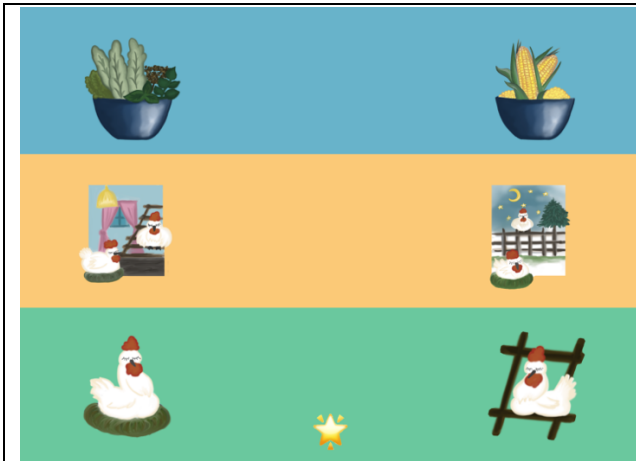


But most of all, the farm was famous because of its chickens. Each year, they laid more eggs than any other farm.

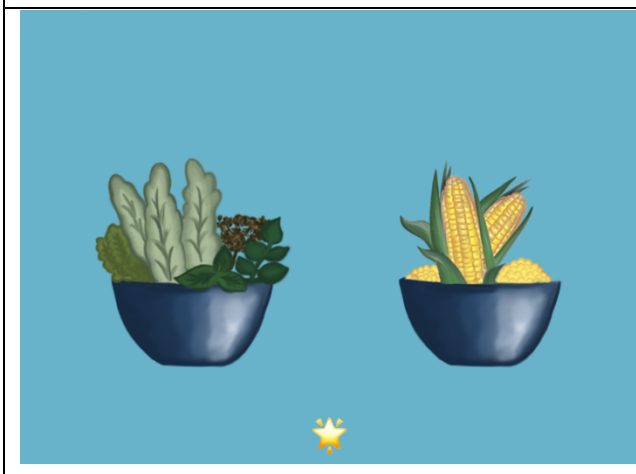


Hello, I am Farmer Meyer. For many years I have a farm and live here with my animals.

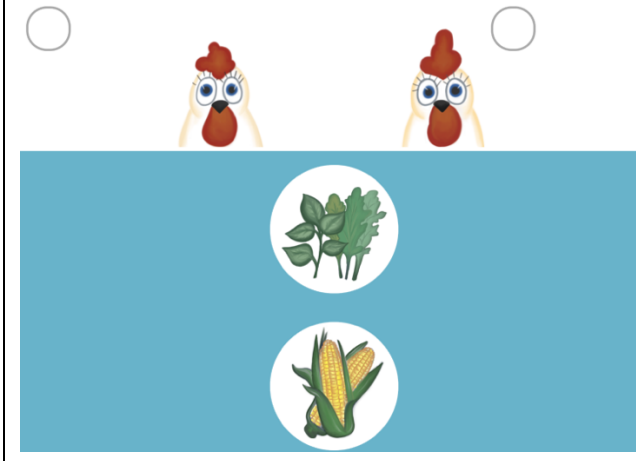
Here is my cow Lucy, my pig Peggy, and my sheep Bob. I have a lot of chickens and I know all their names.



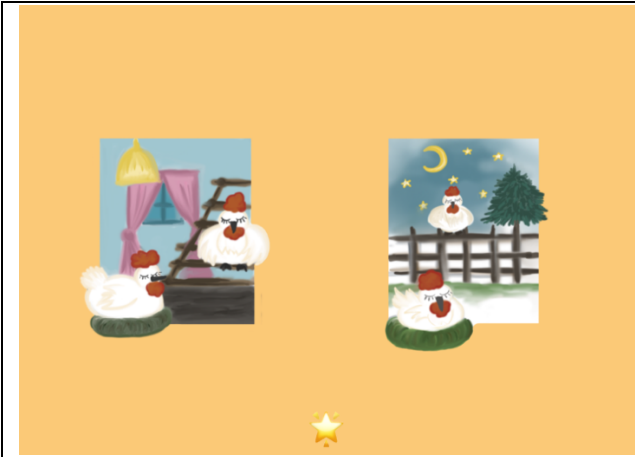
I can feed my chickens herbs or corn.
 I can decide if they sleep inside or outside.
 And I can decide if they sleep in a nest or on a perch.



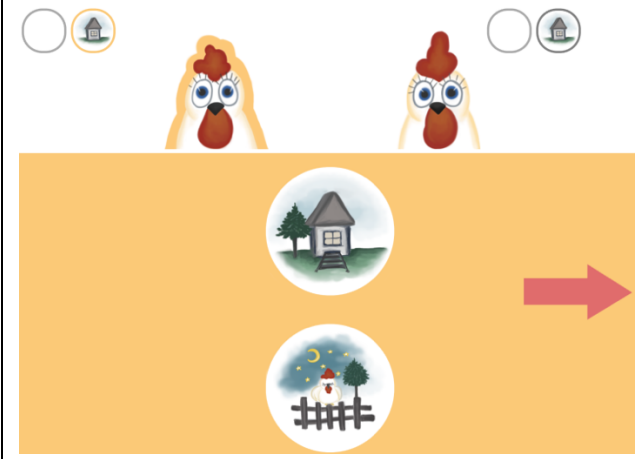
Do you want to help me feed the chickens?



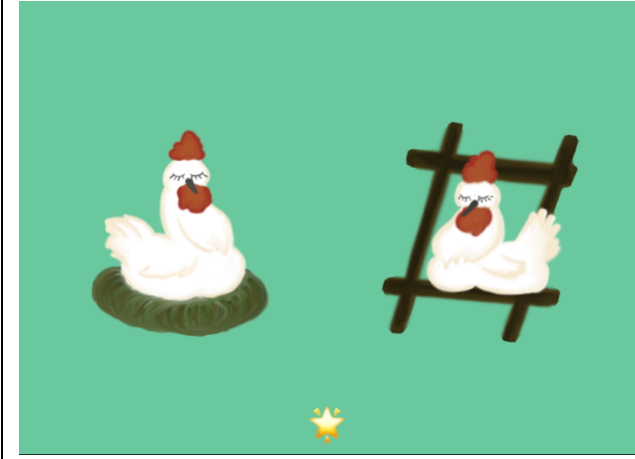
Choose a type of food for each chicken.



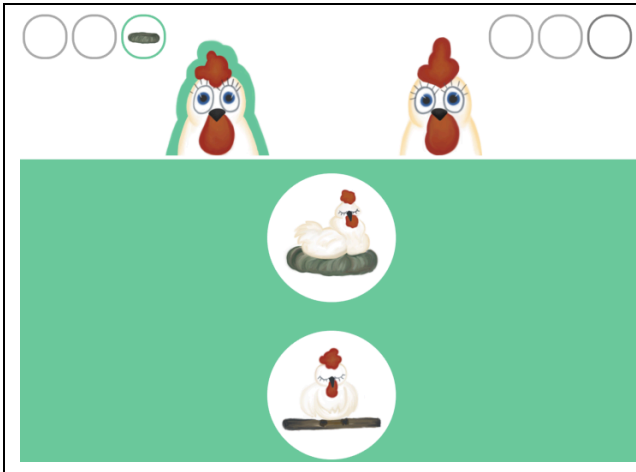
You can also help me decide where they sleep, inside or outside.



Choose where each chicken should sleep.



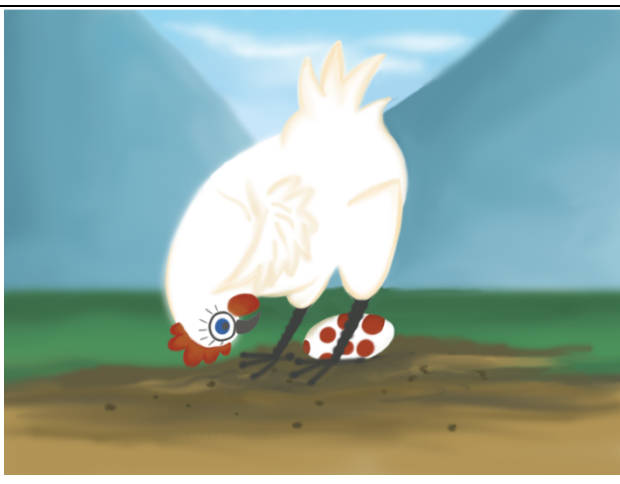
You can also help me decide if they sleep in a nest or on a perch.



Choose how each chicken should sleep.

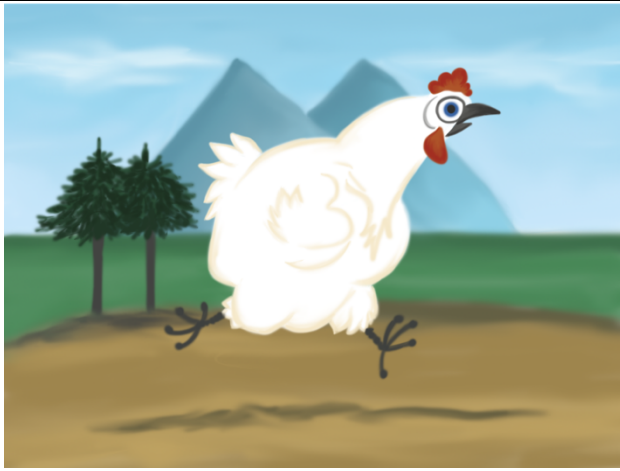


One day something extraordinary happened and nobody could explain it.



Rosy laid a spotted egg!

"My eggs! "- exclaims Rosy. "What happened? I should tell Farmer Meyer!"



Rosy runs as fast as she can to tell Farmer Meyer what happened.



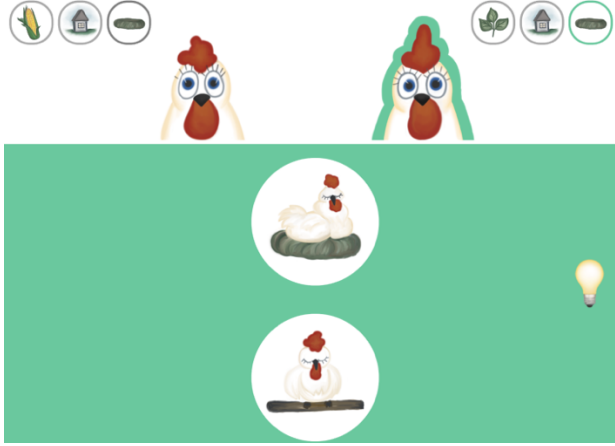
Farmer Meyer! Farmer Meyer! Quick! You have to see it! Something strange just happened. I need to show you. Come with me.




That was the first time that Farmer Meyer discovered that Rosy had laid spotted eggs. He had never seen something like this before. Until now all the chickens' eggs were white.

"Hm, interesting! What could be the reason?"

	<p>Is it the food that Rosy ate? Is it because she was sleeping inside or outside? Or is it because of the way of sleeping? What could it be?</p> <p>I think it was the type of food that makes a difference. Can you help me find out?</p> <p>To do a good experiment, the type of food should be different for each chicken, but everything else should be the same. Do you understand?</p>
	<p>Choose a type of food for each chicken</p>
	<p>Choose a sleeping location for each chicken.</p>

	<p>Choose a type of sleeping place for each chicken.</p>
	<p>[If the design was correct, children were told that it was correct and why; if the design was not correct, children were told that it was not correct and why and given the instructions again. The children could then try again to design a controlled experiment.]</p>

**Appendix K: Tutorial storyboard and script for Study 6 (original in German;
English translation in *italics*)**

Erzähler <i>Narrator</i>	Anmerkungen für die Animation <i>Notes for the animations</i>
Einleitung <i>Introduction</i>	
<p>Hey du! :-) Darf ich kurz vorstellen: Das sind Max und Anna.</p> <p><i>Hey you! Let me introduce Max and Anna.</i></p>	<p>Max und Anna einblenden (Animation 1) <i>Max and Anna appear</i></p> 
<p>Max und Anna sind beide Forscher. Sie interessieren sich sehr für Flugzeuge!</p> <p><i>Max and Anna are both researchers. They are very interested in planes.</i></p> <p>Vor allem für die ganz ganz schnellen! Je schneller desto besser.</p> <p><i>Above all, the really fast planes - the faster the better.</i></p>	<p>Flugzeug fliegt ins Bild (Animation 2) <i>Plane flies into the frame</i></p> <p>Flugzeug fliegt aus dem Bild (Animation 3) <i>Plane flies out of the frame</i></p>
<p>Heute wollen Max und Anna eine ganz wichtige Fragen lösen:</p> <p><i>Today, Max and Anna want to solve an important question:</i></p> <p>Fliegt eigentlich ein <u>eckiges</u> Flugzeug schneller oder ein <u>rundes</u>?</p> <p><i>Do pointy or rounded planes fly faster?</i></p>	<p>Übergang Flieger vs Flieger (Animation 4) <i>Transition between planes</i></p> <p>Einblendung rundes Flugzeug und eckiges Flugzeug (Animation 5) <i>Appearance of rounded and pointy planes</i></p>

Hat also **die Form** des Flugzeuges einen Einfluss darauf, wie schnell das Flugzeug fliegen kann?

*Does **the shape** of the plane make a difference in how fast the plane can fly?*

Max hat nämlich mal gehört, dass eckige Flugzeuge viel schneller fliegen. Anna glaubt aber, dass die Flugzeuge rund sein müssen, damit sie schnell fliegen können.

Max heard that pointy planes fly faster. But Anna thinks that the plane should be rounded so that it can fly fast.



Einblendung Fragezeichen (Animation 6)
Appearance of question mark

Was machen Forscher, wenn sie so eine Frage lösen müssen?

What do researchers do when they want to answer such a question?

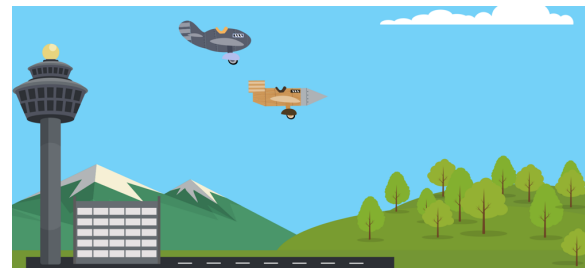
Genau! Ein Experiment!

Exactly! An experiment!

Und wo gibt es Flugzeuge? Richtig! Auf dem Flughafen.

And where are airplanes? That's right! At the airport!

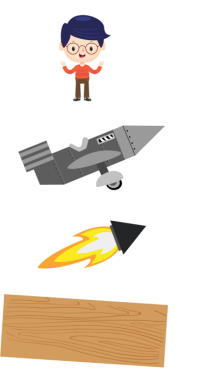
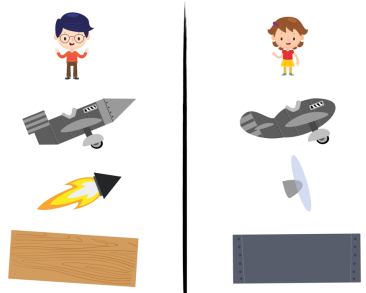
Einblendung Flughafen (Animation 7)
Appearance of airport



Max und Anna beschließen, dass sie 2 Flugzeuge bauen. Max baut ein eckiges und Anna baut ein rundes. Dann können sie ja einfach beide fliegen und sehen welches schneller ist.

Max and Anna decide to build two airplanes. Max builds a pointy airplane and

Flugzeuge bewegen sich in Hangar
Planes move into hangar (Animation 8)

<p><i>Anna builds a rounded plane. Then they can fly the planes and see which one is faster.</i></p>	
<p>Erster Bau First Build</p>	
<p>Max baut also das eckige Flugzeug. <i>So, Max builds the pointy plane.</i></p> <p>Er baut sein Flugzeug außerdem mit einem Düsenantrieb. <i>He uses a jet engine.</i></p> <p>Und als Material nutzt Max Holz. <i>And for the material, Max uses wood.</i></p>	<p>Übergang Tabelle (Animation 9) <i>Transition to the variable table</i></p> <p>Tabelle</p>  <p>Eckig/ <i>pointy</i> - Animation 10 Düsenantrieb/ <i>jet engine</i> - Animation 11 Holz/ <i>wood</i> - Animation 12 Zusammenbauen/ <i>build together</i> - Animation 13</p>
<p>Anna baut das runde Flugzeug. <i>Anna builds the rounded plane.</i></p> <p>Sie verbaut für ihr Flugzeug einen Propeller. <i>She uses a propeller engine.</i></p> <p>Ihr Flugzeug baut sie aus Metall. <i>She builds her plane out of metal.</i></p>	<p>Tabelle erweitern</p>  <p>Linie & Anna/ <i>Line & Anna appear</i> - Animation 14 Flugzeug Rund/ <i>round</i> - Animation 15 Propeller/ <i>propeller</i> - Animation 16 Metall/ <i>metal</i> - Animation 17 Zusammenbauen/ <i>build together</i> - Animation 18</p>
<p>Jetzt bauen beide ihre Flugzeuge zusammen!</p>	

Now they both put their planes together!
Max hat also ein eckiges Flugzeug, mit einem Düsenantrieb aus Holz.

Max has a pointy plane made out of wood and with a jet engine.

Und Anna hat ein rundes Flugzeug, mit einem Propellor aus Metall.

And Anna has a rounded plane made out of metal and with a propellor engine.

Sehr cool! Damit sind beide Flugzeuge fertig! Zeit für ein Rennen! Und los gehts!!!

Very cool! Now both planes are finished. Time for a race - let's go!

WOW! Max Flugzeug fliegt viel schneller.

Wow! Max's plane flies much faster!

Max freut sich riesig und schreit: "Ich hab gewonnen, ein eckiges Flugzeug ist viel schneller!"

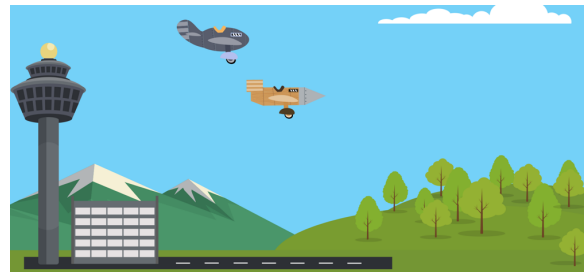
Max is super excited and yells "I won! A pointy plane is much faster!"

Aber Moment! Anna schreit Stopp!

But wait, Anna yells "Stop!"

Mmh... Weißt Du warum Anna Stopp schreit? Ist irgendwas falsch mit ihrem Experiment? Klar!

Hmm.. Do you know why Anna yelled stop? Is something wrong with their experiment? Exactly!

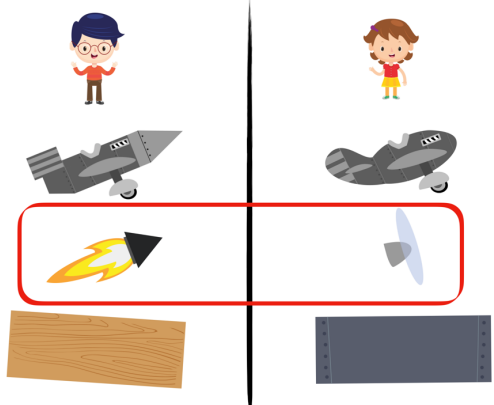


Flughafen einblenden - Animation 19
Appearance of airport

Rennen - Animation 20
Racing planes

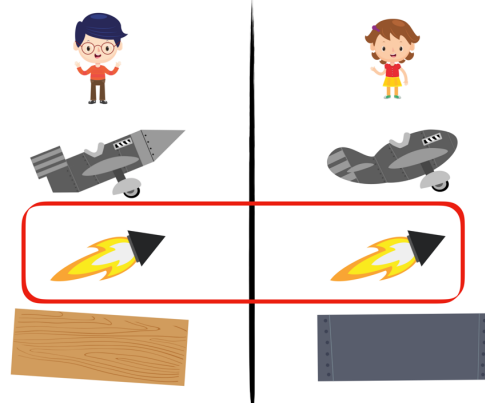
Einblendung Pokals - Animation 21
Appearance of trophy

Anna Stop - Animation 22
Appearance of Anna and stop sign

<p>Das ist doch total unfair! Max Flugzeug hat einen Düsentrieb. Annas Flugzeug nur einen Propellor Vielleicht ist nur deswegen Max Flugzeug viel schneller!</p> <p><i>That is totally unfair! Max's plane has a jet engine and Anna's plane has a propellor engine. Maybe that's why Max's plane was faster.</i></p>	
<p>Zweiter Bau Second Build</p>	
<p>Genau! Das ist ganz wichtig. Gucken wir uns nochmal an, was die beiden da eigentlich gebaut haben.</p> <p><i>Exactly! That is very important. Let's look again at what they both built.</i></p> <p>Max und Anna wollen ja testen, ob ein eckiges oder rundes Flugzeug schneller ist.</p> <p><i>Max and Anna want to test if a rounded or a pointy plane is faster.</i></p> <p>Aber Max hat einen Düsenantrieb und Anna einen Propellor. Vielleicht ist nur deswegen Max' Flugzeug schneller geflogen.</p> <p><i>But Max has a jet engine and Anna has a propellor engine. Maybe that's why Max's plane was faster.</i></p> <p>So können sie also gar nicht vergleichen, ob jetzt rund oder eckig schneller ist... Mmmhh was jetzt...?</p> <p><i>They can't even compare if a rounded plane or a pointy plane is faster... Hmmm what now?</i></p>	 <p>Überblendung Tabelle - Animation 23 <i>Transition to variable table</i></p> <p>Linie zeichnen - Animation 24 <i>Draw line</i></p>

Klar! Anna baut einfach auch einen Düsenantrieb an ihr Flugzeug. So können sie schon besser vergleichen!

Exactly! Anna will also use a jet engine for her plane. Then they can compare better.



Anna Düsenantrieb - Animation 25
Highlighting the difference/ changing to jet engine

Jetzt hat Anna also auch ein Düsenantrieb an ihrem runden Flugzeug.

Now Anna also has a jet engine on her rounded plane.

Übergang & Zusammenbau - Animation 26
building the planes & transition

Super! Jetzt haben beide Flugzeuge den gleichen Antrieb, also können die beiden testen. Los gehts!

Great! Now both planes have the same engine, so they can race. Let's go!

Wow! Diesmal ist Annas Flugzeug schneller!

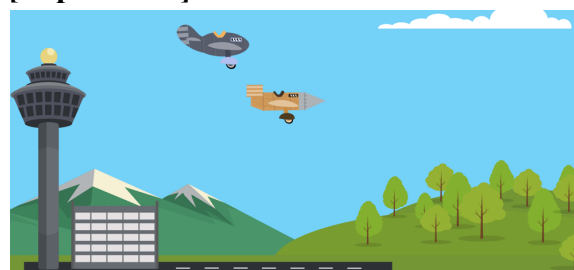
Wow! This time, Anna's plane is much faster!

Anna freut sich riesig! Yeah! Also heißt das, dass ein rundes Flugzeug viel schneller fliegt.

Anna is very excited! Yay! That means a rounded plane is much faster!

Diesmal schreit aber Max: Stopp, halt!

[Experiment]



Übergang Flughafen - Animation 27
transition to airport

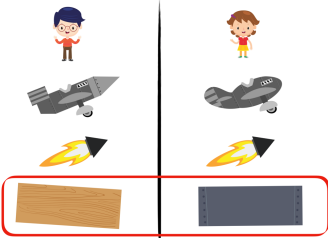
Rennen - Animation 28
racing planes

Einblendung Pokal - Animation 29
Appearance of trophy

Max Stopp - Animation 30
Appearance of Max & stop sign

<p><i>But this time, Max yells “Stop, wait!”</i></p> <p>Wie, was ist denn jetzt?</p> <p><i>Hm, what is it this time?</i></p> <p>Weißt Du warum Max diesmal Stopp schreit? Weißt du was falsch ist?</p> <p><i>Do you know why Max yelled Stop this time? Do you know what is wrong?</i></p> <p>Stimmt! Guck dir mal die beiden Flugzeuge nochmal an! Annas Flugzeug ist ja aus Metall. Max seins ist aus Holz. Vielleicht fliegt ein Flugzeug aus Metall ja viel schneller.</p> <p><i>Exactly! Let’s look at the planes again. Anna’s plane is made of metal and Max’s is made of wood. Maybe planes made of metal fly faster.</i></p>	
---	--

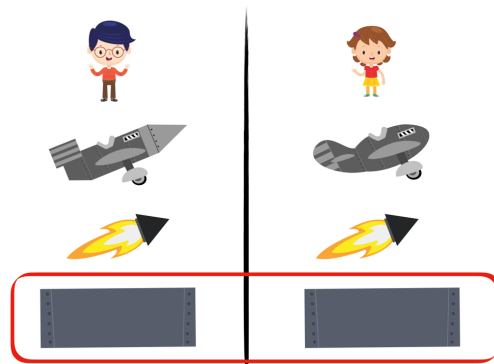
Dritter Bau & Lektion
Third Build & CVS Instruction

<p>Stimmt! Anna und Max wollen ja immer noch wissen, ob ein rundes oder eckiges Flugzeug schneller ist.</p> <p><i>Exactly! Anna and Max want to know if a rounded or a pointy plane flies faster.</i></p> <p>Deswegen haben sie schon darauf geachtet, den gleichen Antrieb zu benutzen.</p> <p><i>That’s why they made sure they both have the same type of engine.</i></p> <p>Aber das Material ist noch unterschiedlich! Max Flugzeug ist aus Holz, Annas aus Metall. Vielleicht macht das ja auch einen Unterschied!</p>	<p>[Tabelle]</p>  <p>Übergang Tabelle - Animation 31 <i>Transition to table</i></p> <p>Linie Zeichnen - Animation 32 <i>draw line</i></p>
--	--

But the material is still different! Max's plane is made of wood and Anna's is made of metal. Maybe that also makes a difference!

Also baut Max sein Flugzeug schnell um.
Jetzt ist es auch aus Metall.

So, Max quickly rebuilds his plane. Now it is also made of metal.



Umbau Metall - Animation 33

Highlighting difference and change to metal

Das ist ganz wichtig! Das musst du dir unbedingt merken wenn du mal Forscher wirst. Erst jetzt haben Max und Anna ein faires Experiment.

That is very important! You have to definitely do this if you want to be a researcher. Only now do Max and Anna have a fair experiment.

Max und Anna wollen ja vergleichen, ob ein rundes oder eckiges Flugzeug schneller fliegt. Dafür ist es wichtig, dass alle anderen Sachen gleich sind.

Max and Anna want to compare if a rounded or a pointy plane flies faster. To do this, it's important that all the other things are the same.

Sie müssen den gleichen Antrieb benutzen. Und auch das gleiche Material. Nur so wissen sie sicher, welche Form jetzt die bessere ist.

They need to use the same type of engine

Gutes Experiment - Animation 34

Appearance of green check mark

<p><i>and the same material. Only this way can they be sure which shape is better.</i></p>	
<p>Ende Ending</p>	
<p>Jetzt können Max und Anna endlich ihr faires Experiment durchführen. Los gehts!!!</p> <p><i>Now, Max and Anna can finally do a fair experiment. Let's go!</i></p> <p>Wow! Anna fliegt viel schneller! Sie gewinnt das Rennen! Und was bedeutet das?</p> <p><i>Wow! Anna flies much faster and she wins the race! And what does that mean?</i></p> <p>Ganz einfach: Anna und Max können jetzt SICHER sagen, dass ein rundes Flugzeug schneller fliegt. Sie haben nur die Form verändert und deshalb ein faires Experiment gemacht.</p> <p><i>Very simple: Anna and Max can now know for sure that a rounded plane flies faster than a pointy plane. They only changed the shape of the plane and did a fair experiment.</i></p>	
<p>Damit gibt sich auch Max geschlagen. Die beiden freuen sich über ihre Flugzeuge und fliegen den restlichen Abend noch ein bisschen herum.</p> <p><i>And Anna won. They are both excited about their planes and spend the rest of the evening flying around.</i></p>	<p>Endanimation Flugzeuge fliegen in den Horizont <i>planes fly off into the horizon</i></p> 

