The Development of Scientific Reasoning in Preschoolers: Hypothesis Testing, Evidence Evaluation and Argumentation from Evidence

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

am Munich Center of the Learning Sciences

der Ludwig-Maximilians-Universität

München

Vorgelegt von

Özgün Köksal Tuncer

München 2017

1st Supervisor/Erstgutachterin: Prof. Dr. Beate Sodian

2nd Supervisor/Zweitgutachter: Prof. Dr. Reinhard Pekrun

Ph.D. thesis submission: 30.11.2017

Ph.D. defense date: 31.01.2018

The research presented in this work was supported by the Elite Network of Bavaria [Project number: K-GS-2012-209]. I would like to extend my sincere gratitude to the opportunities (conference attendance, incubator stay, international knowledge exchange) made possible by the ENB.

To my father

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Dr. Beate Sodian for the continuous support of my Ph.D. study, for guidance, patience and immense knowledge. Your mentoring helped me in all time of research and writing of this thesis. I would like to also thank you for providing the opportunity to continue this fascinating line of research in the Project Explore.

I would like to thank my supervisors: Thank you Prof. Reinhard Pekrun and Prof. Markus Paulus for your insightful comments and challenging questions. Thank you Prof. Cristine Legare for agreeing to collaborate and for the great experience I had when I was visiting the Cognition, Culture and Development Lab in Austin, Texas.

I would like to thank April Moeller, Bahar Tunçgenç, Maryam Alqassab, and Noemi Skala for their feedback during the writing process of this thesis.

It was a great experience for me to be a member of the Reason Team. I would like to thank all of my Reason colleagues who made moving to a new country and doing Ph.D. here easier for me. Special thanks to Maria Fysaraki and Maryam Alqassab for their close friendship; I feel very lucky to have you in my life. Thank you, Sandra Becker, Janina Klemm, Ansgar Opitz, and Andrea Saffran for your friendship, support and for all the fun. I would also like to thank my developmental psychology friends, especially to my office-mates, Saskia Tobias, Kerstin Ganglmayer, and Noemi Skala for their friendship and support during writing this thesis.

I thank my mother and my brother for their continuous support and enduring belief in me.

Last but not least, I thank my husband Ünal, for coming with me to another country, for building a life from scratch to be on my side when I have been pursuing my goals. You have been loving and supportive beyond imagination.

Abstract

Although research on scientific reasoning has shown that young children have poor skills in epistemic activities such as evidence evaluation or experimentation; recent research demonstrated that they have powerful learning mechanisms in making causal predictions from evidence patterns or performing experiments to reveal causal relations that are not readily available to them. Although these abilities are informative concerning early epistemic activities, little is known about whether young children can reason scientifically. The ability to coordinate hypotheses and evidence; and having a metacognitive understanding of the hypothesis–evidence relation are the two foundational abilities for scientific reasoning. In three empirical studies, the present thesis investigated the development of these two abilities in 4- to 6-year-old preschoolers in three epistemic activities; namely, hypothesis testing, evidence evaluation, and argumentation from evidence. Study 1 showed that older preschoolers can differentiate between epistemic goals of hypothesis testing and practical goals of effect production, which suggest that the epistemic categories of hypotheses and evidence; and the ability to coordinate the two is already present in the late preschool years. Study 2 revealed that preschoolers can generate disconfirming evidence in order to refute false causal claims and they can reflect on the relation between beliefs and evidence. Study 3 showed that 5- and 6-year-olds can reflect on the relation between their knowledge states and confounded evidence. The findings of the three studies suggest that the foundational abilities for scientific reasoning, understanding the inferential relation of hypothesis and evidence and the reflective ability over this relation are present in preschoolers.

Extended Summary

There is a discrepancy in the literature concerning preschoolers' scientific reasoning skills. Research has demonstrated that scientific reasoning abilities follow a monotonic development across childhood years (Amsel & Brock, 1996; Bullock, Sodian, & Koerber, 2009; Croker & Buchanan, 2011; Kuhn, Amsel, & O'Loughlin, 1988; Piekny & Maehler, 2013; Tschirgi, 1980). Younger children often show poor performance in epistemic activities such as evaluating evidence or designing unconfounded experiments. These findings suggest that young children may lack the fundamental conceptual understanding of scientific reasoning (Kuhn, 1989). On the contrary, a recent line of research on causal reasoning has demonstrated that preschoolers have powerful learning mechanisms that enable them to learn from evidence patterns and make interventions on the world in order to gain information that is not readily available to them (e.g., Cook, Goodman, & Schulz, 2011; Gopnik, Sobel, Schulz, & Glymour, 2001; Legare, 2012; Schulz & Bonawitz, 2007). These early abilities resemble the epistemic practices in science. This similarity has captured researchers' interest and has brought forward the question whether preschoolers can reason scientifically (Gopnik, 2012; Kuhn, 2011; Schulz, 2012).

The broad nature of scientific reasoning makes it challenging to determine what is foundational for its development. In this respect, two aspects have been proposed as essential for the development of scientific reasoning (Kuhn, 1988; 1989; Sodian, Zaitchik, & Carey, 1991). First is the ability to represent epistemic categories of hypotheses and evidence distinctly and to coordinate them properly. Second is metacognitive understanding—the ability to reflect on and control one's cognitive processes related to knowing and knowledge seeking (Kuhn, 1988). Although findings

on preschoolers' powerful learning mechanisms provide critical information about preschoolers' use of epistemic practices, little is known about the development of hypothesis–evidence understanding and metacognitive abilities for knowledge related processes at this age.

The goal of the present thesis was to investigate the development of scientific reasoning skills in the preschool years (4, 5 and 6 years of age), specifically focusing on the development of hypothesis–evidence relation and reflective abilities over this relation. In this respect, three empirical studies were conducted. Considering that the development of understanding hypothesis–evidence relation may follow different developmental patterns in different epistemic activities, three epistemic activities were investigated; namely, hypothesis testing, evidence evaluation, and argumentation from evidence. The blicket detector paradigm (Gopnik & Sobel, 2000)—a paradigm commonly used in causal reasoning studies to investigate cause–effect relations—was adapted according to the research questions of each empirical study.

Study 1 investigated preschoolers' abilities to differentiate epistemic goals of hypothesis testing from practical goals of effect production with three experimental studies. In two between-subjects conditions, children were presented with identical baseline evidence. Depending on the condition, they were either asked to test a given hypothesis (Hypothesis Testing Condition) or produce an effect (Effect Production Condition). The correct responses in the two conditions were different; Hypothesis Testing Condition required choosing an informative object in order to test a given hypothesis; whereas Effect Production condition required choosing a familiar object. Study 1a investigated this ability in 4- to 6-year-olds. The findings (N = 111) showed that preschoolers selectively chose the correct objects in response to the goals of each condition and that there was a developmental improvement from 4 to 6 years of age.

Two following studies focused on the "younger preschoolers" (4- and younger 5-year-olds) to investigate the research questions whether their poor performance in differentiating hypothesis testing from effect production is due to particular task demands (Study 1b, N = 54) and how they would perform in the case of exploratory epistemic goals (Study 1c, N = 54). The results suggest that younger preschoolers' poor performance was not due to task demands (Study 1b) and that they can differentiate exploratory epistemic goals from effect production (Study 1c). These findings suggest that there is a developmental change in differentiating hypothesis testing from effect production in early childhood. While "older preschoolers" (older 5- and 6-year-olds) can differentiate hypothesis testing from effect production, younger ones differentiate between them successfully only when the epistemic goals were exploratory. Taken together, there seems to be a developmental change in preschool years in their understanding of what it means to test a hypothesis.

Study 1 provided us with information regarding preschoolers' differentiation of hypothesis testing from effect production; however, it does not show whether children of this age have an explicit understanding of the inferential relation between mental concepts and evidence. Study 2 investigated children's understanding of hypothesis–evidence relation and metacognitive understanding with two epistemic activities: hypothesis testing and counterargumentation. In the Hypothesis Testing Phase, children were led to generate a specific hypothesis about the cause of a light effect and then test it. Subsequently, in the Argumentation Phase, they were presented with two false causal claims in order to elicit spontaneous evidence generation and evidence-based verbal counterarguments. The majority of the children (82%) adopted a systematic hypothesis testing strategy (positive or contrastive testing). Furthermore, 83% of the children provided valid evidence to disconfirm a false causal claim and/or valid verbal

counterarguments at least once. Older preschoolers performed better than the younger ones. Unlike the developmental pattern in hypothesis testing, there were no age differences in terms of evidence-based counterargumentation. In sum, Study 2 yielded evidence for a twofold developmental pattern: there was a developmental difference for hypothesis testing from 4 to 6 years; but even 4-year-olds have an explicit understanding of the inferential relation between beliefs and evidence shown by their generation of disconfirming evidence and explicit evidence-based counterarguments in order to refute false causal claims.

Although children's verbal counterarguments in Study 2 provided some information about preschoolers' abilities to reflect on the inferential relation between beliefs and evidence, to date, young children's ability to reflect on their knowledge states as a result of evidence has not been directly addressed. In an evidence evaluation task, Study 3 study aimed to investigate 5- and 6-year-olds' (N = 60) metacognition of their own knowledge and how their knowledge is constructed as a function of evidence when they are presented with conclusive versus confounded evidence. In a within-subjects design, children observed confounded (Confounded Condition) and unconfounded (Unconfounded Condition) evidence regarding the efficacy of a target object. In each condition, children were asked whether they knew the efficacy of the target object or whether they required more information to know that. After each response, children were asked to justify their answer. Children indicated that they required more information to know the efficacy of the target object in the Confounded Condition significantly more often than in the Unconfounded Condition. Forty-percent of the children provided evidence-based justifications at least once, where they referred to the confounded nature of evidence in the Confounded Condition as a reason for their lack of knowledge. Children's explicit reflections on the causal ambiguity in the case of

confounded evidence suggest that they possess some metacognitive understanding of the relation between hypotheses and evidence.

Altogether the findings of this thesis provided critical information regarding our knowledge on scientific reasoning abilities in preschool age. Preschoolers seem to have distinct representations for the epistemic categories for mental concepts (hypotheses, beliefs) and evidence and they understand the inferential relation between hypotheses, beliefs and, evidence. While even younger preschoolers have an understanding of the inferential relation between beliefs and evidence; the understanding of testing hypotheses seems to follow a developmental change from 4 to 6 years of age. Furthermore, the findings show that there is an emerging metacognitive understanding for epistemic states in preschool years. Even the younger preschoolers show reflective abilities for how evidence is relevant to refute claims. Around 5 and 6 years preschoolers show reflective awareness of uncertainty regarding their knowledge states due to confounded evidence. Overall, the foundational abilities for scientific reasoning are developing between 4 to 6 years, older preschoolers can coordinate hypotheses with evidence and show reflective awareness of the relation between their epistemic states and evidence.

Table of Contents

Tables ..... **XV**

Figures ..... **XVII**

1. Introduction **1**

1.1 Overview of the Thesis 3

2 Literature Review **5**

2.1 Scientific Reasoning 5

2.2 Development of Scientific Reasoning 9

2.2.1 Foundational Abilities 10

2.2.1.1 Theory–Evidence Coordination 10

2.2.1.2 Metacognitive Understanding of Theory–Evidence Coordination 12

2.2.1.3 Conceptual Clarification 14

2.2.1.4 Conclusion 16

2.2.2 Empirical Findings on Scientific Reasoning 16

2.2.2.1 Evidence Evaluation 17

2.2.2.2 Hypothesis Testing 24

2.2.2.3 Self-Directed Experimentation 32

2.2.2.4 The Relations between Scientific Reasoning and Other Cognitive Skills 33

2.2.2.5 Conclusion 36

2.3 Causal Learning in Young Children 38

2.3.1 The Theory Theory 38

2.3.2 Empirical Studies on Causal Learning in Young Children 39

2.3.2.1 Sensitivity to Statistical Sampling Patterns 40

2.3.2.2 Forming Causal Relations 41

2.3.2.3 Exploratory Play 48

2.3.2.4 Conclusion 58

2.4 Metacognition in Preschool Age 58

2.5 The Child-as-Scientist View 62

2.6 The Aim of the Thesis 67

3 Study 1: Young Children Selectively Make Interventions in response to Epistemic and Practical Goals 73
3.1 Study 1a 77
3.1.1 Research Questions of Study 1a 79
3.1.2 Method 79
3.1.2.1 Participants 79
3.1.2.2 Materials 80
3.1.2.3 Design 81
3.1.2.4 Procedure 81
3.1.2.5 Coding 85
3.1.3 Results and Discussion 86
3.2 Study 1b 91
3.2.1 Research Question of Study 1b 92
3.2.2 Method 93
3.2.2.1 Participants 93
3.2.2.2 Materials 93
3.2.2.3 Design 94
3.2.2.4 Procedure 94
3.2.2.5 Coding 95
3.2.3 Results and Discussion 95
3.3 Study 1c 97
3.3.1 Research Question of Study 1c 98
3.3.2 Method 99
3.3.2.1 Participants 99
3.3.2.2 Materials 99
3.3.2.3 Procedure 99
3.3.2.4 Coding 99
3.3.3 Results and Discussion 100
3.4 General Discussion 102

4 Study 2: Hypothesis Testing and Argumentation from Evidence in Young Children 108
4.1 Research Questions of Study 2 113
4.2 Method 114

4.2.1 Participants ... 114
4.2.2 Materials ... 115
4.2.3 Procedure ... 116
4.2.4 Coding ... 120
4.3 Results ... 123
4.4 Discussion ... 132

5 Study 3: Young Children's Understanding of Evidence as an Epistemic Category ... 141
5.1 Research Questions of Study 3 ... 146
5.2 Method ... 147
5.2.1 Participants ... 147
5.2.2 Materials ... 147
5.2.3 Design ... 148
5.2.4 Procedure ... 150
5.2.5 Coding ... 154
5.3 Results ... 158
5.4 Discussion ... 165

6 General Discussion ... 170
6.1 Summary of the Three Studies ... 170
6.2 Synthesis of the three studies ... 174
6.3 Implications ... 179
6.4 Future Research Directions ... 182
6.5 Conclusion ... 186

References ... 189

Tables

Table 3.1 Logistic Regression Predicting Likelihood of Choice of Object based on Condition, Age and Condition x Age Interaction 88

Table 3.2 Percentage and Proportion of Children in Three Age Groups' Choosing the Effect-Unknown Object by Condition 88

Table 3.3 Percentage and Proportion of Choice of Correct Object by Memory Score 90

Table 3.4 Percentage and Proportion of Choice of Correct Object by Memory Score in Study 1b 97

Table 3.5 Percentage and Proportion of Choice of Correct Object by Memory Score in Study 1c 101

Table 4.1 Evidence Characteristics in terms of the Salience of the Causal Factors and Objects in Each Phase of the Study 116

Table 4.2 Proportion and Percentages of One-Variable Hypothesis Testing in Phase 3 126

Table 4.3 Proportion and Percentages of Children's Verbal Counter Arguments in the Heavy Counter Argumentation Phase by Age Group and Content 127

Table 4.4 Proportion and Percentage of the Competency Level in Heavy Counter Argumentation Phase by Age Group 128

Table 4.5 Proportion and Percentages of Children's Verbal Comments for False Color Hypothesis by Age Group 129

Table 4.6 Proportion and Percentage of the Competency Level in Blue Counter Argumentation Phase by Age Group 130

Table 4.7 Proportion and Percentages of Children's Individual Level of Competence in Heavy Counterargumentation by Hypothesis Testing Pattern 131

Table 4.8 Proportion and Percentages of Children's Individual Level of Competence in Blue Counterargumentation by Hypothesis Testing Pattern 131

Table 5.1 Proportion and Percentage of Children's Knowledge Status Scores in the Familiarization Trials 160

Table 5.2 Percentage and Proportion of Children`s Type of Justification in the Trials of Confounded Evidence 161

Table 5.3 Frequency and Proportion of Children's Object Choice in the Confounded and Unconfounded Condition 163

Table 5.4 Percentage and Frequency of Children's Justifications Who Chose the Target Object 164

Table 5.5 The Distribution of Children's Justification for Choosing the Novel Object 164

Table 5.6 Individual Consistency for Object Type (Target vs. Novel) in the Forced Choice 165

Figures

Figure 3.1 Depiction of materials of Study 1a. 81

Figure 3.2 Schematic display of the effects of the cubes in the learning phase. 83

Figure 3.3 Schematic display of the experimental procedure of Study 1a. 85

Figure 3.4 Percentages of object choice by condition in Study 1a. 87

Figure 3.5 Percentages of object choice by condition in the younger group. 89

Figure 3.6 Percentages of object choice by condition in the older group. 90

Figure 3.7 Percentages of object choice by condition in Study 1b. 96

Figure 3.8 Percentages of object choice by condition in Study 1c. 101

Figure 4.1 Exemplars of materials of Study 2. 115

Figure 5.1 Schematic display of the main procedure of Study 3. 150

Figure 5.2 The presentation order of the cubes and their effects in the learning phase. 152

Figure 5.3 Questions in the familiarization trials in response to children's answers. 153

Figure 5.4 Questions in the experimental phase in response to children's answers. 154

Figure 5.5 Frequencies of information seeking and knowledge statements in the Confounded and Unconfounded Condition. 159

1. Introduction

Understanding the development of scientific reasoning is important both for increasing our knowledge about how the human mind works and fostering scientific reasoning skills starting from early ages. Scientific reasoning is a product of human psychology arising from our motivation to understand and explain the world we live in; it allows us to make informed predictions about the future or hypothetical worlds. The accumulated knowledge gained as a result of scientific reasoning has a great impact on our lives. Understanding how the developing human mind gives rise to such extensive knowledge and advanced predictive power about the world is a compelling endeavor in itself. Furthermore, fostering scientific reasoning skills has become one of the important goals of education. In today's world, with easy access to knowledge of all kinds via the internet, memorizing knowledge is no longer meaningful. Instead, evidence-based, critical reasoning skills gained critical importance (e.g., Trilling & Fadel, 2009). Understanding the nature and the development of scientific reasoning is essential in order to inform educational approaches with respect to learners' strengths and weaknesses.

Although many of the cognitive processes used in scientific reasoning are also common in daily life reasoning (e.g., inductive reasoning, analogy, logical thinking); research on scientific reasoning has demonstrated that this skill is challenging across all ages, especially for younger children (e.g., Dunbar & Klahr, 1989; Kuhn, Amsel, & O'Loughlin, 1988; Penner & Klahr, 1996; Schauble, 1990; Tschirgi, 1980). Children often mistakenly evaluate evidence, as they show poor performance in designing unconfounded experiments or providing evidence-based arguments although the tasks do not require particular background knowledge. Research shows marked developmental changes in early adolescence (e.g., Bullock, Sodian, & Koerber, 2009;

Croker & Buchanan, 2011; Kuhn et al., 1988). This brings forward the claim that, apart from their deficits in executive function, attentional focus, and control, young children may be lacking foundational cognitive mechanisms to be able to reason scientifically (Kuhn, 1989).

Interestingly, a recent line of research has demonstrated that young children have powerful learning mechanisms that resemble scientific reasoning. Young children can form causal relations from covariation patterns, they are sensitive to ambiguity in evidence, and they make informative interventions in order to reveal causal relations that are not readily available to them (e.g., Cook, Goodman, & Schulz, 2011; Gopnik, Sobel, Schulz, & Glymour, 2001; Legare, 2012; Schulz & Bonawitz, 2007). These early abilities closely resemble the epistemic practices in science (e.g., Gopnik, 2012; Schulz, 2012). This brings forward to the question: can young children reason scientifically? Imagine minimizing the necessary background knowledge; to what extent the strategies and mechanisms that young children use (e.g., isolation of variables in the case of causally ambiguous evidence; making predictions based on covariation data) are similar to the epistemic practices of science (e.g., experimentation, evidence evaluation)?

The answers to these questions are surely based on how scientific reasoning is conceptualized. Theoretically, it has been proposed that there are two foundational abilities: the ability to coordinate theories, hypotheses and evidence; and metacognition of knowledge seeking and formation processes (Kuhn, 1988; 1989). The aim of this thesis is to investigate the development of hypothesis–evidence coordination and reflective abilities at preschool age. The development of these two essential abilities was investigated in three epistemic activities; namely, hypothesis testing, evidence evaluation, and argumentation from evidence. These three epistemic activities were selected due to their significance for knowledge gathering processes. This thesis aims to

contribute to the existing literature in two ways. Firstly, few studies investigated the development of scientific reasoning in the preschool years. As a result, our knowledge of the phenomenon is very limited. Investigating the development of the hypothesis–evidence relation in the preschool years expands our knowledge on the development of early scientific reasoning skills. With three empirical studies, this thesis aimed to shed light on the development of essential reasoning skills in different epistemic activities. Secondly, this research has been built upon both the findings from causal reasoning and scientific reasoning research. Causal reasoning is critical for scientific reasoning skills; however, to date, the two research areas remain distinct. In this respect, this thesis can be considered as a bridge between the two areas and may pave the way for future investigations informed by findings of both areas.

1.1 Overview of the Thesis

This thesis starts with a Literature Review mainly on scientific reasoning and causal learning (Chapter 2). In this chapter, theoretical views and empirical studies on scientific reasoning and early causal learning will be provided. While the scientific reasoning part comprises a short summary of approaches to scientific reasoning throughout out the history, the remaining parts focus on the development of scientific reasoning from the preschool years till adulthood; and young children's causal reasoning abilities. This section ends with a discussion of the child-as-scientist metaphor and the aim of the present thesis. Next, in three separate chapters (Chapter 3, Chapter 4, & Chapter 5), three empirical studies on preschoolers' scientific reasoning skills will be provided. Each of these chapters introduces the state-of-the-art in the existing literature with respect to particular research questions. This is followed by a description of the methodologies, results, and discussion of the findings of each empirical study. In the

General Discussion section (Chapter 6), the findings of the three studies will be summarized, and a synthesis of the three studies with respect to existing literature will be provided. At the end, potential implications of the present findings to practical areas will be proposed, and open research questions for future research will be presented.

2 Literature Review

2.1 Scientific Reasoning

Scientific reasoning can be shortly described as "…the reasoning and problem-solving skills involved in generating, testing and revising hypotheses or theories, and in the case of fully developed skills, reflecting on the process of knowledge acquisition and knowledge change that results from such inquiry activities" (Morris, Croker, Masnick, & Zimmerman, 2012, p. 61). It has been conceptualized as both reasoning about content knowledge produced within science domains such as physics and biology; and some domain-general cognitive abilities and strategies that can be used across different disciplines for knowledge construction (Dunbar & Fugelsang, 2005). Content knowledge in specific fields has been the result of the interplay of knowledge accumulation and conceptual change in specialized domains throughout years (e.g., quantum theory, evolutionary theory) and expertise in such domains is a result of long years of training. Apart from scientific content knowledge, there are domain-general scientific reasoning skills which are some general epistemic practices that can be applied across different disciplines (also in daily life) for knowledge seeking purposes. In this section, our goal is to lay out some prominent approaches to scientific reasoning in the areas of learning and cognitive sciences.

The very first systematic investigations on scientific reasoning were mostly on the scientific discovery process with the special focus on discoveries in scientific practice domains such as physics and biology. For instance, the Gestalt psychologist Wertheimer (1945) interviewed Einstein about his discovery of the theory of relativity. The aim of this investigation was to examine how Einstein overcame thinking within the

theoretical boundaries of Newtonian physics, which paved the way for theory change in the domain of physics. In the 1950s, the idea that scientific reasoning relies on cognitive processes that are frequently used in the daily life such as problem-solving, causal reasoning, and analogy was proposed. Bruner, Goodnow, and Austin (1956) suggested that science is a process of hypothesis testing and collection of evidence in order to assess category membership of various phenomena. Wason (1968) investigated the nature of hypothesis testing strategies by adopting Popper's criteria for the scientific method—science should focus on disconfirming theories rather than trying to confirm them. Wason developed a task in which participants generated and tested hypotheses. Their study was the first showing a confirmation bias in people: the tendency to conduct tests to confirm one's own hypothesis rather than to disconfirm it. Simon and Newell (1970), who are the pioneers of the computational approach in the cognitive sciences, conceptualized scientific reasoning as a form of problem-solving. Similar to earlier researchers, they also suggested that it is not entirely different from the reasoning strategies used in daily life (induction, concept formation) but it is the implementation of such strategies in more systematic ways. According to their framework, scientific reasoning is a problem-solving activity which takes place as a search between two spaces: the space of instances and the space of rules (Simon & Lea, 1974).

Klahr and Dunbar (1988) revised Simon and colleagues' framework and presented the Scientific Discovery as Dual Search (SDDS) framework, suggesting 'science as search between two spaces.' Based on their empirical studies on adult participants who were asked to find out how a robot worked; they suggested that scientific reasoning is a problem-solving activity that takes place between the space of hypotheses and the space of experiments. In this framework, the *hypotheses space* is the set of all possible hypotheses and the *experiment space* is the set of possible

experiments; each of these spaces gives feedback to each other—scientific reasoning is a cycle between these two hypothesis and experimentation spaces. Restrictions of hypotheses in the hypothesis space narrow down the possible experiments in the experiment space. The results of the experiments give feedback to the hypothesis space; which in turn leads to further restriction of hypotheses. Furthermore, their empirical studies suggested that there are individual differences in how people employ these two spaces. Some people, *theorists*, first formulate hypotheses and then conduct experiments to test those hypotheses; whereas some people, *experimenters*, begin with exploratory experimentation without setting hypotheses. According to the feedback they receive from exploration, they formulate hypotheses.

Studying the development of scientific reasoning raises the question of what is the essential aspect of scientific reasoning that needs to be developed. Kuhn (1988) suggested the most essential and indispensable aspect of scientific reasoning is theory and evidence coordination. As the core of science is about developing theories, generating evidence to test those theories, and revising them based on evidence, the key competence is the understanding of the interplay between theories and evidence. Kuhn's contemplation of scientific reasoning has been influential on theorizing and research on the development of scientific reasoning. This approach to the development of scientific reasoning will be explained in further detail in the development of scientific reasoning part of the thesis.

Scientific reasoning is not only what scientists do, but it is a standard for accurate and reliable means of knowledge seeking. In this respect, science education has been gaining more and more importance in the recent years (Next Generation Science Standards Lead States, 2013). Understanding the nature of science and being able to evaluate the reliability of knowledge claims are defined as 21st-century skills (Trilling &

Fadel, 2009). Use of scientific knowledge and epistemic activities in practice domains (e.g., medicine, social work) is a goal in order to facilitate thinking and decision making in domains different than science (e.g., Ghanem, Kollar, Fischer, Lawson, & Pankofer, 2017; Patel, Arocha, & Zhang, 2005). Fischer et al. (2014) presented an overarching framework of scientific reasoning for different science and practice domains. They classified three modes of scientific reasoning: (a) theory building which would serve for further knowledge acquisition, (b) science-based reasoning and argumentation in practice in which practitioners use scientific epistemic activities in their decision making processes, and (c) the use of scientific theories and reasoning strategies in artifact development process. Additionally, Fischer and colleagues presented eight epistemic activities which encompass scientific reasoning activities across different disciplines and areas; namely, problem identification, questioning, hypothesis generation, construction and redesign of artifacts, evidence generation, drawing conclusions, and communicating and scrutinizing. Their framework proposes a broad, common framework which can be used across several different practices and aims to establish better communication between the domains.

Scientific reasoning is a complex set of abilities and activities. It is challenging to provide one definition that would encompass the complex nature of it. The activities mentioned earlier—discovery, assessing category membership of phenomenon, hypothesis testing—are all epistemic activities of scientific reasoning but the scientific reasoning is not limited to them. This thesis does not aim to provide a conclusive definition of scientific reasoning. It adopts a broader definition for scientific reasoning as *intentional knowledge seeking* (Kuhn, 2010) due to its developmental focus, although such a broad definition bears the danger of being too general and encompassing other phenomena that would not be considered as scientific reasoning. This broader definition

provides a suitable framework since this thesis focuses on the developmental origins. For the purposes of this thesis, the core epistemic practices, namely, hypothesis generation, hypothesis testing, and evidence evaluation are the focus in characterizing scientific reasoning. The following sections will provide further discussion about the approaches to the development of scientific reasoning.

2.2 Development of Scientific Reasoning

Theorizing on the development of scientific thinking goes back to Piaget's theory of cognitive development. Based on their studies with children of different ages, Inhelder and Piaget (1958) suggested that scientific thought only appears in adolescence with the beginning of the formal operational stage, as a consequence of the development of second-order thinking—the ability to *operate on operations*. Piaget's theory characterizes scientific thinking as a domain-general skill mostly based on logical thinking. Post-Piagetian research, however, generated evidence contrary to Piaget's conceptualization of scientific reasoning from several aspects. Firstly, researchers argued that scientific reasoning cannot be equated to logical thinking (Cheng & Holyoak, 1985; Cheng, Holyoak, Nisbett, & Oliver, 1986). Participants in reasoning studies do not approach problems with deductive logic; rather they have a pragmatic approach considering goals and actions (Cheng & Holyoak, 1985; Kuhn, Pennington, & Leadbeater, 1983). Moreover, studies demonstrated that children possess causal and scientific reasoning skills contrary to the characterization of *stage theory* claiming to show that young children are irrational, noncausal thinkers. Young children are able to form causal relations (Bullock, Gelman, & Baillargeon, 1982) and understand analogical relations (Goswami, 1991). When they were presented with age-appropriate tasks, elementary school children were found to have an understanding of hypothesis testing

(e.g., Bullock & Ziegler, 1999; Sodian, Zaitchik, & Carey, 1991). Put together, Piaget's theory for the development of scientific reasoning does not accurately characterize the psychological mechanisms for scientific reasoning and the developmental trajectory of scientific reasoning abilities.

2.2.1 Foundational Abilities

Scientific reasoning is a combination of a set of various cognitive skills and performance is very much dependent on task complexity and content domain. Due to its broad nature, it is challenging to point out a developmental trajectory of scientific reasoning. In this respect, one approach is to characterize what foundational abilities are indispensable for scientific reasoning in order to track down the developmental origins and the trajectory of scientific reasoning.

2.2.1.1 Theory–Evidence Coordination

> "A central premise underlying science is that scientific theories stand in relation to actual or potential bodies of evidence, against which they can be evaluated. Reciprocally, scientific 'facts' stand in relation to one or more actual or potential theories that offer a vehicle for their organization and interpretation." (Kuhn, 1988, p. 3)

Kuhn (1988) proposed that the essential characteristic of science is the interplay of theories and evidence. The role that evidence plays in knowledge construction is the key aspect of the scientific approach that differentiates it from other knowledge construction practices such as philosophy. Scientific knowledge is constructed as a result of a continuous cycle between theory and evidence. Evidence is a means to make inferences about the veracity of theories. Theories are formulated based on evidence, which in turn leads to the collection of new evidence which further feeds back into

theories and results in theory revision. This cycle between theories and evidence is at the core of the scientific approach and therefore, the understanding of the inferential relation between theories and evidence is essential for scientific reasoning (Kuhn, 1988; Sodian et al., 1991).

Theory–evidence coordination is at the heart of scientific reasoning; therefore, mature forms of scientific reasoning can only develop if people can have distinct concepts for the epistemic categories of theories and evidence; and if they can understand the inferential relation between the theories and evidence. This necessity for distinct concepts for theories and evidence stems from empirical studies of Kuhn and colleagues in which people often confuse theories with evidence. For instance when the task was to provide a causal theory for a phenomenon (Kuhn, 1991), e.g., the reasons for why some children fail at school, and participants were asked to generate evidence, they often confused their subjective theories with evidence. Rather than referring to evidence as empirical means to support or oppose theories, many participants just elaborated on their theoretical view without referencing to evidence; some participants provided examples or descriptive instances in line with their subjective theory (called *pseudo-evidence*). Only a small group of participants were able to refer to evidence as a means to support or oppose the veracity of theories. These findings emphasize how essential it is to differentiate theories and evidence and to properly coordinate them for scientific reasoning.

Kuhn (1989) argues that if the development of scientific reasoning is considered as a continuum, at the lowest end, there is the inability to differentiate theories from evidence. In this level of understanding, individuals do not have distinct representations of theories and evidence; rather "…there is melding of the two into a single representation of 'the way things are." (Kuhn, 1989, p. 679). In this respect, evidence is

considered as an instance of the theory and how evidence bears on the veracity of theories is not understood. Kuhn argues that young children are at this level, they have a conceptual deficit in representing theories and evidence separately. Due to this conceptual deficit, young children cannot reason scientifically. At the highest end of the continuum, there is the successful differentiation of theories from evidence. At this level, individuals understand how evidence bears on the veracity of theories, they are able to *bracket* evidence as separate concept different from their prior theories, and they can correctly evaluate theory-violating evidence. Studies by Kuhn and colleagues (Kuhn et al., 1988; Kuhn, 1991) demonstrated that although adults are better at this distinction, they are far from perfect. They often confuse theories and evidence and meld them into one single representation. College students have shown the best performance in differentiating and coordinating theories and evidence.

2.2.1.2 Metacognitive Understanding of Theory–Evidence Coordination

Kuhn (1988) postulated that metacognitive understanding—being able to reflect on knowledge acquisition and formation processes—is necessary for scientific reasoning. She expressed that this idea has been built on Piaget's theory of second-order thinking that is characteristic of the formal-operational stage: the ability to make operations on operations; and it is similar to Vygotsky's (1962) conceptualization of the higher intellectual functions: *reflective awareness* and *deliberate control*. According to Kuhn (1988), scientific thinking requires metacognitive awareness and control of the theory formation and revision process. Put differently, scientific reasoning necessitates being able to think *about* theories (Kuhn, 2010). People should be able to reflect on the relations between theory and evidence and understand the nature of epistemic categories.

Kuhn (1988) explained the importance of metacognition with an example from scientific reasoning. The *control of variables strategy* (CVS) has been frequently studied in the literature examining experimentation skills. It is the strategy to reveal causal effects of a *predictor variable* on an *outcome variable* by manipulating the different levels of the predictor variable while keeping other causally relevant variables constant and observing the changes in the outcome variable. Considering the steps of CVS that should be executed—keep other variables constant, manipulate the predictor—this strategy seems relatively simple in regard to information processing demands (Case, 1978b as cited in Kuhn et al., 1988). However, studies demonstrated that transferring CVS skills that are learned in one context to other contexts is low (e.g., Dean Jr. & Kuhn, 2007). Kuhn (1983) argued that successful execution of CVS requires, firstly, the ability to execute the manipulation and, secondly, knowing why this strategy works at all; and the challenging task is actually the latter—understanding why certain ways of knowledge seeking are better than others, and being able to choose informative strategies for knowledge gain. This requires a metacognitive understanding of epistemic processes and it is crucial for knowledge formation. Without this understanding, executing the proper manipulations of variables would not make sense because, from the perspective of the reasoner, it is unclear why this strategy is informative at all.

Empirical studies and educational approaches have supported the view that metacognitive abilities are important for scientific reasoning. To begin with, studies showed that metacognitive understanding is positively related to the use of appropriate reasoning strategies (Amsel et al., 2008) and metaconceptual understanding is a predictor of better scientific argumentation skills (Bullock et al., 2009; Kuhn, Iordanou, Pease, & Wirkala, 2008) and scientific problem solving (Rozencwajg, 2003). Furthermore, metacognitive abilities play a critical role in learning and education (See

Sodian & Frith, 2006 for a discussion). It has been frequently shown that facilitating students' metacognitive understanding via interventions has positive effects on scientific understanding (e.g., Khishfe & Abd-El-Khalick, 2002; Michalsky, Mevarech, & Haibi, 2009; Sodian, in press; White & Frederiksen, 2000; Zion, Michalsky, & Mevarech, 2005; Zohar & Peled, 2008). Considering that scientific reasoning is challenging and some form of instruction is necessary for the development of mature forms of scientific reasoning, metacognitive abilities gain even more importance.

Put together, from a developmental perspective, examining when and how the ability to reflect on theory–evidence distinction and coordination develops is critical for the investigations on the development of scientific reasoning.

2.2.1.3 Conceptual Clarification

Before moving further, it is critical to clarify the term *theory* and the related concepts. According to Kuhn and Pearsall (2000), the central characteristic of theories is that they are falsifiable by empirical evidence. They proposed that there are four levels of theories which vary based on their complexity. The first level constitutes category claims such as "plants are living things." The second level constitutes event claims such as "the plant died." The first and second level of theories can be considered as similar to beliefs. The third level constitutes causal explanatory claims; they state a causal relation between a phenomenon to another, such as "the plant died because of inadequate sunlight." The fourth level constitutes causal explanatory systems which describe and explain the interaction of multiple variables. The term theory in Kuhn and Pearsall's operationalization encompasses all these four levels of mental concepts.

For several reasons, this thesis will not follow Kuhn and Pearsall's terminology. Firstly, the development of the ability to represent uncertainty and falsifiability is important for developmental research; however, this is not emphasized in Kuhn and

Pearsall's conceptualization. Theory of mind research demonstrates that around 4 years, children start to represent that others might have false beliefs (different from reality) based on evidence they observe. This development is informative for the development of scientific reasoning because it shows the developmental origins of representing that people may have different mental representations based on evidence (Sodian et al., 1991; Kuhn & Pearsall, 2000). It is a plausible hypothesis that this understanding is developmentally preceding the ability to represent that the falsifiability and uncertainty of propositions and this is critical for hypothetical thinking (Kuhn & Pearsall, 2000; Ruffman, Perner, Olson, & Doherty, 1993). In this respect, using terminology that explicitly emphasizes the difference between mental concepts regarding the representation of falsifiability and uncertainty would be more useful. Secondly, this thesis mainly focuses on causal relations and do not aim to investigate children's theories in different levels of complexity. Therefore, such differentiation in complexity is not required for the purposes of the present thesis.

For the reasons mentioned above, this thesis will follow conceptualization described by Sodian et al. (1991) with the focus on beliefs, hypotheses, and theories as forms of different mental concepts. Beliefs are mental representations of the world. They can be any representation of how the world works, e.g., the world is flat. Hypotheses are also mental representations, but their critical characteristic is that they are subjected to confirmation or disconfirmation. Having a hypotheses entails being able to think about alternative states of reality since a hypothesis, in the simplest terms, can be correct or incorrect. Therefore, having a hypothesis that the world is flat entails thinking about alternative possibilities, e.g., the world is round. Having a hypothesis, rather than a belief, is acknowledging that there is uncertainty about the veracity of the hypothesis. In this respect, being able to represent statements as hypotheses is critical

for the development of scientific reasoning. This thesis reserves the term theory for the Level 4 theories in Kuhn and Pearsall's hierarchy that theories are a set of coherent, interrelated beliefs or hypotheses with causal explanatory functions. Put together, beliefs, hypotheses, and theories are mental representations that fall into one epistemic category. From the perspective of developmental research, using distinct terms (i.e., beliefs, hypotheses, and theories) is less confusing because how each of these mental representations develops are principally informative for the investigations on the development of scientific reasoning[1].

2.2.1.4 Conclusion

The ability to differentiate beliefs and hypotheses from evidence and to coordinate beliefs, hypotheses, and theories with evidence is essential for scientific reasoning. Additionally, it is necessary to have a metacognitive understanding of this distinction and the epistemic relations between mental concepts and evidence. Theories of development of scientific reasoning should examine the developmental origins of the ability to distinctly represent hypotheses and evidence, as well as the ability to understand the empirical relation between hypotheses and evidence. Following sections will provide an overview of the studies on the development of scientific reasoning. Although evidence from middle childhood, adolescence, and adulthood will be presented, the main focus will be on the studies on early childhood.

2.2.2 Empirical Findings on Scientific Reasoning

Empirical studies on scientific reasoning have been distinguished based on the epistemic activities they examine. Several studies explored the whole cycle of the

[1] These concepts have been used interchangeably in the literature. When describing studies in the literature, the Literature Review chapter of this thesis will be faithful to the terms used in the original papers.

knowledge acquisition process. These studies investigated more than one epistemic activity and examined how several epistemic activities play a role in the inquiry process (e.g., Klahr, Dunbar, & Fay, 1990; Klahr, Fay, & Dunbar, 1993; Kuhn et al., 1995; Kuhn, Pease, & Wirkala, 2009; Penner & Klahr, 1996; Schauble, 1996). Other studies have investigated children's competence for specific epistemic activities. Across all the epistemic activities, the two particularly investigated epistemic activities in the scientific reasoning literature are evidence evaluation and experimentation (Zimmerman, 2000, 2007). The ability to successfully evaluate evidence or test hypotheses is very much dependent on the background knowledge in the specific area in question. To illustrate, it is very unlikely for a layperson to evaluate evidence or design experiments in science domains. In this respect, studies on the development of scientific reasoning mostly use simple daily life contexts where participants do not require background knowledge. In the following section, empirical studies on evidence evaluation, hypothesis testing, self-directed experimentation; and studies on the relations between scientific reasoning and other cognitive abilities will be summarized.

2.2.2.1 Evidence Evaluation

Evidence evaluation is the assessment of how certain pieces of evidence bear on the truth of beliefs, hypotheses, and theories. Skillful evidence evaluation necessitates a differentiated understanding of hypotheses and evidence, and the ability to coordinate hypotheses with evidence. Studies in the field mostly investigated evaluation of covariation data patterns from the preschool years to adulthood in order to investigate its developmental progression. Participants in these studies have been presented with covariation data patterns and were asked to make causal judgments based on the data patterns. There has been particular emphasis on the effects of intuitive theories on how people evaluate evidence—how data patterns consistent or inconsistent with people's

prior theories influence the way they evaluate evidence. Another line of research investigated the role of causal mechanism information on evidence evaluation and how people integrate causal mechanism and covariation information when they evaluate evidence.

***Understanding covariation*.** The ability to make causal judgments from covariation data patterns has been the most frequently studied ability within the area of evidence evaluation. Kuhn et al. (1988) had a comprehensive study on evidence evaluation skills of 10- to 12-years-olds, 14- to 16-year-olds, and adults. They investigated how participants' prior theories influence their data evaluation. Based on the results of a pilot study, researchers used two variables that people think are causally-related (e.g., type of fruits) and two variables that people think are causally irrelevant (e.g., type of potato) to catching colds. In the original study, participants were presented with covariation data and asked to evaluate evidence patterns. Their responses were coded as either *evidence-based* or *theory-based* response. Evidence-based responses consisted of the cases when participants made their evaluations based on evidence. Theory-based responses were the cases when participants referred to their prior theories without referring to evidence as a means to support or oppose to theories. In further studies, Kuhn et al. investigated the effects of explicit instruction, the presentation format of evidence (real objects vs. pictorial presentation), task instructions, and the reciprocal cycle of the task on evidence evaluation skills.

The main findings of the studies suggested a marked developmental trend in the coordination of theories and evidence from middle childhood to adolescence; yet even the performance of adults was far from perfect. Participants' prior theories influenced very much how they evaluated the evidence at hand. When evidence was inconsistent with their prior theories, there was a tendency to ignore or distort inconsistent evidence

and selectively attend to consistent evidence. Moreover, even the participants who revised their theories after seeing inconsistent evidence did not show metacognitive understanding of the revision of their theories as a consequence of inconsistent evidence. Kuhn and colleagues argued that this poor performance was due to having an undifferentiated representation of theories and evidence. In a similar study, Amsel and Brock (1996) investigated evidence evaluation with fewer variables and less complex covariation patterns (perfect covariation vs. zero covariation) in a similar age range. Results suggested a developmental progression in evaluating inconsistent evidence. Children provided less evidence-based justifications and more theory-based justifications than adults when evidence was inconsistent with their prior beliefs. On the other hand, there was no developmental difference in the evaluation of consistent evidence. The findings of Amsel and Brock demonstrated that children's increased tendency to give theory-based justifications and to ignore inconsistent evidence is present also in less complex covariation patterns.

One medium for presenting covariation evidence is contingency tables, and abilities for evaluating covariation evidence presented by contingency tables have been one of the research questions. There have been several strategies that people make judgments based on contingency tables (i.e., the sum of diagonals, conditional probability). In two studies, Shaklee and colleagues (Shaklee & Mims, 1981; Shaklee & Pazsek, 1985) investigated children's, adolescents', and college students' abilities to evaluate contingency tables. Shaklee and Mims (1981) found out that participants' strategies changed depending on age and the complexity of the required strategy to make correct judgments. In another study, Shaklee and Pazsek (1985) investigated the abilities of elementary school children and found that, although even the 7- and 8-year-olds were able to make judgments based on contingency tables, the most advanced

strategy was observed only in very few children. A recent study by Saffran, Barchfeld, Sodian, and Alibali (2016) investigated how symmetry and asymmetry of variables influence children's evaluation of contingency tables. Children performed better in the symmetrical than asymmetrical data patterns and 10-year-olds performed better than 8-year-olds. Many children in the symmetrical condition also provided justifications for their judgment by comparing the frequencies of all four cells. In a similar study with contingency tables, Saffran, Barchfeld, Alibali, Reiss, and Sodian (submitted, as cited in Sodian, in press) investigated elementary schoolers' ability to provide explanations when they were provided with correct judgments. The results revealed that children were more accurate and consistent in providing correct justifications for their judgments in comparison to a condition when they were asked to provide their own judgments and justifications for their judgments. Put together, these results suggest that elementary school children have some basic competence in making judgments from contingency tables and they can also reflect on why they reach certain conclusions as a result of evidence which was shown by their explicit judgments.

There is far less evidence on evidence evaluation skills in preschool age. Ruffman et al. (1993) showed that 6-year-olds understand that people's beliefs depend on patterns of evidence available to them. In a series of experiments, children between 4- to 7-years of age were shown perfect and imperfect covariation evidence patterns in simple contexts (e.g., people liking green or red food). Children were first shown the true state of the evidence (e.g., imperfect covariation: five out of six people like red food). Later, the experimenter manipulated the evidence in a way that it suggested the opposite belief and asked children what a protagonist would think if he sees this manipulated pattern of evidence. If children understand that people's beliefs are shaped by the covariation evidence they observed, children's statements of the reality and

children statements of the protagonist's belief should be different. The results showed that by the age of 6 years most children have the understanding that beliefs are formed based on the patterns of evidence people observe. Since children represent two different judgments based on different evidence patterns, authors considered this as evidence for young children's differentiation and coordination of hypotheses and evidence in simple contexts.

Koerber, Sodian, Thoermer, and Nett (2005) investigated preschoolers' evidence evaluation skills further with several different tasks. One of the tasks was similar to Ruffman et al. (1993)'s "fake evidence task". Similar to Ruffman et al.'s results, in the case of perfect covariation both 5- and 6-year-olds performed better than chance, but not the 4-year-olds. Koerber et al. argued that the fake evidence task might be cognitively demanding because children need to hold two contradictory representations of the same phenomenon simultaneously. Researchers developed an easier task which does not require keeping two beliefs simultaneously but requires keeping track of a protagonist's belief revision process in response to different patterns of evidence. Researchers argued that such a task measures belief–evidence differentiation without introducing false-belief understanding demands.

In the study by Koerber et al., children were introduced to a protagonist who believed that the type of chewing gum matters for having healthy teeth. Later, children and the protagonist together observed evidence contrary to the protagonist's early belief. Children were asked what the current and early belief of the protagonist was. If children answer both questions correctly, this would suggest that they have, at least, a basic understanding of how evidence bears upon beliefs. Children at all ages performed better than chance in this task. When the false belief demands decreased, four-year-olds were able to track down their belief revision as a result of a change in evidence. In their

second experiment, Koerber et al. investigated the role of covariation pattern and prior beliefs on 5-year-olds' evidence evaluation skills using the belief revision task and found that both variables have an effect on performance. When children did not have prior beliefs about the context, they performed better compared to when the evidence was contradictory to their prior beliefs. Children performed best in the perfect covariation condition, which was followed by noncovariation; and they performed worst in imperfect covariation[2].

As part of a comprehensive cross-sectional study on the development of scientific reasoning, Piekny and Maehler (2013) conducted similar belief revision/fake evidence tasks with children from 4- to 11-year-olds. The 4- and 5- year-olds performed best in perfect covariation pattern which was followed by imperfect covariation. The participants were worst in the noncovariation data pattern. The big difference between performance on noncovariation between Piekny and Maehler and Koerber et al. was probably due to the difference in the questions used in the noncovariation condition. In the second experiment by Koerber et al., the possibility of noncovariation was explicitly mentioned in the experimental question, whereas it was not mentioned in Piekny and Maehler 's study. Children in Piekny and Maehler's study might have thought that they were expected to choose one of the choices (red or blue chewing gum) even though the correct answer was none of them. Furthermore, Saffran, Barchfeld, Sodian, and Alibali (2017) investigated preschoolers' abilities to evaluate evidence presented in contingency tables. They used age appropriated data presentations. Their results revealed that children as young as 6-year-olds have some basic ability to judge evidence patterns in

[2] In the first experiment of (Koerber et al., 2005) children's performance was very poor in the noncovariation evidence pattern. However, the results of the second experiment suggest that the low performance was due to children's expectation that they were required to state a choice. In the second experiment, firstly, children were presented with a protagonist who believes that the variable does not have an effect (noncovariation); and secondly, the main test question was asked by explicitly emphasizing that noncovariation was also a possible answer.

the case of symmetrical variables suggesting that preschoolers can coordinate their judgments with evidence in the case of simple contingency tables.

Understanding causal mechanisms. Koslowski (1996) argued that prior theories play an important role in the evidence evaluation process. Theories provide a framework for the evaluation of the overwhelming amount of evidence. Scientists frequently use theories in the process of evidence evaluation. According to Koslowski, it is misleading to conceptualize the ideal reasoner as one who does not take theories into consideration. She proposed that causal mechanism information plays a critical role in evidence evaluation and she criticized the former studies on their overemphasis on covariation evidence because covariation does not entail causation. Her argument was based on the idea that there are many covarying patterns in the world, yet we consider them causal only if there is a causal mechanism between the variables. She suggested that scientific reasoning necessitates the use of covariation and causal mechanism information interdependently. A series of studies with sixth-graders, ninth-graders, and college students showed that causal mechanism information plays a critical role in evidence evaluation judgments. In all groups, participants gave more importance to causal mechanism information than covariation. In the face of identical evidence, participants considered an implausible event causal, if they were provided with causal mechanism information.

In this section, our goal was to present an overview of the studies on evidence evaluation throughout development. The results show that there is a developmentally increasing trend starting in preschool years (Amsel & Brock, 1996; Koerber et al., 2005; Kuhn et al., 1988; Piekny & Maehler, 2013; Ruffman et al., 1993). Prior theories have great influence on people's evaluation of covariation evidence; people show worse performance when they need to evaluate data patterns inconsistent with their prior

theories in comparison to when they need to evaluate data patterns consistent with their prior theories (Amsel & Brock, 1996; Koerber et al., 2005; Kuhn et al., 1988). The younger the age, the bigger the difference between the evaluation of consistent and inconsistent patterns of data suggesting that the discrepancy between prior theories and covariation data patterns makes it harder to coordinate theories and evidence in younger ages (Amsel & Brock, 1996). The causal mechanism has been found to play an important role in causal judgments: it was found that people give more importance to causal mechanism information than covariation information when they make causality judgments (Koslowski, 1996).

2.2.2.2 Hypothesis Testing

In the simplest definition, the goal of hypothesis testing is to generate or observe evidence to decide on the veracity of a hypothesis. Properly representing epistemic goals, having distinct epistemic categories for hypotheses and evidence, being able to coordinate given hypotheses with relevant pieces of evidence and making judgments about the veracity of a given hypothesis based on evidence are at the core of hypothesis testing activity. There are several testing strategies that have been frequently used in the literature (i.e., isolation of variables, contrastive testing, positive testing). The aim of this section is to provide a general overview of the studies in the literature.

Control of variables strategy. The ability to execute the CVS has been the most frequently investigated ability across hypothesis testing skills. As mentioned before, it is the strategy to find out the effects of a predictor variable on an outcome variable by manipulating the different levels of the predictor variable while keeping other causally relevant variables constant and observing the changes in the outcome variable.

Tschirgi (1980) investigated the use of CVS in 7- to 11-year-old children. Researchers developed a story which was about a protagonist who bakes a delicious

cake with three new ingredients. The protagonist does not actually know which of the ingredients make the cake delicious, but has the hypothesis that using honey as a sweetener is the cause of the tasty cake. Children were given three choices of possible experiments and asked to choose which one of the experiments would work to find out whether honey was truly the cause of the tasty cake. One of the options, the correct one, was keeping the other two variables constant and using sweetener instead of honey. This comparison enables comparing the two results: honey is used vs. honey is not used. Another option was using honey and changing the other two variables. The third option was changing all of the ingredients. Researchers also investigated how the outcome (good vs. bad) of a hypothesized variable would influence people's choice of tests. The results revealed that outcome has great influence children's responses. When the outcome was bad (i.e., untasty cake) and the goal was to test the hypothesis that honey is the cause of untasty cake, children chose the correct test, keeping other variables constant and varying the hypothesized variable. On the other hand, when the outcome was good (i.e., tasty cake) and the goal was to test the hypothesis that honey is the cause of tasty cake; children kept the hypothesized variable constant and changed the other two variables. Authors argued that children aimed to produce a good result instead of finding out the actual cause. This strategy is considered as an engineering approach where the goal is to have a good outcome.

The Munich Longitudinal study LOGIC showed that children are better at evaluating the quality of experiments rather than generating experiments themselves (Bullock et al., 2009). In one task, children were instructed to design an experiment to test a given hypothesis. Even though third- and fourth-graders were good at designing contrastive tests (varying the hypothesized variable and contrasting different levels of the hypothesized variable), they did poorly in the controlled experiments (manipulating

one variable and keeping other variables constant). Only over 20% of the fifth-graders and over 40% of the sixth-graders were able to perform a controlled test. In another task, children were asked to choose a good experiment from several options of well- and poorly-designed experiments. Interestingly, their performance for choosing a good experiment was significantly better than their performance for designing experiments: over one-third of the third-graders, two-thirds of the fourth- and fifth-graders, and 80% of the sixth-graders chose the correct test. Moreover, among those children who chose the correct test, the majority of the children gave correct justifications for their choice (with performance increase with an age trend), indicating that they could reflect on how evidence bears on the veracity of hypotheses.

Croker and Buchanan (2011) investigated the use of CVS further in younger children, in 3- to 11-year-olds. They used a similar design as Tschirgi (1980), and additionally, they examined the interaction between outcome (good vs. bad) and children's prior beliefs. The story context was about having healthy teeth, and three variables were presented as potential causes. Researchers both manipulated the outcome (healthy vs. unhealthy teeth) and the plausibility of the hypothesis in terms of children's prior theories (Cola is the cause of good teeth vs. Milk is the cause of good teeth.). The results revealed that the plausibility of the hypothesis and the outcome (good vs. bad) influenced children's strategies at all ages. Even 4-year-olds chose an appropriate test strategy (manipulate one variable) when the evidence was consistent with their prior belief, and the outcome was good, or when the evidence was inconsistent with their prior belief, and the outcome was bad. In contrast, when the evidence was inconsistent with children's prior beliefs, children chose manipulations that were likely to lead to a positive outcome. Thus, it appears that young children do not firmly distinguish the goal

of testing a hypothesis from the goal of producing a positive effect (or avoiding a negative one).

Van der Graaf, Segers, and Verhoeven (2015) investigated the use of CVS with 4- to 6-year-old children. Researchers adapted Chen and Klahr's (1999) so-called "wooden ramp task." There were four variables; each with two levels that could be manipulated to investigate which factors influence how far the ball rolls. The variables were the ball type (heavy vs. light), the slope (steep vs. less steep), the starting point (top vs. mid-slope), and the surface of the slope (smooth vs. rough). The dependent variable was how far the ball rolls. Children could easily measure how far the ball rolls by counting the number of steps on the further part of the slope. The study examined children's performance in constructions of varying complexities. For instance, in Level 1 the experimenter asked how one variable (e.g., the ball type) influences how far the ball would roll. The experimenter set all three variables to constant and children were allowed to manipulate the ball type. The correct response, here, would be performing a contrastive test: choosing one heavy ball and one light ball and comparing how far each would roll. In each level, children were asked to arrange four of the variables one by one. In Level 2, similar to Level 1 children were asked to find out the effect of one variable; however, in this level, the experimenter only arranged two variables to constant. Therefore, correct performance requires varying the hypothesized variable and keeping the levels of one variable constant. In Level 3, children were asked to set three variables, and in the level 4, they were asked set four variables. In each level, children were asked for the influences of all four variables one-by-one. Sessions were terminated if children did all experiments at one level incorrectly.

There were some important aspects of the CVS assessment in the study by van der Graaf et al. (2015) which are important to mention. Firstly, earlier studies showed

that performance is better when children are given direct instruction about executing CVS compared to when they freely explore the task themselves (Klahr et al. 2011; Lorch Jr et al., 2010). This study was also structured in a way that the experimenter prompted children to find out the efficacies of the variables before each experiment. It is cognitively demanding to keep track of the goals when there are many variables, and prompting children may facilitate their performance. Secondly, the assessment of van der Graaf et al. (2015) was dynamic. Children received feedback after each construction. If they set the variables correctly, the experimenter told them that they did it correctly and why it was correct. If they set the variables incorrectly, the experimenter told that it was incorrect and showed children the correct way. Therefore, each of the experiments was a learning opportunity before further experiments.

Researchers both investigated the reliability of the CVS ramp task in this age group and children's CVS skills. In each level, there were four experiments. In Level 1, children did many correct experiments which showed that children were successful at contrasting two different levels of the target variable. All children did at least one correct test, so they all passed to Level 2. In Level 2, 40 out of 45 children correctly designed at least one experiment. This required keeping one variable constant and contrasting the two different levels of the target variable. In this level, although 40 children did at least one test, they were not able to perform correct tests for the most of the variables. At Level 3, 21 out of 40 children were able to correctly design at least one experiment which required keeping two variables constant and contrasting the two different levels of the target variable. At Level 4, 14 out of 21 children did at least one

correct experiment which required keeping three variables constant and contrasting the levels of the target variable[3].

One important question is how children would perform in this task if they would not receive feedback after each experiment design. As part of a bigger study investigating the relation between use of CVS and other cognitive skills; van der Graaf, Segers, and Verhoeven (2016) presented data that might be useful for answering this question. In this study, the task was exactly the same as the ramp task (van der Graaf et al., 2015); the only difference was that children did not receive any feedback for their performance. Moreover, it is important to note that the sample of this study was younger than the sample of van der Graaf et al.'s (2015) study. Van der Graaf et al. (2015) reported that there was an age effect; older preschoolers (K2 level) performed better than younger preschoolers (K1 level)[4]. The younger group in van der Graaf et al. (2015) and the sample of van der Graaf et al. (2016) was in similar age range (older 4- and younger 5-year-olds). Therefore, the comparison of these groups across the two studies might give insight regarding the effect of feedback on children's CVS use. Although we cannot reach conclusions, the descriptive results suggest that there was not a big performance difference between CVS task with and without feedback[5]. Taken together, the results of these two studies suggest that 4- to 6-year-old children have a basic

[3] We should be careful in generalizing the results of this study because the sample was recruited from a school following so-called "talent hot-bed school" approach in which science and technology were given special importance in education (van der Graaf et al., 2015).

[4] K1 and K2 correspond to two-year schooling grades in Netherlands. Authors did not provide information about the ages of children in the K1 and K2 groups. Since the age range of the full sample was ranging from 4;6 to 6;6; we assume that K1 group consisted of older 4- and younger 5-year-olds; and K2 group consisted of older 5- and younger 6-year-olds.

[5] The maximum number of possible experiments across the four levels was 16 in both studies. The total mean correct experiment score in van der Graaf et al. (2016) was 4.79 (out of 16) and the total mean correct experiment score of the performance of the younger children in van der Graaf et al. (2015) was 5.14.

understanding of coordinating hypotheses and evidence and that there is a developmental increase from 4 to 6 years of age.

Conducting conclusive tests. Sodian et al. (1991) investigated the ability to generate conclusive tests in children from 6 to 9 years. In the so-called "mouse house task," children were told a story about two brothers who disagree about the size of a mouse which lives in their cellar. The brothers want to find out whether the mouse is a big mouse or a small mouse (Hypothesis Testing condition); however, they cannot see the mouse because it only comes out at night. The brothers decide to find out the size of the mouse by putting a mouse house in the cellar with some cheese in it. They can either use a house which has a small opening that only a small mouse can fit through or a big opening that both a big and small mouse can fit through. Children in the study were asked whether to use a mouse house with a big opening or a small opening to find out whether the mouse is a small or a big mouse. The logic behind this task is if there is the small opening and cheese is gone, this means that the mouse is small because only a small mouse can enter through the small opening. Similarly, if there is the small opening and the cheese is still there in the morning, this suggests that the mouse is big so that it cannot enter the house and eat the cheese. Therefore, choosing the house with small opening provides a conclusive test to test the hypotheses whether the mouse is big or small. On the other hand, if there is the big opening and cheese is gone does not provide any information about the size of the mouse; therefore it is an inconclusive test. The study also included "Effect Production" condition as a control, where children were asked to feed the mouse, to see whether children could differentiate between hypothesis testing and effect production. In this case, the correct answer is to choose the house with the big opening so that it does not matter whether the mouse is big or small, it can enter the house and eat the cheese.

This task requires children to understand the goals of the two conditions and choose different options depending on the goal of the task. The study was within-subjects, and the best performance was differentiating the two goals and selectively choosing the correct option in both conditions. Fifty-five percent of the 6- and 7-year-olds and 86% of the 8- and 9-year-olds chose the correct option and provided correct justifications for their choice. Therefore, the results suggest that young elementary school children have the ability to choose the conclusive test in order to generate informative evidence in line with an epistemic goal. Moreover, children's justifications suggest that they can reflect on the differentiation and coordination of hypotheses and evidence.

Piekny, Grube, and Maehler (2014) and Piekny and Maehler (2013) replicated the mouse house task with preschoolers. The former study was a longitudinal study in which children were tested when they were 4-, 5-, and 6-years-old. At all measurement points, children performed better than chance level when they were asked to feed the mouse. On the other hand, only at the sixth-year measurement point, children were able to choose the conclusive test to find out whether the mouse is small or big. Piekny and Maehler (2013) replicated the task with children from 4- to 13-years. They found a developmental increase in hypothesis testing performance. A specific look at preschooler age shows that 4-year-olds performed at chance level in both conditions and 5-year-olds performed better than chance level only in the feeding condition. The results of these two studies suggest that 4- and 5-year-olds are still not competent in selectively choosing the correct choice depending on the goal of the conditions. In both conditions, they chose the big opening, suggesting that their goal was to produce an effect rather than test hypotheses. Only around 6 years children selectively choose the correct choice.

In sum, this section provided an overview of the hypothesis testing studies with a special focus on studies on preschoolers. Studies showed that there is a developmental change in hypothesis testing abilities: older children are better than younger ones but their performance is far from perfect (e.g., Bullock et al., 2009; Croker & Buchanan, 2011; Piekny & Maehler, 2013; Tschirgi, 1980). Important factors that influence testing performance are the outcome of the hypothesized variable and prior theories (e.g., Tschirgi, 1980); and interestingly the interaction of these two factors also lead to different performance (Croker & Buchanan, 2011). Although there is a tendency for the engineering approach rather than hypothesis testing across all ages (Tschirgi, 1980), even 6-year-olds were found to be able to differentiate hypothesis testing from effect production (Piekny et al., 2014; Piekny & Maehler, 2013). A special focus on preschoolers suggests that preschoolers have basic skills to differentiate hypotheses from evidence shown by their ability to design simple tests and there is a developmental increase in the abilities of hypothesis testing from 4- to 6-years of age (Piekny et al., 2014; Piekny & Maehler, 2013; van der Graaf et al., 2015; van der Graaf et al., 2016).

2.2.2.3 Self-Directed Experimentation

Instead of focusing on one epistemic activity such as experimentation or evidence evaluation, there are studies in the literature which investigated scientific reasoning performance across several epistemic activities. In most of these studies, participants followed the whole cycle of epistemic activities. In these studies, participants generated their own hypotheses, designed experiments, evaluated the results of their experiments, made inferences based on evidence, and repeated the cycle if necessary. Some of the studies used real-world contexts (e.g., Penner & Klahr, 1996; Schauble, 1990, 1996) while some others used artificial contexts (Dunbar & Klahr, 1989; Kuhn et al., 1988). Most of these studies are with middle school children,

adolescents, and adults. The complexity of self-generated experiments makes it hard to study it in early childhood.

The main results of the studies on self-directed experimentation (e.g., Klahr et al., 1990, 1993, 1993; Kuhn et al., 1995; Schauble, 1996; see Zimmerman, 2007 for a review) suggested that middle school children did not generate hypotheses; rather they frequently provided beliefs which they aimed to confirm rather than test. They often did confounded experiments. They were unable to evaluate entire evidence. They make inferences based on a very few number of evidence cases. Their evaluation of evidence was highly distorted by their prior beliefs. When there was inconsistent evidence, they were likely to either distort or ignore it. In the cases when they revised their prior beliefs, they were not aware of the belief change which suggests that they were not able to reflect on knowledge change process due to evidence. Adolescents, on the other hand, showed better performance compared to middle school children. They were better at generating hypotheses and testing them. Data from think-aloud measures suggested that they have an understanding of what it means to have a hypothesis and to test one. In these studies, adults showed the best performance across all age groups (e.g., Klahr et al., 1993; Kuhn et al., 1995; Schauble, 1996). They were better at putting aside their prior beliefs and making inferences based on evidence. Yet, both adolescents' and even adults' performance was far from perfect.

2.2.2.4 The Relations between Scientific Reasoning and Other Cognitive Skills

The development of scientific reasoning is interrelated with the development of other cognitive abilities. Investigations on the relations between scientific reasoning skills and other cognitive abilities are particularly important because they can provide information whether scientific reasoning is a unitary construct in itself. Moreover, such investigations inform us about which cognitive abilities are related to scientific

reasoning across development and whether there are changes throughout development with respect to the relations of individual cognitive abilities and scientific reasoning. Several studies on elementary school children, and a few on preschool children, investigated the relations between scientific reasoning and other cognitive skills (Astington, Pelletier, & Homer, 2002; Mayer, Sodian, Koerber, & Schwippert, 2014; Osterhaus, Koerber, & Sodian, 2017; Piekny, Grube, & Maehler, 2013; Sodian, Kristen-Antonow, & Koerber, 2016; van der Graaf et al., 2015).

Mayer et al. (2014) investigated scientific reasoning skills, intelligence, inhibition, problem-solving, spatial skills, and reading skills in 10-year-olds. All cognitive factors but inhibition was correlated with scientific reasoning performance. Although reading comprehension, nonverbal- and verbal-intelligence had significant influence on scientific reasoning performance; they were found to be separate constructs different from scientific reasoning. Van der Graaf et al. (2016) found that inhibition and verbal working memory are indirectly related to experimentation and evidence evaluation skills through grammatical ability in 4-year-olds. They did not find any relations between scientific reasoning and visuospatial working memory, cognitive flexibility, vocabulary, and spatial visualization. The discrepancy between the results of Koerber et al. and van der Graaf et al.—the relation between scientific reasoning and inhibition—is intriguing. Considering that the two studies focused on different age groups (4-year-olds and 10-year-olds), one potential hypothesis may be that the relationship between inhibition and scientific reasoning changes across development. As children's inhibition skills develop across the childhood years, inhibition may lose its influence on older children's scientific reasoning performance. However, this hypothesis is unlikely because findings in the literature document that inhibition ability is associated with scientific reasoning in the middle childhood and adolescence years

(Kwon & Lawson, 2000; Osterhaus et al., 2017). In this respect, a more likely hypothesis, suggested by Mayer et al., is that the paper-and-pencil test used for measuring scientific reasoning skills in their study may not require high inhibition demands because the task may not trigger any prior theories or other individual motivations to act differently.

Theory of mind, the ability to represent alternative beliefs, might be a precursor of scientific reasoning (see Sodian et al., 1991; Kuhn, 2010). Osterhaus et al. (2017) assessed the relation of second-order false belief reasoning, experimentation, understanding the nature of science, inhibition, intelligence, and language abilities in 8- to 10-year-olds. Both experimentation and understanding the nature of science was correlated with general information processing skills (i.e., inhibition, intelligence, language abilities). Second-order false belief reasoning was a predictor of the nature of science understanding which was a predictor of experimentation skills. Notably, these relations were independent of the general information processing skills. Astington et al. (2002) investigated the relationship between second-order false belief reasoning and reasoning about evidence with general language and nonverbal reasoning as control variables in 5- to 7-year-olds. Second-order false belief understanding accounted for a significant amount of the variance in reasoning about evidence which was not explained by control variables. Piekny et al. (2013) found that false belief understanding at 4 years was a predictor for experimentation skills at 5 years when language, executive function, working memory, and intelligence were controlled for. Sodian et al. (2016) found that both first-order and second-order false belief reasoning at 5 years are related to experimentation skills at 8 years when verbal intelligence was accounted for. These findings support the hypothesis that development of both first-order and second-order false belief reasoning precedes the development of scientific reasoning abilities.

2.2.2.5 Conclusion

This section reviewed the empirical studies on scientific reasoning abilities across development. Contrary to the early conceptions of scientific reasoning as a late developing skill, plenty of research has shown that scientific reasoning skills emerge long before adolescence years (Bullock et al., 2009; Piekny & Maehler, 2013; Piekny et al., 2014; Ruffman et al., 1993; Sodian et al., 1991; van der Graaf et al., 2015; van der Graaf et al., 2016). The findings from evidence evaluation, hypothesis testing, and self-directed experimentation studies document that children`s abilities follow a developmental trajectory: the younger the age, worse the performance. Although adults and adolescents perform better in scientific reasoning tasks, their performance is far from perfect: they often confuse theories and evidence; they ignore or distort inconsistent evidence. These findings suggest that adults are also far from being the ideal reasoner.

The hypothesis that young children have a conceptual deficit for the foundational skills of scientific reasoning (Kuhn, 1989) is not supported by evidence. Although studies investigating young children`s scientific reasoning skills are scarce, they suggest that there is a developmental change between 4- to 6-years (Koerber et al., 2005; Piekny et al., 2014; Piekny & Maehler, 2013; Ruffman et al., 1993; van der Graaf et al., 2015; van der Graaf et al., 2016). The hypothesis that the theory of mind development is a precursor of scientific reasoning skills (Kuhn & Pearsall, 2000; Ruffman et al., 1993) has been supported by evidence (Astington et al., 2002; Osterhaus et al., 2017; Piekny et al., 2013; Sodian et al., 2016). Moreover, studies on preschoolers` evidence evaluation skills suggest that 4-year-olds have some basic abilities to coordinate evidence and beliefs; and they can reflect on the belief revision process as a consequence of a change in evidence patterns (Koerber et al., 2005; Piekny & Maehler, 2013). This finding is in

line with the theory of mind abilities which show that there is an emerging understanding in the preschool years that people might hold different beliefs about the world (e.g., Wellman, 2011). The ability to represent alternative hypotheses in relation to evidence seems to emerge later, around 6 years (Ruffman et al., 1993; Piekny & Maehler, 2013; Piekny et al., 2014). Other than the development of representing beliefs and hypotheses, several other factors have critical influence how preschoolers perform in scientific reasoning tasks. The influence of prior theories, the number of variables, evidence patterns (deterministic vs. probabilistic), the symmetry of variables are the main factors that has been empirically shown to influence preschoolers` scientific reasoning performance (Koerber et al., 2005; Saffran et al., 2017; van der Graaf et al., 2015; 2016).

Since hypothesis–evidence coordination lies at the core of scientific reasoning, studies investigating preschoolers` abilities naturally focus on children`s representations of beliefs and hypotheses by using tasks where children are expected provide judgments based on evidence or test given hypotheses in order to judge the veracity of the hypothesis. Interestingly, there is a recent line of research on causal reasoning showing young children`s powerful abilities in learning from evidence patterns. Although this research line generally does not directly investigate children`s abilities to make explicit judgments based on evidence or ability to test given hypotheses, the findings are informative for the research on the development of scientific reasoning. It is a plausible hypothesis that those early competences may precede mature forms of scientific reasoning. In the next section, causal reasoning abilities of young children will be summarized.

2.3 Causal Learning in Young Children

Assessing causal relations between variables is one of the main goals of science. In fact, causality-based accounts are highly favored by philosophers in attempts to characterize the nature of scientific explanations (Okasha, 2002). Recent studies on causal reasoning revealed that children's causal reasoning skills are underestimated by earlier accounts of cognitive development. Causal reasoning abilities are relevant and informative for the development of scientific reasoning skills, showing both human's intrinsic abilities for forming causal relations and their biases that influence this process. In this section, young children`s causal learning abilities will be summarized.

2.3.1 The Theory Theory

The theory theory has been the theoretical background for most of the studies that will be mentioned in the causal learning section. In this regard, this section will briefly summarize early and recent accounts of the theory. Early accounts (Gopnik & Wellman, 1992; Gopnik & Wellman, 2012) of the theory suggested that people have coherent, abstract, and structured representations of the world which are similar to scientific theories. Contrary to nativist theories, the theory theory argued that these theories are learned. Based on their observations and their interactions with the world, young children develop abstract, and in some way, coherent theories about the world. These theories need not to be correct, yet they provide an implicit understanding of how things are and enable them to make predictions. The accumulation of contradictory evidence results in revision or rejection of theories. The theory theory claims that learning is a consequence of constant theory formation and revision.

According to this framework, (a) children form and revise their theories based on evidence. (b) These intuitive theories have a distinctive hierarchical structure; specific

theories might be embedded in more general framework theories. (c) Theories serve for distinctive cognitive functions: they allow making predictions about the future or explain the present evidence. (d) Theories have distinctive dynamic characteristics. They are subjected to revision in the face of inconsistent evidence. (e) Theory change is a gradual process: it does not take place as dichotomous acception or rejection, rather there is a gradual change as a result of accumulating evidence. Until recently, there was no clarification in the theory with regard to the mechanisms for learning. Later, theory theorists (see Gopnik & Schulz, 2004; Gopnik & Wellman, 2012 for a detailed overview) were influenced from recent computational developments in computer science and suggested a mechanism regarding how learning and theory change take place. Especially causal Bayes nets, a subcategory of probabilistic models, have been adopted and frequently used for modeling cognitive development (e.g., Bonawitz, Griffiths, Schulz, Sun, & Miyake, 2006; Bonawitz et al., 2006; Buchsbaum, Gopnik, Griffiths, & Shafto, 2011; Sobel et al., 2004).

2.3.2 Empirical Studies on Causal Learning in Young Children

In the following section, recent empirical studies on young children's causal learning will be summarized. Most of the studies mentioned further in this part used adaptations of an experimental paradigm, so-called "the blicket detector paradigm," which was developed by Gopnik and Sobel (2000). In this paradigm, children are presented with a machine called the blicket detector, which is commonly a box in some particular form that generates an effect (plays music or lights up) when it is in interaction with certain objects. Typically, the objects range from small wooden boxes to geometrical shapes. Children observe that when objects interact with the detector, in most cases, the detector lights up or plays music; and this creates the feeling that the objects cause the effect. In most of the studies, no mechanism information was available

to participants. In fact, the objects do not cause the effect by themselves but another experimenter activates or deactivates the detector with a hidden remote control. The popularity of the paradigm is due to its flexibility for adaptations to different research questions. Moreover, it is novel and engaging for young children; and it does not require any background knowledge.

2.3.2.1 Sensitivity to Statistical Sampling Patterns

Infants are sensitive to the statistical patterns in language (e.g., Gómez, 2002; Saffran, Aslin, & Newport, 1996); and in perception (e.g., Fiser & Aslin, 2002; Wu, Gopnik, Richardson, & Kirkham, 2011). They make sampling inferences based on the statistical information (Kushnir, Xu, & Wellman, 2011; Xu & Tenenbaum, 2007; Zu & Denison, 2009). For instance, Xu and Garcia (2008) measured 8-month-old infants' looking time when the probability of a sample drawn from a population is low, and when the probability of a population was low based on a given sample that was supposedly taken from that population. The *population* in the study was a container full of white and red ping-pong balls. The ratio of one color of balls (e.g., red) to the ratio of another color of balls (e.g., white) was 70:5. In "from sample to population" experiments, infants observed a sample of four red balls and one white ball. Later, they were either shown a container mostly full of red balls or mostly full of white balls. The container with mostly red balls was a more likely population for the given sample in comparison to the container mostly full of white balls. In a "violation of expectation paradigm," infants looked significantly more to the unexpected population compared to the expected population. Subsequent experiments investigated whether infants make the same predictions in the opposite direction, predicting a sample from the population. Similar results were found in this condition, too. These findings show that infants have an inherent sensitivity for the relations between samples and populations.

2.3.2.2 Forming Causal Relations

Studies on young children's causal learning abilities demonstrated that they correctly form cause-effect relations from probabilistic or deterministic evidence patterns. The next section will summarize the recent findings on implicit evidence evaluation skills.

Understanding conditional probability. The logic of scientific experimentation and many statistical analyses (e.g., partial correlation, regression) are based on *screening off* reasoning—in the case of more than one causal variable, keeping constant one of the variables in order to see how the outcome variable is influenced[6]. Studies in adults demonstrated that adults can do the screening of reasoning (e.g., Cheng & Novick, 1990; Shanks, 1985). Gopnik, Sobel, Schulz, and Glymour (2001) investigated 2-, 3-, and 4-year-olds on the use of this type of reasoning when learning cause–effect relations from patterns of covariation. In the experimental condition (One- Cause) children saw a light box being lit up when Object A was placed on it; it was not activated when Object B was placed on it, and it was activated again when Object A and Object B were placed on it together. In the control condition (Two-Cause), children observed that Object A was placed on the box three times and the machine lit up at all three times; Object B was placed on the box three times, and the machine lit up two out of three times. Subsequently, children were asked whether Object B is an activator (blicket) or not. The number of times that Object B was associated with the light (alone and together with Object A) was equal across the two conditions.

[6] In order to investigate the causal relations between variables, conditional probabilities of the variables should be investigated. Imagine the scenario, a person has headache at nights whenever she goes to parties and drinks wine. To find out whether going to parties alone or drinking wine alone causes headache, one should look at the conditional probabilities of having wine without going to parties or going to parties without drinking wine. Reichenbach (1956) coined the term screening of reasoning for this type of reasoning (Sobel et al., 2004).

The associationist account on causal learning would predict that children would consider Object B as an activator in both of the conditions because the number of times it was associated with the light was equal across the two conditions. On the other hand, if children form causal relations considering screening off relations (i.e., the box lights up when Object A and Object B are on the box together because of Object A) they would predict that Object B was an activator in the Two-Cause condition, but not in the One-Cause condition. Results showed that young children take into account the screening-off information and do not consider Object B as an activator because it was associated with the light only when it was placed together with Object A. The same procedure was replicated in the domains of psychology, biology, and a novel domain; and children showed a similar performance (Schulz & Gopnik, 2004). Therefore, the ability to use screening-off in covarying relations seems to be a domain-general skill in early childhood[7]. Sobel et al. (2004) found that that 4-year-olds infer the effect of an object from covariation evidence even in the cases when they did not observe direct evidence for the effect of the object. Children observed that Object A and Object B were placed on the box and the box lit up; Object A was placed on the box, and the box did not light up. From this data, children inferred that Object B was an activator. Therefore, children do not necessarily need to see the direct evidence; they can make inferences based on the conditional probabilities, even in the case of indirect evidence.

In the studies mentioned so far, children were presented with deterministic evidence. Individual objects always caused the same effect. However, not all causal relations in the world appear in a deterministic way. Imagine a scenario that a person has a headache three out of five times when she drinks wine. In this case, there is no

[7] One criticism to this methodology is that maybe children do not attend to covariation evidence but only attend to the effects when the objects were placed on the machine in isolation (Cheng & Novick, 1992 as cited in Sobel et al., 2004)

deterministic relation between drinking wine and having headaches. However, the probability of the two events appearing together suggests that there is a causal relationship between drinking wine and having a headache. Our everyday life is full of such cases; causal relations between variables appear probabilistically. Even though they are not deterministic, they are informative in forming causal relations. In this respect, the ability to evaluate causal strength in the case of probabilistic causal relations is fundamental. Adults form cause–effect judgments based on probabilistic evidence (Cheng, 1997; Waldman & Hagmayer, 2001). Kushnir and Gopnik (2005) demonstrated that 4-year-olds chose an object which activated a machine two out of three times as more effective than an object which activated the machine one out of three times. This finding suggests that young children form causal relations not only from probabilistic evidence but they can take into consideration the causal strength based on the conditional probabilities.

Naïve theories. Several studies on scientific reasoning demonstrated that participants' prior theories have a significant influence on how they evaluate evidence. When evidence is contradictory to their prior theories, participants only attend to evidence which is in line with their prior theories; they distort inconsistent evidence and make faulty inferences (e.g., Koerber et al., 2005; Kuhn et al., 1988). Cognitive developmentalists suggest that even very young children have naïve domain-specific theories in several domains such as physics, psychology (Carey, 1987), or biology (Wellman & Gelman, 1992). Young children already form domain-specific theories which help them to build new information. Considering cognitive development in relation to theories has been a discussion in the field (Leslie, 1994; Scholl & Leslie, 1999; Carey & Spelke, 1994; Keil, 1995; Gopnik & Wellman, 2012), this thesis will not focus on this discussion. The critical point for the purposes of this thesis is how prior

theories influence the process of learning new causal relations. Depending on the consistency of the evidence with prior theories, prior theories might reinforce learning from evidence when they are consistent with evidence, or they might hinder the evaluation of novel evidence when they are inconsistent with evidence.

Studies investigated whether children can learn implausible causal relations from evidence. Prior research showed that 4- and 5-year-olds prefer to categorize objects based on their perceptual features over distance causality (Sobel & Buchanan, 2009) suggesting that it is implausible that the objects at a distance activate the blicket detector. Based on this finding that distance causality is implausible, Kushnir and Gopnik (2007) presented 3- and 4-year-olds with the blicket detector getting activated by an object at a distance. After observing the evidence, children learned that a toy at a distance (implausible hypothesis) is actually causally more efficacious in activating the detector rather than an object that is in direct contact with the detector (plausible hypothesis). In another study by Schulz and Gopnik (2004), children learned that talking to a machine (implausible hypothesis), rather than pressing a button (plausible hypothesis), activates a machine. The authors of these two studies argued that children can overcome their naïve theories when they observe contrary evidence and form causal relations in line with the evidence. However, in these studies the relative strength of the naïve theories concerning evidence (distance causality over contact causality or talking to a machine rather than pressing a button) is weak. Although children may come across contact causality more frequently in everyday life, there are still many examples that they can observe distance causality (e.g., remote controls, light buttons). Based on the results of these two studies, it would be premature to generalize that children can overcome or revise their naïve theories when they are exposed to contrary evidence.

Early research showed that it is challenging for young children to think that psychological states may cause physiological reactions (e.g., stress causes stomachache) (Estes, Wellman, & Woolley, 1989; Hatano & Inagaki, 1994; Notaro, Gelman, & Zimmerman, 2001; Wellman & Estes, 1986). Based on this early finding, in a study by Schulz et al. (2007), 3- to 5-year-old preschoolers were read story books where there was either a domain-plausible cause–effect relation (i.e., within-domain, physical cause, physical effect: running in the cat-tails causes itchy spots on Bambi's legs) or a domain-implausible cause-effect relation (i.e., cross-domains, psychological cause, physical effect: Bunny being scared about show-and-tell and having tummy ache). The story narrated a week of the protagonists (Bunny or Bambi), repeating the particular actions and effects across days. Stories of each day also included another event (e.g., running in the pine grove, running in the cedar trees); however, those events did not repeat over the days. In sum, children heard stories of seven days in a week in the format: First day, event AB and outcome X; second-day, event AC and outcome X; third-day event AD and outcome C ... took place. At the end of the stories, children were asked why Bambi had itchy spots in the within-domain story condition and why Bunny had a tummy ache in the cross-domains story condition. Children in the baseline condition did not hear any stories. The results revealed that 4- and 5-year-olds showed a ceiling effect in choosing the repeating event as the cause of the outcome. In the cross-domains story, they chose the repeating event as the cause more often than the baseline; however, the frequency of children choosing the cross-domains cause was significantly lower than the within-domain condition. These results suggest that children, to some extent, can form cause-effect relations from evidence which is, indeed, contrary to their naïve theories. However, learning from evidence is lower compared to evidence which is in line with their naïve theories.

In Schulz et al.'s (2007) study, contrary to the 4- and 5-year-olds, 3-year-olds did not show any belief revision effect in the cross-domain scenario. Researchers investigated if this age group's performance in cross domains would benefit from an intervention (Bonawitz, Fisher, & Schulz, 2011). The results revealed that, after the intervention, children were more likely to form cross-domain cause-effect relations even though the relation was implausible. Put together, these results suggest that prior theories have an influence on children's formation of causal relations, yet there is evidence that the effect of prior theories may be weakened to some degree after observing theory-violating evidence.

Inferring unobserved causes. Frequently in daily life, causes are not always apparent; therefore, inferring unobserved causes has a vital role in the formulation of causal theories about phenomena. Adults infer hidden causes from the causal structure (Kushnir, Gopnik, Lucas, & Schulz, 2010). Gopnik et al. (2004) demonstrated that 4-year-olds infer a hidden common cause for an effect when the patterns of evidence suggested that the individual cases are not the cause of an effect. Schulz and Sommerville (2006) showed that 4-year-olds infer unobserved hidden causes when the evidence was not deterministic. They infer generative or inhibitory hidden causes based on the evidence pattern. Evidence from earlier studies also showed that children infer unobserved causes. In a "balance study" by Bonawitz et al. (2012; will be explained in detail in Section 2.3.2.3), when children were presented with theory-violating evidence, they provided explanations referring to auxiliary variables. These results suggest that the ability to infer unobserved causes is already present in early childhood.

Diagnostic reasoning. People need to not only make predictions in the direction from causes to effects but they also need to reason in the opposite direction: what are the causes of an effect? This type of reasoning is called diagnostic reasoning and plays a

critical role in scientific reasoning. Since most of the time, there are many possible causes, diagnostic reasoning requires having an understanding of uncertainty and being able to think about alternative possibilities. The *first-order diagnostic reasoning* is when the efficacy of possible causes are known. In the *second-order diagnostic reasoning*, diagnoses are made when the efficacy of one or more possible causal variables is uncertain. This requires taking into consideration the knowledge at hand and making inferences that would best explain the effect. Both first-order and second-order diagnostic reasoning require being able to think about alternative possibilities. Yet, the latter is harder than the former; because in the latter, one does not have any direct evidence regarding variables' efficacy and this necessitates the ability to reason in the case of uncertainty (Fernbach, Macris, & Sobel, 2012).

Fernbach et al. (2012) found that even 3-year-olds have skills for first-order diagnostic reasoning. Three- and 4-year-olds were successful at revising their diagnoses when they were told that their diagnoses were wrong. While some studies suggest that second-order reasoning appear only around 8 years (Bindra, Clarke, & Schultz, 1980), Fernbach et al. (2012) showed that 4-year-olds can revise their diagnoses in the case of uncertainty in simple problems (one unknown cause). Erb and Sobel (2014) and Sobel, Erb, Tassin, and Skolnick Weisberg (2017) investigated second-order diagnostic reasoning further with 3- to 7-year-old children in several task formats (simpler: one-unknown cause; more difficult: two-unknown cause or additive effects). Both studies demonstrated a developmental trend in the preschool years in understanding uncertainty. Although 4- and 5-year-olds showed some competence in simpler tasks (Erb & Sobel, 2014); they performed around chance in tasks with increased difficulty (Sobel et al., 2017). Only around 6- and 7-years children performed better than chance in more difficult tasks. Put together, these findings suggest that even 3-year-olds have some

abilities for first-order diagnostic reasoning; they do not persist on one possibility but understand that there might be alternative causes to an effect. However, their understanding is limited to the cases when they know the possible causes and their effects. Between 4- to 7-years, there is an emerging understanding of uncertainty which affects children's diagnostic reasoning performance. Four- and 5-year-olds show basic competence when there is only one-unknown cause; whereas, only 6- and 7-year-olds can successfully perform in the case of two-unknown causes or additive tasks.

Overall, the growing number of studies has shown that preschoolers have precocious abilities for forming cause-effect relations based on both deterministic and probabilistic evidence. Children's naïve theories affect how they evaluate evidence. In the case of weaker prior theories, preschoolers revise their theories when they are presented with theory-violating evidence (Sobel & Buchanan, 2009; Kushnir & Gopnik, 2007). In the case of stronger theories, there is a significant effect of prior theories. In comparison to theory-consistent evidence, it is harder to form causal relations in the case of theory-inconsistent evidence (Schulz et al., 2007). Although there is such a difference in forming causal relations between theory-consistent and theory-inconsistent evidence; children can transfer the outcomes of evidence from theory-inconsistent evidence to novel tasks. Moreover, the abilities to infer unobserved causes based on the evidence patterns (Bonawitz et al., 2010; Gopnik et al., 2004; Kushnir et al., 2010) and to reason diagnostically about uncertain causes (Erb & Sobel, 2014; Fernbach et al., 2012; Sobel et al., 2017) suggest that young children already possess critical skills for dealing with complex causal problems in daily life.

2.3.2.3 Exploratory Play

Play is an essential source of learning, and it has a fundamental role in the childhood years. Its importance for development has been acknowledged by many

cognitive developmentalists (Piaget, 1962; Singer, Golinkoff, & Hirsch-Pasek, 2006; Vygotsky, 1978). Play behavior in childhood period is characterized as a means of practice for the adulthood years (Buchsbaum, Bridgers, Weisberg, & Gopnik, 2012). There have been numerous findings on how play is related to the development of various cognitive skills such as emotion regulation (e.g., Galyer & Evans, 2001), metacognition (e.g., Whitebread & O'Sullivan, 2012), language development (Tamis-LeMonda, Shannon, Cabrera, & Lamb, 2004), and collective intentionality (Rakoczy, 2007).

Exploratory play is also crucial for causal reasoning. Children perform interventions on their surroundings and learn from the results of their own interventions. This type of learning is considered differently than adult learning. Exploration is considered as "…learning about environment for its own sake" (Buchsbaum et al., 2012, p. 2203) and it is different from exploitation learning in which there are specific goals to achieve as a consequence of learning. In fact, Buchsbaum et al. (2017) proposed a hypothesis that an interaction between causal reasoning and exploratory behavior during more extended childhood period may have paved the way for the evolution of specialized causal reasoning skills of the humans.

Even though theories of development emphasize how children "learn from doing," there has been few empirical research until recently (Bruner, Jolly, & Sylvia, 1976; Singer, Golinkoff, & Hirsh-Pasek, 2006). In the last ten years, there have been many studies on children's learning of causal information from their exploration. Some of these studies are especially important for the development of scientific reasoning research because they provide evidence that young children are sensitive to the informativeness of evidence and they differentially intervene on their environment depending on the ambiguity of evidence. This sensitivity and differential exploratory

play is one of the important reasons why young children are considered as "little scientists" (Gopnik, 2012). In this part of the thesis, the empirical findings on preschoolers' exploratory play will be summarized.

Children intervene selectively depending on whether evidence is causally ambiguous vs. unambiguous, or whether evidence is in accordance with their naïve theories or not. Schulz and Bonawitz (2007) found out that 4- to 7-year-old children prefer playing with a familiar toy over a novel toy only when the familiar toy is causally ambiguous. When there was missing knowledge regarding the causal structure of a toy, children spent more time in exploring the toy compared to the cases when the causal structure was apparent. Furthermore, Schulz, Gopnik, and Glymour (2007) demonstrated that 3- to 5-year-old children make interventions by isolating parts of a gear mechanism to learn the causal mechanism. They found that children who isolated the machine's parts learned the causal structure better than the children who did not isolated its parts.

Naïve theories and exploration. Children's naïve theories of the world influence their exploratory play. Prior beliefs about a phenomenon restrict the quantity and the quality of the hypotheses children have. Bonawitz, van Schijndel, Friel, and Schulz (2012) investigated the influence of children's naïve theories on their exploration and explanation in an object balance paradigm. Earlier studies on children's intuitive balance theories demonstrated that there is a developmental progression: children who are younger than 6 years have no specific theories about balance, children who are at around 6 or 7 years have the theory that the geometric center is critical for balance; and children who are at around 7 or 8 years of age develop an understanding that the center of mass is critical for balance (Karmiloff-Smith & Inhelder, 1974). Based on these earlier findings, Bonawitz et al. (2012) designed an exploration paradigm which required balancing

objects on a scale toy. At the beginning, children were categorized into three different groups according to their naïve theories: "mass theorists" who had the prior theory that the block would be balanced when the objects were placed at the center of mass, "center theorists" who had the prior theory that the block would be balanced when the objects were placed at the geometric center, and "no theory" children who did not have a particular prior theory. The dependent measure was an indirect measure: children's play duration with the familiar scale toy and a novel toy. The results of the study revealed that when children were presented with evidence contrary to their existing theories, they played more with the familiar toy. To illustrate, when mass theorists observed that the block was balanced at the geometric center (inconsistent with intuitive theory), they played more with the familiar toy; whereas when they observed that the block was balanced at the center of mass (consistent with intuitive theory), they played more with the novel toy. The opposite was the case for the center theorists. On the other hand, independent of the consistent vs. contradictory evidence, no theory children played longer with the novel toy. These results demonstrate that children's naïve theories influence the way they evaluate and interact with new evidence. Children selectively explored more when the evidence was contrary to their naïve theories than when the evidence was in line with their naïve theories.

The studies by Bonawitz et al. (2012) and Schulz and Bonawitz (2007) demonstrated that young children explore longer or prefer familiar toys over novel toys when evidence is ambiguous, confounded, or inconsistent. However, the dependent variables of these studies were a quantitative aspect of exploration: the time children spend on playing with the ambiguous toy. However, playing longer does not necessarily provide information regarding the quality of exploration and learning outcomes. To illustrate, a child might do the same intervention repeatedly and play longer; yet if this is

not an informative intervention, there would not be any learning outcomes at the end. Learning from one's interventions is possible only when the interventions are informative.

The isolation of objects. Several studies investigated young children's patterns of exploration. Cook, Goodman, and Schulz (2011) familiarized 4-year-olds with perceptually different beads activating a light box. Depending on the condition, children were familiarized with different baseline evidence: Children in the Some Beads condition saw that two out of four different beads activated the light suggesting that only some beads were efficacious whereas children in the All Beads condition saw that all four beads activated the light suggesting that all beads were efficacious. Next, children were given two pairs of beads. One of the bead pairs was stuck together; the other bead pair was also stuck but they were separable. Subsequently, children observed that the bead pairs activated the light. In a free exploration phase, children were given chance to play freely with the bead pairs. The results revealed that 65% of the children in the Some Beads condition isolated the separable pair and tested both of the beads separately, whereas only 5% of the children in the All Beads condition separated the bead pairs. In sum, when the baseline information suggested uncertainty, children performed informative interventions by isolating the bead pair which reveals information about the causal efficacies of the individual beads. In a following similar experiment children also invented novel ways to reveal causal efficacies of the individual beads: when the bead pairs were inseparable, children isolated the beads by rotating the pair so that each time only one of the beads touched the surface of the machine. These two studies suggest that children perform a strategy similar to *the isolation of variables strategy* during exploratory play.

Unconfounded experiments. Another study which investigated children's patterns of exploration was conducted by van Schijndel, Visser, van Bers, and Raijmakers (2015). Different from the study by Cook et al. (2011), this study investigated children's exploration patterns in the case of theory-violating evidence in a specific domain in which children already had naïve theories: the domain of shadow size. Shadow size depends on the size of objects and the distance of light source from objects; therefore, correct prediction of shadow size depends on taking into account both of these factors. However, early research showed that young children tend to overlook the distance variable and make predictions based on only the size of the object (Chen, 2009; Ebersbach & Resing, 2007; Siegler, 1981). To investigate how intuitive theories influence children's exploration in the case of consistent and inconsistent evidence, van Schijndel et al. only included the children who had the intuitive theory that only the size of the objects is critical for the shadow size[8]. Researchers used a so-called "shadow machine" which enables putting puppets side-by-side; then, allows children to observe and compare the shadows of the puppets reflected on a screen. Two variables were used: puppet size with two levels (small and big) and distance of the puppets from the light source with three levels (close, middle, and further away). Before a free play session, children in the Inconsistent Evidence condition observed theory-violating evidence where one big and one small puppet were placed in different distances and the small puppet had a larger shadow than the bigger puppet. Children in the Consistent Evidence condition observed evidence consistent with their intuitive theory: the small puppet and the big puppet were placed at the same distance, and the big puppet had a bigger shadow.

[8] Children were administered a pretest about their prior theories on shadow size. Researchers used rule assessment methodology (Siegler, 1976; 1981) and latent class analysis (McCutcheon, 1987; Rindskopf, 1987) in order to assess children's naïve theories. Thirty-nine out of 102 children who mainly performed correctly in the size items but did wrongly in the distance items were included in the further analyses.

Van Schijndel et al. both investigated children's exploration patterns during free play and measured their learning outcomes via a post-test that was administered after the free play. During the free play, experimenters coded the number of times children conducted *experiments*. An experiment in this study was defined as placing one or more puppets and turning the light source on. In this context, one can do unconfounded, confounded, irrelevant, or equal experiments. The primary focus of the study was to see whether children would do unconfounded experiments: keeping one variable constant and varying the other variable. One can do either unconfounded size experiments or unconfounded distance experiments. The results revealed that children who observed theory-violating evidence performed more unconfounded size experiments than the children who observed theory-confirming evidence. On the other hand, there was no difference between the numbers of unconfounded distance experiments across the two conditions. On closer inspection, the results shows that 20% of the children in the confirming and 73% of the children in the conflicting condition performed at least one unconfounded size experiment, and approximately 40% of the children in both conditions performed at least one distance experiment.

These results are interesting, firstly, because one hypothesis would be that children who observed conflicting evidence would look for an alternative hypothesis to explain the inconsistent evidence. In this respect, distance from the light source is a potential variable to consider. With such a hypothesis, the expected experimental design would be performing distance experiments. However, this was not the case. Indeed, children who observed theory-violating evidence performed more size experiments which suggest that children were trying to confirm their naïve theories when they came across inconsistent evidence. Secondly, doing confirmatory experiments, to some extent, might explain children's low learning outcomes assessed at the posttest. In the total

sample, only 10% of the children revised their naïve theory (size matters) to a more advanced theory (both size and distance matter), and this did not differ across the two conditions. A close look at children's experiments showed that the children who revised their theory to a more advanced one performed at least one unconfounded distance experiment. On the other hand, none of the children who did not perform distance experiments revised their theory. Put together, these results reveal important insights about young children's exploration patterns. Even though the learning outcomes were not high, this study presents evidence on young children's spontaneous use of CVS during exploration. Taken into account that this study employed an exploration paradigm and there were no specific prompts provided for experimentation, the natural appearance of CVS in children's exploratory play is intriguing. Children`s experiments are not only focused on the variables critical in their naïve theories but they make experiments considering a variable which they did not consider causally efficacious before.

Causal explanation and exploration. In the exploration studies mentioned so far, children tended to interact with toys when the causal relations were either ambiguous or contrary to their naïve theories. When the causal relations were ambiguous, children lacked information about the causal relations of the toy. When the evidence was contrary to their theories, there was, again, some information missing because the observed evidence was not coherent with the existing theories. However, one open question is what kind of cognitive mechanisms yield this exploratory play behavior. Legare, Gelman, and Wellman (2010) investigated children's explanations in the case of inconsistent and consistent evidence. In the learning phase, the objects that activated the light box were given a novel label "toma"; and the objects that did not activate the light box were labeled "not-a-toma." Children observed that two

perceptually identical tomas (e.g., red squares) and two perceptually identical not-tomas (e.g., blue hexagons). In two within-subjects experimental conditions, children were presented with consistent and inconsistent evidence with the object effects they observed in the learning phase. The results of the study revealed that children were more likely to give causal explanations for the inconsistent evidence than consistent evidence. It seems that children not only tend to explore more in the case of inconsistent evidence as demonstrated by the former studies but also have a tendency to give causal explanations after observing inconsistent evidence.

In the following study, Legare (2012) investigated the relationship between children's causal explanations and their exploratory behavior in the case of consistent and inconsistent evidence. Legare (2012) hypothesized a relationship between children's explanation and exploration patterns in the case of inconsistent evidence. The first part of the study was identical to Legare et al. (2010). Additionally in this study, after children observed consistent or inconsistent evidence and provided causal explanations, they were given a chance to play freely with the toys they observed consistent or inconsistent evidence for. Results revealed that children who provided causal explanations (e.g., it is broken) for the inconsistent evidence displayed longer and more variable play behavior than children who did not provide causal explanations for inconsistent evidence. The dependent measures were the duration of play, stacking objects, opening objects, and evidence evaluation statements during play. According to these results, an underlying mechanism for selective exploration in the case of ambiguous or inconsistent evidence might be due to children's motivation for causally explaining evidence. Children may tend to explain inconsistent evidence which triggers further exploration to gain more information regarding the causal relations and enables explaining the reasons for the inconsistency. Further studies have also shown that

children revise their explanations when they are presented with novel patterns of evidence (Legare, Schult, Impola, & Souza, 2016), and prompted to explain facilitates learning causal relations (Legare & Lombrozo, 2014; Walker, Lombrozo, Legare, & Gopnik, 2013). These studies have shown that explanation have positive influences on children's causal learning.

Legare (2014) argues that the tendency for explaining and exploring is similar to hypothesis generation and hypothesis testing processes. In science, these two epistemic activities work in tandem. Scientists generate hypotheses and conduct experiments to test these hypotheses to gain information. The findings of the experiments pave the way for new hypotheses and new experiments. In the study by Legare (2012), children saw evidence that an object that was expected to activate a machine did not activate the machine. Some of the children provided causal explanations for the inconsistent evidence, such as the object is broken, and those children were more likely to open the object, sort that object with other objects etc. The tendency to causally explain the inconsistent evidence is similar to hypothesis generation process. Stating a causal explanation triggers the tendency to learn whether the causal explanation is indeed correct which results in further exploratory behavior.

Taken together, exploration studies demonstrated that children do not only passively evaluate evidence, but they perform active interventions to reveal causal information. When there is causal information to be gained, they prefer to play with causally ambiguous toys (Schulz & Bonawitz, 2007), and they play longer (Bonawitz et al., 2012). Naïve theories have a significant influence on their exploration. Theory-violating evidence triggers exploration (Bonawitz et al. 2012; van Schijndel et al., 2015). Importantly, their interventions are informative. Epistemic strategies such as the isolation of variables (Cook et al., 2011) or CVS (van Schijndel et al., 2015) naturally

appear in children's exploratory play. All these findings demonstrate that young children have at least an implicit sensitivity to ambiguity, and they can make informative interventions. From the perspective of scientific reasoning, it remains an open question whether children generate and test any particular hypotheses during their exploration. In this respect, Legare's (2012) findings are informative because they demonstrate that children form explicit hypotheses in response to inconsistent evidence, which later informed their exploration behavior. With respect to learning from interventions, van Schijndel et al. (2015) suggested that young children's natural appearing experimentation is similar to scientific reasoning studies (e.g., Perner & Klahr, 1996), as shown by their tendency to perseverate on their own intuitive theories and making experiments in order to confirm their own theory rather than testing it.

2.3.2.4 Conclusion

This section on early causal learning skills aimed to provide an overview of the studies on young children's early "evidence evaluation" and "experimentation" abilities. Taken together, young children have, at least, an implicit awareness of evidence as an epistemic category. In this regard, these findings bear great importance for informing research and theories on the development of scientific reasoning. However, when and how the abilities to make explicit judgments based on evidence and to test hypotheses with an explicit understanding of the alternative hypotheses develops; and how the development of mature forms of scientific reasoning is related to early causal learning competences in the preschool years remain unexplored.

2.4 Metacognition in Preschool Age

Based on the evidence on the early childhood years, researchers consider young children as little scientists who revise theories in accordance with evidence and perform

interventions on the world to reveal causal relations (Gopnik, 2012). However, this kind of learning is often implicit, in that children may use evidence to build and revise hypotheses and theories without representing their own or others' theories as mental constructs. Yet, one important component of mature scientific reasoning, which differentiates it from early forms of theory formation and revision, is the ability to reflect on one's own thinking processes, in other words, a metacognitive understanding of the knowledge acquisition process (Kuhn, 1988; Kuhn, Iardonau, Pease, & Wirkala, 2008; Sodian & Frith, 2008). In this respect, existing literature on the development of metacognitive abilities are informative concerning the goals of this thesis. This section will briefly describe metacognition and provide information about empirical studies on the development of metacognitive abilities in the preschool years.

Although the importance of reflective thinking and having deliberate control over one's own cognitive processes were mentioned by earlier theorists of cognitive development (e.g., Vygotsky, 1962), the term metacognition was firstly coined by Flavell (1979) and described as:

> '. . . any knowledge or cognitive activity that takes as its cognitive object, or that regulates, any aspect of any cognitive activity . . . [T]his conceptualization refers to people's knowledge of their own information-processing skills, as well as knowledge about the nature of cognitive tasks, and about strategies for coping with such tasks. Moreover, it also includes executive skills related to monitoring and self-regulation of one's own cognitive activities" (Schneider, 2008, p. 114)

Schneider (2008) presented a taxonomy of metacognition components: the two main components in this taxonomy are knowledge about the mental world (Kuhn, 2000) and knowledge about memory (Flavell & Wellman, 1977). Knowledge about the mental

world consists of understanding false belief, desires, and knowledge. Knowledge about memory has a declarative and a procedural component, and the procedural component has components of monitoring, control, and self-regulation.

A significant number of the studies in metacognition research have focused on *metamemory* component of metacognition. Early research on metamemory showed that these skills show a slow progression of development from early to late childhood (Schneider & Pressley, 1997). Further studies showed that 4- to 6-year-old preschoolers spontaneously employ memory strategies; there is development in preschoolers' strategy use during the preschool years (Schneider & Sodian, 1988; Sodian, Schneider, & Perlmutter, 1986), and preschoolers' *feeling of knowing* judgments predicts their recall performance (Cultice, Sommerville, & Wellman, 1983). More recent studies investigating implicit metacognition in young children using physiological measures (e.g., pupil dilation) suggest that even 3.5-year-old children have an implicit understanding of their memory (Paulus, Proust, & Sodian, 2013).

The other component of metacognition, knowledge about the mental world, being able to reflect on one's own knowledge and ignorance is especially critical for scientific reasoning and it is the focus of the present thesis. There is empirical evidence that infants and toddlers have some form of understanding of knowledge and ignorance of others. Toddlers produce more communication acts when their mother is ignorant than knowledgeable (O`Neill, 1996). Infants and toddlers selectively share information when the person whom they are interacting with is ignorant rather than knowledgeable (e.g., Harris, Ronfard, & Bartz, 2017; Liszkowski, Carpenter, & Tomasello, 2008). Around 3 years, children start to use mental state words such as *think* and *know* in their daily speech. Bartz, Rowe, and Harris (2016 as cited Harris et al., 2017) investigated 16- to 37-month-old toddlers' expressions of ignorance. Children were shown either familiar

or unfamiliar item drawings and asked to name those items. In the unfamiliar drawings, children expressed ignorance either in implicit ways such as looking at the mother, saying filling words such as “um”, and some children also provided explicit reflections of ignorance. These findings show that the sensitivity to other’s knowledge states is already present early in development.

Kim, Paulus, Sodian, and Proust (2016) investigated preschoolers’ implicit understanding of their ignorance by looking children’s tendencies to share knowledge with others when they were knowledgeable or ignorant. Although 3-year-olds were worse at explicitly reporting their ignorance, they were less likely to share information with the other when they were ignorant than when they were knowledgeable. Lyons and Ghetti (2013) employed a forced-choice task when success was dependent on some form knowledge, and they found that 3-to 5-year-old children were less likely to make a choice when they were uncertain about their knowledge compared to when they were confident. Taken together, these findings suggest that even 3-year-olds have an implicit understanding of their epistemic states.

Children’s explicit judgments about their epistemic states may provide insights about the development of explicit metacognitive skills. Rohwer, Kloo, and Perner (2012) investigated 3- to 7-year-old children’s knowledge judgments for their knowledge states (ignorant vs. knowledgeable) when children were entirely and partially ignorant. In the Total Ignorance condition, children were asked whether they knew what was inside of a closed box that they had never seen before. In the Partial Ignorance condition, children were shown several objects, and the experimenter put one of the objects in the box without showing the child which object it was. Children`s explicit judgments about their knowledge or ignorance were investigated. Even 3-year-olds correctly reported their epistemic states correctly in the case of complete knowledge or

total ignorance. The results revealed that only the children who were older than 5 years reported ignorance in the partial knowledge task. These results suggest a development around 5 years in evaluating one`s own epistemic states in the case of uncertainty.

Scientific reasoning requires the ability to explicitly reflect on the hypothesis–evidence relation and the ability to control and coordinate hypotheses and evidence (Kuhn, 1988). In Kuhn's sense, this is similar to conscious awareness. Although a few studies on preschoolers' explicit metacognition may give insights about the development of reflective abilities on their own epistemic states, little is known about preschoolers' ability to reflect on the hypotheses–evidence relation.

2.5 The Child-as-Scientist View

Piaget was the first cognitive developmentalist who proposed the metaphor of little scientists (Piaget, 1952). This was a bold argument in his time because scientific thought was considered the specialization of philosophy of science which mainly focused on the logical validity and the justification of scientific theories (Reichenbach, 1989 as cited in Kuhn, 1992), and the psychology of science—how scientific knowledge and theories are product of human psychology—was not yet an area of investigation in itself. The "child as scientist" metaphor, in his sense, was arising from children's curiosity and intrinsic motivation for exploration. Later in his writings, Piaget also emphasized the continuity in development from children to scientists—how early childhood development later give rise to formal thinking and mature scientific thinking as an endpoint (Piaget & Garcia, 1989). The implication of the metaphor, however, was very limited since Piaget's theory depicted children as irrational thinkers who are unable to form causal relations.

More recent accounts of cognitive development have gone further in characterizing the similarity between development in the early years of life and scientific thinking. The theory theory account suggested that conceptual changes in development take place as theory formation in science and that infants and children have representations and rules which are structurally and functionally similar to theories of science (see Carey, 1985; Gopnik & Wellman, 1992 for alternative views). One argument was that the structure and the characteristics of scientific theories might guide cognitive developmentalists in revealing the nature of cognitive development (e.g., Carey, 1985; Gopnik, 1996). Different from other theories of cognitive development (e.g., information processing), this view has been concerned with children's concepts about phenomena and how these concepts evolve and change into different concepts. The investigation of the development of physical concepts for material kind, for instance, showed that children's conceptions for weight, surface, or density show a similar process of knowledge reconstruction as in science (Smith, Carey, & Wiser, 1985)[9]. Similar conceptual changes have also been shown taking place in other areas of cognitive development; for instance, the concepts for the early understanding of biology (Carey, 1985; Hatano & Inagaki, 1994), and psychology (Gopnik, Meltzoff, & Bryant, 1997; Gopnik & Wellman, 1994). From this point of view, children are little scientists because they represent knowledge structurally and functionally similar to scientific theories; and development is the consequence of changes in the theories, similar to processes of theory revision in science.

Recent empirical work on causal learning, presented earlier, has demonstrated that the similarities between young children and scientists go beyond the similarity in

[9] Children first have a concept of felt weight concept without having a concept of other aspects such as density. Later they develop an undifferentiated weight/density concept, which is followed by differentiated conceptions for weight and density separately (Smith et al., 1985).

the theory-like representation of knowledge; there are additional similarities in terms of the mechanisms and processes of knowledge acquisition and formation process (for reviews see Gopnik & Wellman, 2012; Schulz, 2012). In a nutshell, empirical results have shown that (a) children make use of sophisticated statistical learning mechanisms in the process of learning, (b) they perform interventions on the world to reveal information which is not readily available to them, (c) and they (rationally) learn from other people around them (Gopnik, 2012). Putting aside the complexity and the difficulty of the scientific content domains and the specific methodologies used in scientific practice, the core epistemic practices that children and scientists follow seem to be very similar at an abstract level. Moreover, the mechanisms between children's explanation and exploration processes were proposed to be similar to hypothesis generation and hypothesis testing in scientific reasoning (Legare, 2014). Some researchers also argue that the psychological processes giving rise to core epistemic practices of knowledge formation processes are not specific to science but "…they are universal abilities at the foundations of human cognition." (Schulz, 2012, p. 382).

Researchers studying the development of scientific reasoning approach the child-as-scientist account cautiously. Their interest lies not only in the metaphor but in whether children can reason scientifically. Children's sensitivity to ambiguous evidence (Schulz & Bonawitz, 2007), their informative interventions, such as isolating variables (Cook et al., 2011) or designing unconfounded tests (van Schijndel et al., 2015), suggest that children have, at least, an implicit understanding for the informativeness of evidence. However, certain abilities have been theoretically considered as foundational for scientific reasoning (Kuhn, 1988; 1989): these are the understanding of the hypothesis–evidence relation and having a metacognitive understanding of this relation. To be able to reason scientifically, firstly, it is necessary to have an understanding of

mental concepts (beliefs, hypotheses, and theories) and evidence as distinct epistemic categories, as well as their relation to each other—evidence is a means to gain knowledge about the truth or falsity of mental concepts. Secondly, it is necessary to have a metacognitive awareness of the processes of knowledge acquisition and formation. So, scientific reasoning requires being able to think about theories, not only with them (Kuhn, 2010). Although causal reasoning studies have presented evidence that children have some implicit understanding of the epistemic goals, and of the relationship between hypothesis and evidence (e.g., Cook et al., 2011; Legare, 2012), little is known about young children's representation of distinct epistemic categories and their metacognitive understanding of the epistemic categories and processes. In this respect, early causal reasoning skills are not sufficient to claim that young children reason scientifically.

Different from studies demonstrating early competences, older children's, adolescents', and adults' performance in scientific reasoning studies would not suggest a developmental pattern which is in favor of the child-as-scientist account. In scientific reasoning studies, people often show poor performance in designing unconfounded experiments and evaluating evidence. However, these tasks are often difficult: they include multivariable problem contexts, present data contrary to intuitive theories or require familiarity with the content domain of the problem (e.g., see Zimmerman, 2007 for a review). These tasks are good at showing people's abilities in complex problems which people often face in real-life or in formal education; however, it is highly likely that the complexity and difficulty of the tasks hinder the detection of early abilities of scientific reasoning. Few studies, investigating scientific reasoning in preschool age with reduced task demands, demonstrate that preschoolers have some basic abilities for

hypothesis testing (Piekny et al., 2014; Piekny & Maehler, 2013; van der Graaf et al., 2015) and evidence evaluation (Koerber et al., 2005; Ruffman et al., 1993).

All in all, evidence has shown that the resemblance between children and scientists goes beyond the similarity of the exploratory processes as Piaget first proposed. Young children have powerful cognitive processes for learning (e.g., sensitivity to ambiguity, informative interventions) which are similar to core practices of science (e.g., evidence evaluation, experimentation). Although these findings suggest that the child-as-scientist is an appropriate metaphor in itself, it is unclear whether it is more than a metaphor—whether children can reason scientifically. Theoretically, scientific reasoning requires having epistemic concepts of hypotheses and evidence, as well as metacognitive awareness of epistemic processes. However, little is known about these abilities in the preschool years.

2.6 The Aim of the Thesis

This thesis aimed to investigate the abilities for hypothesis testing, evidence evaluation, and argumentation from evidence in 4- to 6-year-old preschoolers by focusing on their metacognitive understanding of their own epistemic states and processes, that is, their abilities for representing epistemic goals and understanding the hypothesis–evidence relation. Based on early studies of scientific reasoning, it has been claimed that scientific reasoning is a late developing skill, and preadolescent children lack the foundational skills to reason scientifically (e.g., Kuhn et al., 1988; Kuhn & Franklin, 2006). Later studies with reduced demands—although they are few—documented that preschoolers have basic abilities of hypothesis testing and evidence evaluation (Koerber et al., 2005; Ruffman et al., 1991; Sodian et al., 1991). Recent findings on causal learning have documented that preschoolers possess powerful learning abilities from evidence, and these abilities resemble epistemic practices of science. Moreover, young children show a metalevel understanding of their epistemic states (Rohwer et al., 2012) and have an understanding uncertainty (e.g., Sobel et al., 2017). In the light of recent findings on young children's abilities, there is a gap in the literature with respect to the investigations on the development of scientific reasoning abilities in the preschool years.

Recent research on causal learning has demonstrated that young children have powerful learning mechanisms for learning from evidence which resemble epistemic practices in science. Young children form cause–effect relations based on evidence, they make causal predictions, and they even make interventions on their environment to reveal causal information which is not readily available to them (See Gopnik & Wellman, 2012 for a review). These findings in early childhood have been interpreted to

support the view "that very young children's learning and thinking skills are strikingly similar to much learning and thinking in science: Children test hypotheses against data and make causal inferences, they learn from statistics and informal experimentation …" (Gopnik, 2012, p. 1623). However, despite many similarities between the ways young children and scientists acquire new knowledge; it is unclear whether and to what extent young children are similar to scientists in terms of intentionally guided knowledge seeking processes that characterize scientific reasoning (Kuhn, 2010). In this respect, it is an open question whether young children can differentiate and coordinate hypotheses and evidence, and whether they have a metacognitive awareness of this relation.

The theory of mind research has demonstrated that around 4 years of age a critical development takes place which has been considered as a prerequisite to being able to reason scientifically. Before the development of such an understanding, children's explicit responses suggest that they do not, at least explicitly, represent that people might hold beliefs different from reality. However, around 4 years of age, children show the understanding that people may hold different beliefs about the same phenomenon or people may hold false beliefs. This is evidence for the ability to differentiate beliefs from evidence and understand the epistemic relation between evidence and mental concepts. However, in the typical theory of mind tasks, children themselves are presented with "real" evidence, and a protagonist is presented with fake evidence. In such tasks, it is sufficient if children represent what they know and what the protagonist knows without any acknowledgment of the uncertainty regarding the veracity of the statements. In this respect, although the theory of mind literature

provides evidence on preschoolers' coordination of beliefs and evidence, it is unclear whether young children can represent hypotheses[10].

Understanding uncertainty is crucial to be able to reason scientifically. The fundamental requirement for scientific reasoning is to be able to represent alternative hypotheses. In the case of information gain, a knowledge seeker should be able to acknowledge that the truth value of a statement (hypothesis) is uncertain. Different from the early research suggesting that an understanding of uncertainty develops around 8 years of age (Bindra et al., 1980), recent findings have demonstrated that preschoolers have a developing understanding of causal uncertainty earlier than 8 years of age. Even 4- and 5-year-olds represent uncertainty in the case of simpler tasks (Erb & Sobel, 2014) and around 6 and 7 years children successfully pass causal uncertainty tasks with increased difficulty (Sobel et al., 2017). However, in these tasks children are required to make predictions about the efficacy of objects, and this does not necessarily require a metacognitive awareness of uncertainty. In this respect, in order to understand whether young children have foundational abilities to reason scientifically, it is necessary to find out whether young children have a metacognitive awareness of their knowledge in the case of uncertainty.

Our knowledge on preschoolers' early metacognitive awareness of their epistemic states had been very limited. A recent study by Rohwer et al. (2012) is

[10] Some studies on scientific reasoning also used paradigms similar to theory of mind tasks in the sense that children were required to represent true and fake evidence. For instance, in the studies by Ruffman et al.(1993) and Koerber et al. (2005) children have to evaluate their own beliefs and beliefs of another person based on real or fake covariation evidence. These findings demonstrated preschoolers' abilities to evaluate covariation evidence, however, the paradigm does not require children to represent hypotheses as possible alternative states with uncertainty.

informative with respect to preschoolers' metacognitive awareness of their epistemic states in the case of uncertainty. The study demonstrated that children around 5.5 years but not the younger ones show a metacognitive awareness of their ignorance in the case of partial information. This study demonstrates that young children can reflect on their epistemic states and acknowledge their ignorance in the case of partial information. This shows that they are able to reflect on the epistemic states as a function of evidence. However, this task does not provide information for children's understanding of forming causal relations as a function of evidence. Investigations on the foundational skills for scientific reasoning should examine whether young children show such reflective awareness of their epistemic states in forming causal relations from evidence.

(a) Preschoolers' sensitivity to the informativeness of evidence shown by causal learning studies, (b) their abilities to reason in the case of causal uncertainty, and (c) their awareness of their ignorance in the case of partial information suggest that preschoolers already possess cognitive skills that are akin to, and possibly precursors of, the foundational abilities of scientific reasoning. In this respect, young children's scientific reasoning abilities might have been underestimated, and warrant further empirical investigation. By being informed by the recent findings on cognitive development, the present thesis aimed to shed light on the early development of foundational abilities for scientific reasoning.

Different from typical scientific reasoning tasks, the empirical studies in this thesis employed the blicket detector paradigm in which children can directly observe cause–effect relations, interact with objects, and observe the effects of their own interventions. Scientific reasoning studies on preschoolers often ask children to evaluate evidence which has been presented via pictures or use tasks which require making inferences from indirect evidence. Although one would require these abilities for mature

forms of scientific reasoning, such requirements might lead to the underestimation of early foundational scientific reasoning abilities. Numerous studies on causal learning have shown that the blicket detector paradigm is convenient and useful for experimental studies with young children. Firstly, the paradigm does not require or contradict with domain-specific content knowledge which has frequently been shown to have a significant effect on children's performance (e.g., Croker & Buchanan, 2011; Koerber et al., 2005). Secondly, children can directly observe cause–effect relations and interact with the objects themselves. Considering that young children learn better from their own interventions and discoveries (Sobel & Sommerville, 2010), the blicket detector paradigm seems to be a better-suited paradigm for investigating preschoolers' early abilities for scientific reasoning.

This thesis investigated the development of scientific reasoning abilities in three epistemic activities namely hypothesis testing, evidence evaluation, and argumentation. We directed our focus towards these three activities firstly because the empirical investigations on the development of scientific reasoning in older children mostly focused on these epistemic activities (See Zimmerman, 2000; 2007); therefore they are informative for both shaping the research questions of the present studies and relating the present findings to the earlier findings in the area. Secondly, children's early skills for forming causal relations from covariation evidence (e.g., Gopnik et al., 2001) or making interventions in the case of ambiguous evidence (e.g., Legare, 2012; van Schijndel et al., 2015) very much resemble the epistemic activities of hypothesis testing and evidence evaluation. To our knowledge, the present thesis is one of the first to bring together the early causal learning and the research on the development of scientific reasoning. In this respect, it is best to focus our studies on epistemic activities those of

which causal learning research provided accumulated evidence on early implicit abilities.

It is important to note that this thesis focused on the elemental, domain-general skills for scientific reasoning and does not claim that the presence of such skills would enable children to solve more complex problems. It is certain that preschoolers would not be able to reason scientifically in the domain of quantum physics or evolutionary biology. We acknowledge that science is difficult; and by studying basic abilities in young children we neither argue that preschoolers can perform successfully in more difficult problems, nor ignore how challenging it is to develop mature scientific reasoning skills. It remains an open question, in this respect, how mature forms of scientific reasoning unfold from less mature forms and which mechanisms and processes at different levels (cognitive, social, and educational) give rise to change. Although we do not investigate this question directly, we believe our findings will be informative nonetheless, as a theory for development of scientific reasoning would not be complete without showing its developmental origins (Kuhn & Pearsall, 2000).

Altogether the present thesis examined the development of scientific reasoning in 4- to 6-year-old children by investigating their abilities for core epistemic practices: hypothesis testing, evidence evaluation, and argumentation from evidence. Together with children's performance in these epistemic practices, we investigated their abilities for representing epistemic goals, having differentiated concepts for hypothesis (or belief) and metacognitive abilities for their epistemic processes. In the following chapters, we will present three empirical studies on the early development of these abilities.

3 Study 1: Young Children Selectively Make Interventions in response to Epistemic and Practical Goals

Scientific reasoning can be described as intentional knowledge seeking (Kuhn, 2010). Kuhn explained this as "…any instance of purposeful thinking that has the objective of enhancing the seeker's knowledge" (2010, p. 498). The fundamental nature of hypothesis testing is characterized by being directed towards epistemic goals: using evidence to gain knowledge about the truth or falsity of a certain state of the world. Having a proper representation of epistemic goals, and being able to differentiate them from practical goals is indispensable for hypothesis testing and for scientific reasoning in general. Without an understanding of epistemic goals, it would not be possible to understand how evidence bears on the truth or falsity of hypotheses. In this respect, it is critical to understand how cognitive capacities for understanding epistemic goals develop in the early childhood years in order to understand the development of hypothesis testing abilities.

There are several scientific reasoning studies showing that children show poor performance in hypothesis testing tasks because they confuse the epistemic goals of hypothesis testing with the practical goal of producing or demonstrating an effect. To illustrate, in a study by Penner and Klahr (1996), 10-, 12-, and 14-year-olds were presented with a so-called "sinking paradigm" and were asked to find out which variables (e.g., weight, material type) effect the sinking rate of a given object. Think aloud statements during experimentation suggested that participants, especially the younger ones, had a tendency for demonstrating a predicted effect, such as showing that heavy objects would sink faster, rather than performing interventions in order to gain knowledge. Several other studies on hypothesis testing similarly show that participants

misrepresent the goals of hypothesis testing and aim to produce a positive effect or demonstrate what they think is true instead (e.g., Croker & Buchanan, 2011; Dunbar, 1993; Dunbar & Klahr, 1989; Tschirgi, 1980).

Although older children and adolescents frequently have problems with differentiating the goals of hypothesis testing from effect production, these studies often employed contexts in which children already had strong prior knowledge and expectations (e.g., sinking paradigm in Penner & Klahr, 1996), or they used difficult procedures in testing (e.g., constructing test objects in Schauble, 1990). The poor performance of older children in earlier tasks might be due to the influences of domain-specific knowledge or high task demands. Studies with neutral contexts and reduced demands have shown that even young elementary school children have the basic abilities for understanding what it means to test a hypothesis. Sodian et al. (1991) demonstrated that 7- and 8-year-olds are able to make conclusive tests to test a given hypothesis in a simple task. Children chose conclusive tests only when they were asked to test a hypothesis but not when they were asked to produce an effect (i.e., feeding a mouse). This shows that children of this age can purposefully choose a conclusive test in order to test a hypothesis.

There is also evidence for early competencies at preschool ages. For example, van der Graaf et al. (2015) presented evidence that preschoolers were able to design contrastive tests comparing the two levels of a variable. Piekny et al. (2014) and Piekny and Maehler (2013) replicated the mouse house task (Sodian et al., 1991) with preschoolers and demonstrated that 6-year-olds are able to choose a conclusive test over an inconclusive one, whereas 4- and 5-year-olds were not better than chance in the hypothesis testing condition. Put together, there is evidence for early competences of hypothesis testing in early childhood. However, existing studies are far from sufficient

to reach a conclusion regarding preschoolers` understanding of epistemic goals and their hypothesis testing skills.

Studies in the last decade demonstrated that young children, indeed, do "little experiments" suggesting that they have some form of implicit representation of epistemic goals, yet these findings do not answer whether children can purposefully do hypothesis testing. Two common methods were used to reveal children's implicit sensitivity to evidence are investigating (a) children's selective exploration of ambiguous evidence (Bonawitz et al., 2012; Schulz & Bonawitz, 2007) and (b) their informative interventions that reveal information during exploratory play (Cook et al., 2011; Legare, 2012; Van Schijndel et al., 2015, see Section 2.3.2.3 for details). Causal learning studies generally give children exploratory prompts but they do not investigate whether children can generate relevant evidence to *test a hypothesis*. The most relevant causal learning study investigating preschoolers' exploratory play behavior which resembles hypothesis testing is by Legare (2012). In this study, children were asked for their explanations in the case of inconsistent evidence. Some children's explanations very much resemble hypotheses, and children's explanations predicted their exploration patterns. Children who provide causal explanations (e.g., the object is broken, see Section 2.3.2.3 for details) for inconsistent evidence showed more variable play that might provide them with informative evidence. These findings suggest that as young as 3 years of age preschoolers have an implicit understanding of the informativeness of evidence. However, scientific reasoning requires the purposeful coordination of a hypothesis with a relevant piece of evidence with the explicit goal of gaining information. In this respect, these findings do not show whether preschoolers can successfully differentiate epistemic goals from practical ones and whether they can generate evidence to test a claim.

Preschoolers' implicit understanding of the informativeness of evidence brings forth the question whether preschoolers' abilities for scientific reasoning have been underestimated. Although the mouse house task is convenient for investigating young children's hypothesis testing abilities by comparing their selective responses in hypothesis testing and effect production conditions, several aspects of the task are challenging for younger children. Firstly, it is a story-based task which brings forth increased linguistic demands. It has frequently been shown that language skills are related to preschoolers' other cognitive abilities (e.g., Gooch, Thompson, Nash, Snowling, & Hulme, 2016; Jenkins & Astington, 1996; Müller, Jacques, Brocki, & Zelazo, 2009). Secondly, it is highly likely that making inferences from indirect evidence to test a hypothesis is more difficult than making conclusions based on direct evidence. In the mouse house task, children have to infer that if the cheese is gone when there is a small door to the mouse house, this means that mouse is small. Although hypothesis testing generally requires making inferences from the indirect evidence, the investigation of the early abilities might be underestimated in the case of indirect evidence. Therefore, investigations on preschoolers' early abilities would be deficient without examining preschoolers' evaluation of direct evidence.

The present study aimed to investigate the early abilities of preschoolers' hypothesis testing abilities with the blicket detector paradigm—which has frequently been shown as a successful paradigm for investigating younger children's abilities. We employed the two conditions of the mouse house task, namely hypothesis testing and effect production in order to investigate preschoolers' purposeful differentiation of epistemic goals of hypothesis testing from effect production.

3.1 Study 1a

The present study aimed to investigate the development of hypothesis testing in early childhood by focusing on preschoolers' ability to differentiate the epistemic goals of hypothesis testing from the practical goals of effect production. In two conditions (Hypothesis Testing vs. Effect Production), children were given two different goals and their selective interventions to reach these goals were assessed. The blicket detector paradigm (Gopnik & Sobel, 2000) was adapted for the study purposes. First, during the familiarization phase, children were introduced with a light box and familiarized with the evidence that some of the objects (based on their color) deterministically activated the light, whereas some others did not. In the test phase, children were presented with two novel objects. They first observed that one of the objects (e.g., the blue object) turned the light on; however, children were not allowed to test whether the other object (e.g., the red object) would also turn on the light or not. Next, children were given two object options (blue: "effect-known" and red: "effect-unknown") to choose from to reach a specific goal in each one of the two conditions.

In the Hypothesis Testing condition, the goal was to test the hypothesis that blue objects turn on the lights, while red objects do not. In the Effect Production condition, the goal was to make sure that the light is activated. Given children's prior knowledge about the blue and the red objects, the correct object to pick was different across the two conditions. Although children knew that the blue object activated the light, they had not seen the necessary evidence to conclude whether the second part of the hypothesis (i.e., that red objects do not turn the light on) was also correct. Thus, in the Hypothesis Testing condition (epistemic goal), the correct object choice would be the red object. On the other hand, in the Effect Production condition (practical goal), the best move would be to choose the blue object, as it had proven to turn on the lights and there was no

evidence that the red object does so. If children can selectively choose the correct intervention in response to hypothesis testing and effect production goals, this would suggest that they can understand the differential nature of the goals and have a basic understanding of generating evidence in order to test a hypothesis.

As the present study focuses on hypothesis testing, it is critical to know whether children represent the statement we present (i.e., blue objects activate the box, red objects do not) as a hypothesis. In other words, it is important to know whether children represented the truth of that hypothesis statement as uncertain rather than as a statement describing the true state of the world. In order to encourage children's recognition of the uncertainty of the hypothesis statement in the test phase, whenever we introduced novel pieces of evidence, the experimenter explicitly stated the possible alternative effects of the objects by saying "Maybe only the blue objects make the box light up, maybe only the red objects make the box light up, maybe both of them make the box light up. We don't know." Moreover, we formulated the hypothesis statement in such a way that children's correct object choice would support the idea that they have properly represented the statement as a hypothesis, and not as a description of how the things really are. The wording we used for the hypothesis statement was "Only blue objects activate the box, red objects do not." Considering that children would be more motivated to try an object that produces an interesting effect (i.e., makes the box light up) over an object that supposedly does not produce any effect, choosing the red object over the blue object is more likely to be a result of true representation of this statement as a hypothesis rather than a description of the reality. In these respects, the present study was designed in a way (a) to encourage children's representation of the uncertainty of the hypothesis statement and (b) to indirectly test for their understanding of the statement as a hypothesis.

We also tested lower-level explanations that might motivate the children's object choice. The dependent variable was a binomial choice, and it was possible that children make their choices not with the intention to test the hypothesis or to make the box work again, but with lower-level cognitive motivations such novelty or familiarity preferences. Since children saw the effect of the effect-known object, it was a familiar object, and the effect-unknown object was a novel object. If there was a novelty preference, we would expect children in both conditions to choose the novel object more. If there was a familiarity preference, we would expect children in both conditions to choose the familiar object more. For this reason, we firstly tested the novelty-familiarity hypothesis independent of the condition.

3.1.1 Research Questions of Study 1a

1. Do children selectively choose objects in line with the epistemic goal of the Hypothesis Testing condition and practical goal of the Effect Production condition?
2. Is there a developmental change from 4 to 6 years of age in terms of selectively choosing the correct intervention in the Hypothesis Testing and the Effect Production conditions?

3.1.2 Method

3.1.2.1 Participants

A total of 117 children participated in the study. Five children were excluded due to experimenter error, and one child was excluded because she was younger than the planned age range. The final sample included 111 children (51 females, M_{age} = 66 months; range: 49 months–81 months). Participants were randomly assigned to the two experimental conditions (Hypothesis Testing vs. Effect Production). There were no

significant differences between the distributions of females and males assigned to each group. All children were typically-developing children of lower- to upper-middle class backgrounds from a large German city. Parents signed a written consent form for their children's participation and children were asked for their verbal assent before the study commenced.

3.1.2.2 Materials

A custom-built, 30 cm x 20 cm x 14 cm wooden box with a LED light strip attached around it was used as the "light box". Children were told that certain objects activated the light box when they were put on it while some other objects did not. In reality, a confederate sat behind children and controlled the light box via a remote control. The objects were paper boxes which had a wired interior (similar to a socket), in which wire-wrapped cubes could be inserted. The paper boxes varied in shape (hexagon, square, round) and color (blue, red, yellow, green); and their color or shape was not deterministically related to the light effect that they allegedly produced. The wire-wrapped cubes were 3 cm x 3 cm x 3 cm and came in 12 different colors (See Figure 3.1 for a depiction of the study materials). The critical variable was the color of these cubes; the effects of the cubes of the same color were always the same. Children were told that they could activate the light box by first inserting the cubes into the socket in the paper box, and then putting the paper box on the light box to activate the light.

In both the learning phase and the test phase of the study, two cubes in different colors were introduced in pairs. In total, each child was presented with 13 different cubes in 6 different colors (three color-pairs) in 10 paper boxes. In the familiarization phase, children saw evidence that the same paper box activated and did not activate the light box depending on the cube inserted inside. Photos of four children (2 males, 2 females) were used in the experimental phase in a counterbalanced order.

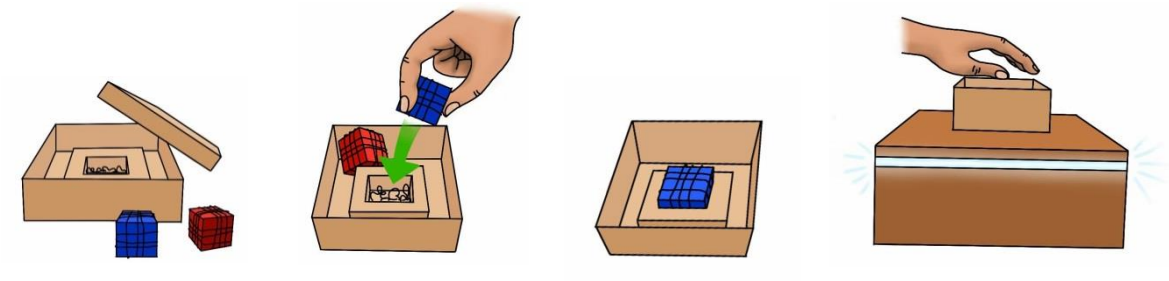

Figure 3.1. Depiction of materials of Study 1a

3.1.2.3 Design

In a between-subjects design, children's evidence generation performance was studied in two conditions, namely, hypothesis testing and effect production. While the evidence that children observed was identical in both conditions, the instruction given in the test phase was different. In the hypothesis testing phase, children were instructed to test a specific hypothesis, which requires choosing the effect-unknown object because it provides the necessary information to test the hypothesis. In the effect production condition, children were instructed to ensure that they activate the light box, which requires choosing the effect-known object because children had already seen that the objects of that color activate the light.

3.1.2.4 Procedure

All sessions were carried out in quiet rooms of kindergartens and recorded by a video camera. Each child was tested individually in a session lasting approximately 10 minutes. The child was seated across from the experimenter. A confederate, who sat behind the children outside of their view, controlled the light box. At the beginning, the experimenter and the child played a warm-up game together (a puzzle, matching the

animals with their habitats). In all phases of the study, children interacted with the objects themselves. The study included a learning phase and an experimental phase.

Learning phase. In this phase, children saw evidence that cubes of the same color have the same effect and it might be that only one of the colors in a pair activates the light box or both of the colors in a pair activate the box (Figure 3.2. Schematic display of the effects of the cubes in the learning phase. The yellow light bulb represents that the cube next to it activates the box, the white bulb represents that the cube next to it does not activate the box. Cubes were presented in pairs but tested on the box individually. The arrow represents time. The yellow-green cube pair represents the one-efficacious-cube pair, and the black-pink cube pair represents the two-efficacious cube-pair. The order whether children first saw the two-efficacious cube-pair and one-efficacious cube-pair was counterbalanced.Two sets of cube-pairs were used in a counterbalanced order. In each set, there were four objects in total (two instances of each color). In one set, only one of the colors activated the light box (one-efficacious cube-pair); in the other set, both of the colors activated the light box (two-efficacious cube-pair); the sets and presentation order were counterbalanced across sessions. First, children were presented with a pair of cubes in a paper box and explicitly told that "Maybe only the green cube makes the light box work, maybe only the yellow cube makes the light box work, maybe both of them make the light box work. We don't know." Children were prompted to insert each cube in the paper box one at a time and to put the paper box on the light box. The experimenter always verbalized the effect ("It made the light box work," "It did not make the light box work"). Subsequently, two novel instances of the cubes of the same color were provided in two different paper boxes. Children put each of the paper boxes on the light box one at a time and observed their effects. An identical procedure was followed for the two more cube-pairs except

the number of efficacious cubes (two-efficacious cube-pair or one-efficacious cube-pair) was different.

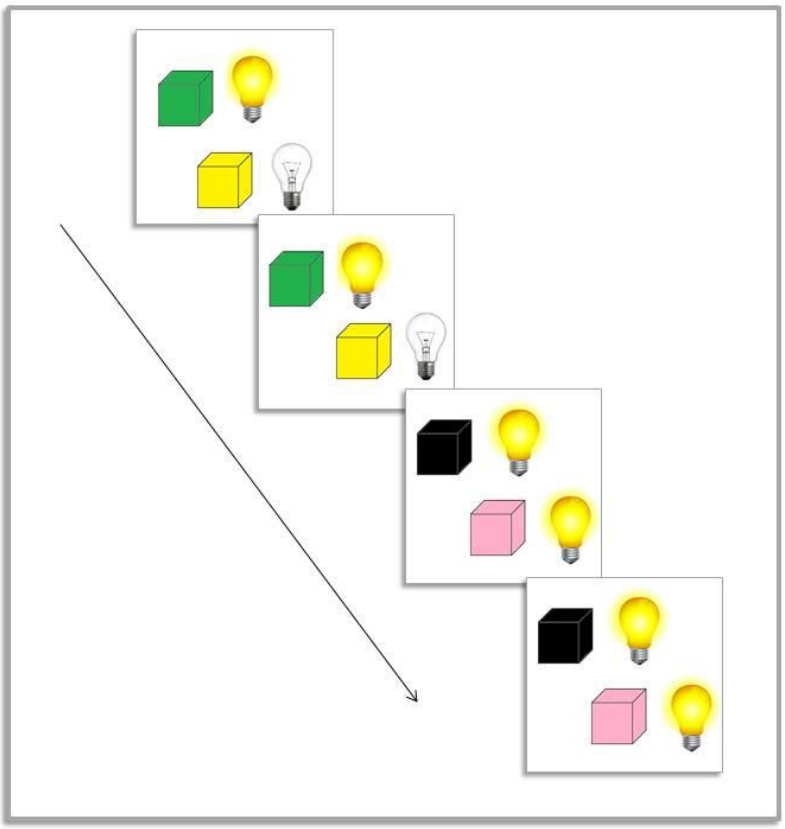

Figure 3.2. Schematic display of the effects of the cubes in the learning phase. The yellow light bulb represents that the cube next to it activates the box, the white bulb represents that the cube next to it does not activate the box. Cubes were presented in pairs but tested on the box individually. The arrow represents time. The yellow-green cube pair represents the one-efficacious-cube pair, and the black-pink cube pair represents the two-efficacious cube-pair. The order whether children first saw the two-efficacious cube-pair and one-efficacious cube-pair was counterbalanced.

Experimental phase. Figure 3.3 depicts the main steps of the procedure of the experimental phase. In this phase, children were not allowed to test all of the objects. In order to rationalize why they were not allowed to try all the objects, this part was presented as a special game with certain rules since children of this age have the understanding that games have certain rules and that they should be played accordingly (Rakoczy, Warneken, & Tomasello, 2008). A cube-pair of two novel colors was

introduced but children were not allowed to try them. As in the learning phase, the experimenter explicitly said it might be that both of the colors activate the light box or only one of them activates it. The experimenter showed a protagonist's photo and said, "This is Laura. She doesn't really know. She thinks only the blue cubes make the box work and the red cubes don't make the box work." Subsequently, children saw evidence that a novel instance of the cube with the hypothesized positive effect (e.g., blue) actually activated the light box. In order to encourage generalization of the effects of the same color cubes, the experimenter said, "It makes the box light up. So, the x (e.g., blue) cubes make the box light up." Next, children were given instructions depending on the condition that they were in. In the hypothesis testing condition, children were told that the aim of the game was to find out whether the protagonist was right or wrong whereas in the effect production condition, they were told that the aim was to make sure that the light box is activated. The experimenter brought two novel paper boxes with their lids closed and explained that there was a cube in each one of the boxes and that they could choose only one of them and try. In order to encourage children to actively think about the aim of the condition before they made their choice, a memory question was asked: In the hypothesis testing condition, the children were asked what Laura thinks and in the effect production condition what the aim of the game is. Independent of children's response, the experimenter repeated the correct answer one more time and simultaneously opened the lids of the paper boxes and revealed the object options.

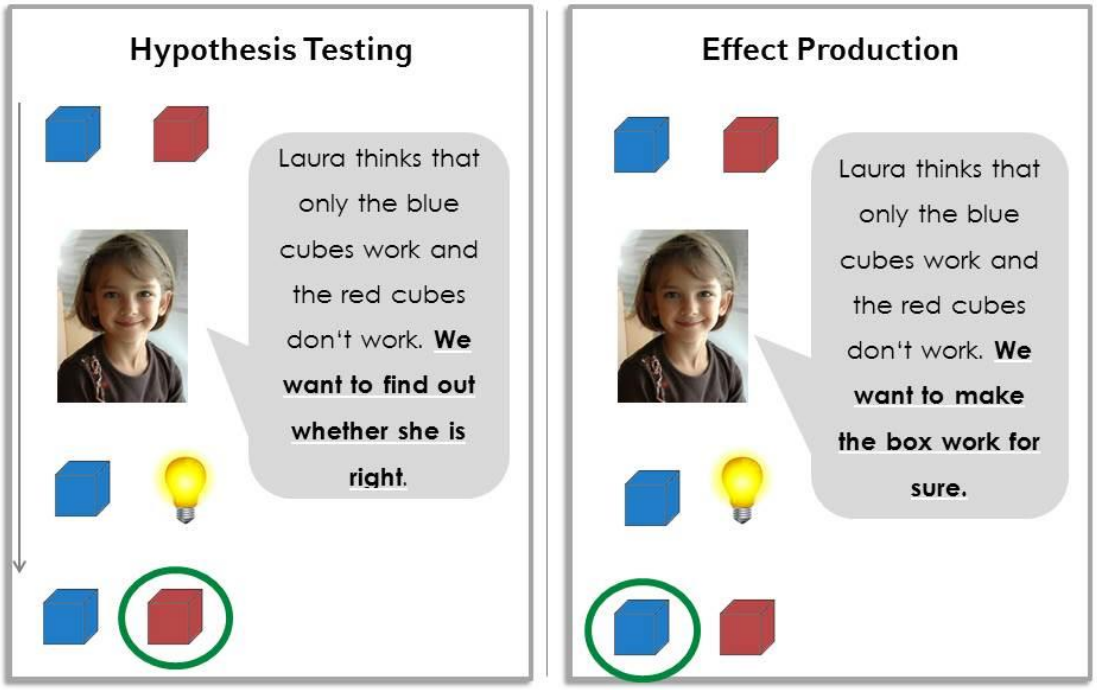

Figure 3.3. Schematic display of the experimental procedure of Study 1a. Hypothesis testing condition on the left, effect production condition on the right. The green circle represents the correct response in each condition.

3.1.2.5 Coding

Children's "choice of object" (effect-known vs. effect-unknown) in the experimental phase was coded. Choosing the effect-unknown object was taken as the correct choice in the hypothesis testing condition whereas choosing the effect-known object was taken as the correct choice in the effect production condition.

Children's answers to memory questions were coded (correct vs. incorrect) for exploratory analyses. In the hypothesis testing condition, children's answers to the memory question (i.e., We want to find out whether Laura is right. Do you remember what Laura thinks?) was coded as correct response if children explained in their responses the complete hypothesis (i.e., "the x ones work, and y ones don't work"), or the effect of one object (i.e., "x ones work," "y ones don't work"). In the effect production condition, children's answers to the memory question (i.e., Do you

remember what is the aim of the game?) was coded as correct response if the children responded by explaining the aim (i.e., to make the box light up). In both of the conditions, absence of response, incorrect responses, and indistinctive responses such as uttering a color-word (e.g., red) were coded as "no response/incorrect." In one hypothesis testing condition, the experimenter forgot to ask the question and hence that data point is missing.

3.1.3 Results and Discussion

There was no effect of the order of color pairs or the order of the effect of color pairs in the familiarization phase on children's choice of object in the experimental phase. There was no effect of the color-pair and the protagonist photo on children's choice of object in the experimental phase. There was no significant difference between males and females in terms of choosing the object in the test phase. There was no general novelty or familiarity preference for the effect-unknown and effect-known objects.

A chi-square test of association was conducted between condition (hypothesis testing vs. effect production) and choice of object (effect-known vs. effect-unknown) to investigate whether there were any differences in children's choice of object in the two conditions. All expected cell frequencies were greater than five. There was a statistically significant association between condition and choice of object, $\chi^2(1)= 20.036$, $p < .001$, two-tailed, $\varphi = .425$. Figure 3.4 displays the percentages of children choosing the effect-known and the effect-unknown objects in each condition.

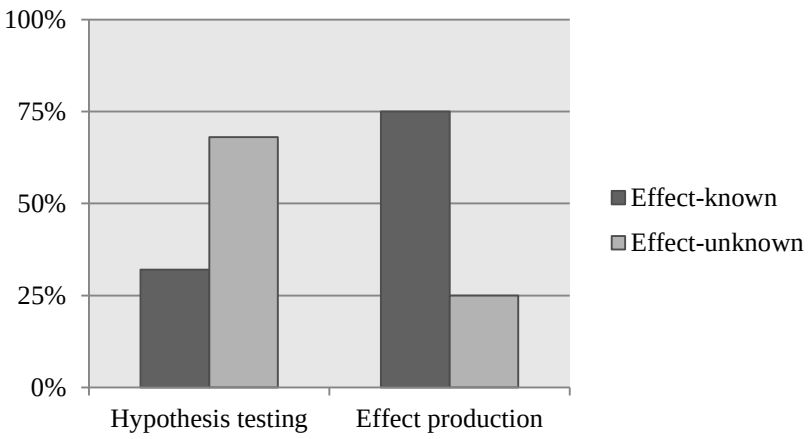

Figure 3.4. Percentages of object choice by condition in Study 1a

In order to investigate how the choice of object varied as a function of age, a binomial logistic regression analysis was conducted on participants' choice of object using condition and age as independent variables. The linearity of the continuous variable, age, with respect to the logit of the dependent variable choice of object (effect-known vs. effect-unknown) was assessed via Box-Tidwell (1962) procedure (p = .579). Based on this assessment, the continuous variable, age, was found to be linearly related to the logit of the dependent variable. The condition variable was dummy coded using the effect production condition as the baseline. Continuous predictor, age, was centered around the mean. The model included condition as the main predictor variable, age as moderator variable and Condition x Age interaction. Results showed that the model was significant, $\chi^2(3)= 27.907$, $p < .001$. The model explained 30% (Nagelkerke R^2) of the variance in children's choice of object and correctly classified 70.3% of the cases. The main effect of condition was significant, $p < .001$, Odds Ratio = 7.08, 95% CI = 2.95-17.01, meaning that the odds of choosing the effect-unknown object in the hypothesis testing condition were 7.08 times higher than the odds of choosing the effect-unknown object in the effect production condition. The main effect of age was not significant, p =

.404. Condition x Age interaction was significant, p = .026, Odds Ratio = 3.59, 95% CI = 1.16-11.05. Table 3.1 displays the details of the logistic regression analyses. The odds of older children choosing the effect-unknown object more in the hypothesis testing condition compared to effect production condition were significantly more than the odds of younger children choosing the effect-unknown object more in the hypothesis testing condition compared to effect production condition. Particularly, the Condition x Age interaction is informative regarding the developmental differences since the correct object for the two conditions were different. Table 3.2 shows the percentage of children who chose the effect-unknown object in the hypothesis testing and effect production conditions.

Table 3.1

Logistic Regression Predicting Likelihood of Choice of Object based on Condition, Age, and Condition x Age Interaction

	β	**SE**	**Wald**	***Df***	***p***	**Odds Ratio**	**95% CI for Odds Ratio**	
							Lower	Upper
Condition*	1.96	45	9.17	1	.000	7.08	2.95	17.01
Age	-.36	43	.70	1	.404	.70	.30	1.62
Condition x Age**	1.27	57	4.94	1	.026	3.59	1.16	11.05
Constant	-1.09	31	2.00	1	.001	.34		

Note. *p<.001, **p <.05

Table 3.2

Percentage and Proportion of Children in Three Age Groups' Choosing the Effect-Unknown Object by Condition

	Hypothesis Testing	**Effect Production**
4-year-olds	52% (11/21)	33% (5/15)
5-year-olds	63% (10/16)	24% (6/25)
6-year-olds	89% (17/19)	20% (3/15)

Note. N = 111

We ran exploratory analyses in order to investigate the details of the developmental change. We made a median-split based on children's months of age and ran two separate post-hoc tests for the younger (n = 54, M_{age} = 58 months, range = 49–65) and older group (n = 57, M_{age} = 66–81, range = 66–81). Due to two post-hoc tests, we used a corrected alpha .025 (.05/2) for multiple tests. For the younger group, (all expected cell frequencies were greater than five); a chi-square test of association between condition (hypothesis testing vs. effect production) and choice of object (effect-known vs. effect-unknown) showed that there was no association between condition and choice of object. For the older group, one of the cells was not greater than five, therefore, Fischer's exact test with the variables condition (hypothesis testing vs. effect production) and choice of object (effect-known vs. effect-unknown) was conducted. The test revealed a significant association between condition and choice of object, $p < .001$, two-sided. Figure 3.5 and Figure 3.6 display the percentages of the younger and the older group's choice of object in the two conditions.

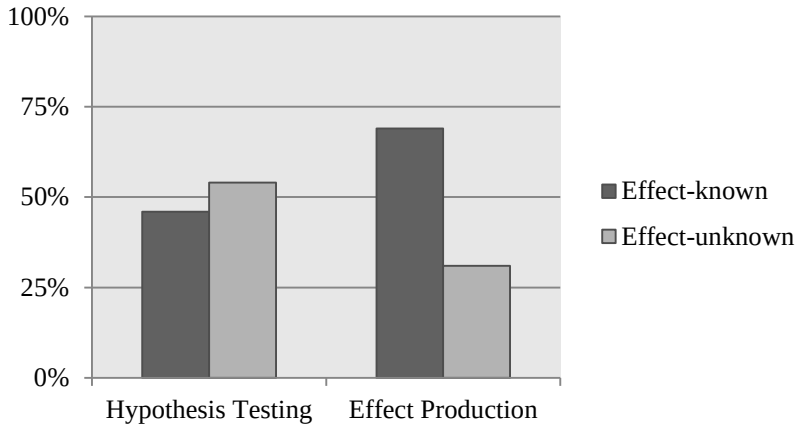

Figure 3.5. Percentages of object choice by condition in the younger group

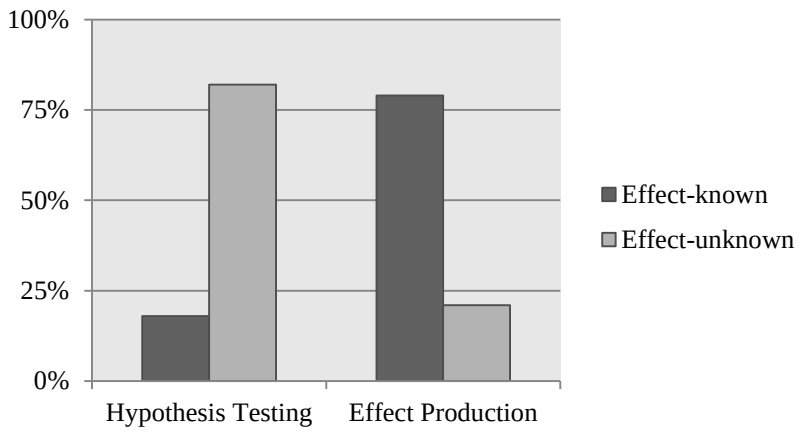

Figure 3.6. Percentages of object choice by condition in the older group

The memory questions were asked to prompt children to keep track of the task goals, not to exclude children based on their response. Exploratory analyses were conducted to investigate whether children's answers to the memory question informed their choice of object in the experimental phase. A chi-square test of association was conducted in order to investigate whether there was an association between memory performance and choosing the correct object in the two conditions. All expected frequencies were greater than five. There was no significant association between memory score and children's object choice in the test phase, $p = .135$, two-sided.

Table 3.3 displays the percentage and proportion of the children's correct object in each condition and their memory score.

Table 3.3

Percentage and Proportion of Choice of Correct Object by Memory Score

	Choice of Object	
Memory Score	Incorrect	Correct
Incorrect/No response	34% (20/59)	66% (39/59)
Correct	21% (10/48)	79% (38/48)

The findings of Study 1a showed that basic competencies for understanding epistemic goals and generating evidence to test a hypothesis are present in the preschool years. Older 5- and 6-year-olds selectively chose the informative piece of evidence in the hypothesis testing condition whereas they chose the object that they knew could produce the desired effect in the effect production condition. However, 4- and younger 5-year-olds did not show any selectivity for object choice across conditions. It was unclear whether this age group's performance is due to their inability to understand and differentially act in the case of epistemic or practical goals, or due to other cognitive demands of the experimental design. In order to investigate this, we conducted Study 1b, in which we decreased certain cognitive demands to see whether this would facilitate the task performance of the younger preschoolers.

3.2 Study 1b

The results of Study 1a showed that older 5- and 6-year-olds were competent in selectively choosing the correct object in the hypothesis testing and effect production conditions. Logistic regression analyses revealed that there was a significant change in performance from 4 to 6 years. Study 1a leaves it unclear whether younger children's poorer performance was due to their inability to selectively choose objects in line with epistemic and practical goals or whether it was due to their inadequacy in learning critical evidence characteristics. Specifically, two evidence characteristics were critical in order for children to perform correctly in the test phase: (1) the cubes with the same color have the same effect, and (2) only one or both of the cubes in a pair might activate the light. Study 1b aimed to rule out lower-level explanations that may stem from these factors and investigate whether younger children's performance would be facilitated when the presentation of the evidence characteristics was made more salient.

Firstly, the learning phase of Study 1a might have been too short for younger children to learn the evidence characteristics. Therefore, the learning phase was prolonged with the rationale that younger preschoolers might benefit from a longer learning phase with an increased number of instances of the each color-pair. Secondly, presenting the evidence characteristics only by means of evidence might have been subtle for the younger children. In order to alleviate this concern, in Study 1b, the experimenter explicitly uttered prompts about critical evidence characteristics. Thirdly, an effect prediction phase was included in order to increase children's active participation in the task. We expected that children who observed the effect of an object and subsequently asked to predict the effects of novel objects with the same color would be more interested in observing whether their choice was correct or false and in return learn the evidence characteristics better. We hypothesized that including such a "prediction-observation" phase would motivate the children to attend to object effects. Lastly, it is possible that the materials used in the Study 1a were confusing for the younger group. Therefore, we simplified Study 1a objects for this study, by using only wire-wrapped cubes in different colors and not providing any information about the wired socket mechanism. These changes were aimed to make the task better suited for testing younger children's differentiation of hypothesis testing and effect production goals.

Previous research (Piekny et al., 2014; Piekny & Maehler, 2013) and our findings from Study 1a suggested that 5 years of age is a critical point in development for scientific reasoning skills. For this reason, the sample of Study 1b consisted of 4-year-olds and young 5-year-olds.

3.2.1 Research Question of Study 1b

Do 4- and younger 5-year-old children selectively choose objects in line with the

epistemic goal of hypothesis testing condition and practical goal of effect production?

3.2.2 Method

3.2.2.1 Participants

In total, 62 children participated in the study. Eight children were excluded due to experimenter error. The final sample included 54 children (29 females, M_{age} = 58 months, range: 48 - 66 months) who were randomly assigned to hypothesis testing (n = 28) and effect production (n = 26) conditions. There was no significant difference between the distribution of gender and months of age in the two conditions. All children were typically developing children of lower- to upper-middle class background from a larger German city. Parents signed written consent for their participation and children were asked for their verbal consent before the study.

3.2.2.2 Materials

A light box which was perceptually similar to the light box in the Study 1a was used. The experimenter controlled the light box via a hidden foot switch. Wire-wrapped cubes in different colors were used as objects. Children were told that the cubes activate the box when they were put on the box. Five color-pairs were used (pink-gray, yellow-green, lilac-orange, black-white, red-blue) in a counterbalanced order in each phase. 7 cm x 7 cm x 3.5 cm brown paper boxes were used to organize the cube-pairs in different phases and to hide the cubes in them before the experimental choice. In total, 25 cubes in 10 different colors (five color-pairs) and 18 paper boxes were used. For the effect prediction phase, two A4 paper sheets were used for sorting the objects: one of them with a picture of an active light box and the other with a picture of an inactive light box. Photos of four children (2 males, 2 females) were used in the experimental phase in a counterbalanced order.

3.2.2.3 Design

The conditions of the study were identical to Study 1a.

3.2.2.4 Procedure

This study consists of three main phases. The general procedure was similar to Study 1a with an additional effect prediction phase.

Learning phase. This phase was similar to Study 1a with two important differences. Firstly, children saw three instances of each color-pair instead of two as in Study 1a. Secondly, the experimenter explicitly uttered the critical evidence characteristics at the end of the phase by saying: "We learned two important things. First, sometimes only one of the colors in a color-pair makes the box light up, sometimes both of the colors make the box light up. Second, when a cube makes the box light up, all the other cubes of that color make the box light up. When a cube doesn't make the box light up, all the other cubes of that color don't make the box light up."

Effect prediction phase. The aim of this phase was to encourage children about the critical evidence characteristics by asking them to predict the effects of novel instances of the cube-pairs after seeing evidence for the cubes of the same color. Children were first given a cube-pair in novel colors (e.g., red-blue) and prompted to try and observe their effects. Subsequently, the experimenter put two sheets of paper in front of the child, one with a picture of an active light box and one with a picture of an inactive light box and instructed children to sort the objects according to their predicted effects. After explaining how to sort the objects, the experimenter asked a memory question to make sure that children understood how the task worked. If children gave an incorrect answer, the experimenter explained one more time and asked again. All children answered the memory question correctly. Subsequently, a novel instance of the cube pair of the same colors as before (e.g., red-blue) was presented, and children were

asked to predict the effects of the cubes in the pair. A similar procedure was repeated a second time with a cube-pair of novel colors. The only difference between the two trials was the effects of each pair. In one of the trials, only one of the colors activated the box whereas in the other set both of the colors activated the box. The order of trials was counterbalanced.

Experimental phase. This phase was identical to Study 1a.

3.2.2.5 Coding

Coding of the choice of object and the memory question was identical to Study 1a. Children's responses in the effect prediction phase were coded for exploratory analyses. Children who sorted both of the cube-pairs correctly got the full score 2, children who sorted only one of the cube-pairs correctly got the score 1, and children who did not sort any of the pairs correctly got the score of 0.

3.2.3 Results and Discussion

There was no effect of the order of color pairs in the familiarization phase on children's choice of object in the experimental conditions. There was no effect of the color-pair and the protagonist`s photo on children's choice of object in the experimental phase. There was no significant difference between males and females in terms of choosing the object in the test phase. There was no general novelty or familiarity preference for the effect-unknown and effect-known objects.

A chi-square test for association was conducted between condition (hypothesis testing vs. effect production) and choice of object (effect-known vs. effect-unknown). All expected cell frequencies were greater than five. There was no significant association between condition and choice of object, $\chi^2(1)= .727$, $p = .394$. Figure 3.7 displays the percentages of children choosing the effect-known and the effect-unknown

object in each condition. Contrary to the predictions, in Study 1b, there was no facilitation in children's selective responses in the hypothesis testing and effect production conditions.

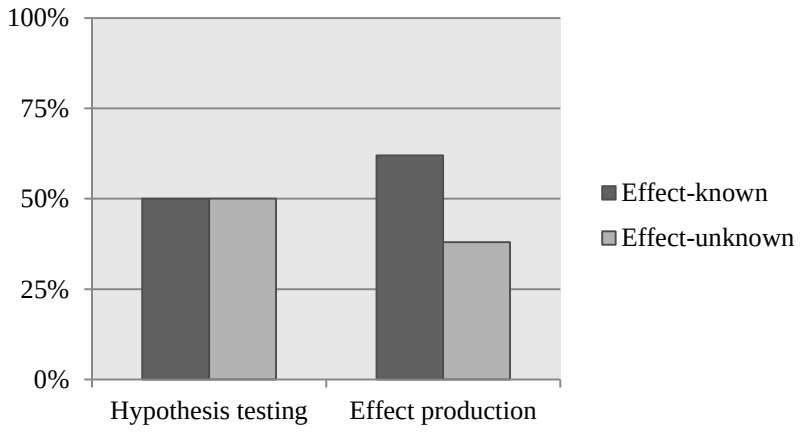

Figure 3.7. Percentages of object choice by condition in Study 1b

Exploratory analyses were conducted to investigate whether there were any associations between children's effect prediction and their choice of object in the experimental phase. A Cochran-Armitage test of trend was conducted to investigate whether a linear trend exists between effect prediction score and choice of correct object depending on the condition. The ordinal response categories ranged from 0 to 2 as children who sorted none of the pairs correctly (n = 3), children who sorted only one pair correctly (n = 24), children who sorted both pairs correctly (n = 27). The proportion of children choosing the correct object was .67, .50 and .59, respectively. The Cochran-Armitage test of trend did not show a statistically significant linear trend between effect prediction score and choosing the correct object, p = .760. Therefore, the results suggested that children's effect prediction performance does not inform how they performed in the test phase. This might be due to the fact that prediction phase was cognitively demanding.

Exploratory analyses were conducted to investigate whether children's answers to the memory question informed their choice of object in the experimental phase. A chi-square test of association was conducted in order to investigate whether there was an association between memory performance and choosing the correct object in the two conditions. One participant's score is missing due to experimenter error. All expected frequencies were greater than five. There was no significant association between memory score and children's object choice in the test phase, $p = .413$, two-sided. Table 3.4 displays the percentage and proportion of the children's correct object in each condition and their memory score. Children's memory scores did not inform their performance in the test phase.

Table 3.4

Percentage and Proportion of Choice of Correct Object by Memory Score in Study 1b

	Choice of Object	
Memory Score	Incorrect	Correct
Incorrect/No response	47% (17/36)	53% (19/36)
Correct	35% (6/17)	65% (11/17)

Note. $N = 53$

3.3 Study 1c

Study 1a and 1b showed that 4- and young 5-year-olds did not make selective interventions in line with the epistemic and practical goals of hypothesis testing and effect production conditions, respectively. This finding is in line with the findings of Piekny et al. (2013) and Piekny and Maehler (2014), in which children younger than 6 years of age did not perform the correct test in the hypothesis testing condition. Importantly, however, previous research indicates no such age difference in exploration tasks (e.g., Cook et al., 2011). Even 3-year-olds have been shown to make informative

interventions in the case of ambiguous evidence (Schulz & Bonawitz, 2007; Legare, 2012). The findings of these exploration studies suggest that younger preschoolers have at least an implicit representation of epistemic goals. That young preschoolers perform poorly in hypothesis testing tasks (Study 1a, Study 1b; Piekny et al., 2013; Piekny & Maehler, 2014), while choosing informative interventions in the case of exploratory goals suggest that perhaps, it is not an understanding of epistemic goals that is lacking at this age, but the concept of hypothesis testing instead. In other words, it may be harder for younger preschoolers to understand what it means to *test a hypothesis*: generating evidence in order to gain information regarding the truth or falsity of a given statement.

In Study 1c, we aimed to investigate younger preschoolers' differentiation of practical goals from exploratory epistemic goals with our paradigm. In order to examine this, we changed a critical aspect of the hypothesis testing condition. Rather than presenting the children with a hypothesis statement to test, we presented them with an exploratory question, "Which ones make the box light up?", and examined younger preschoolers' understanding of exploratory epistemic goals. The correct response, in this case, was the same as in Study 1a and Study 1b, which was choosing the effect-unknown object. Similar to the early studies, success in this condition requires an evaluation of the early evidence and choosing the informative object regarding the research question. If children selectively chose the correct objects in each condition, this would suggest that they have an understanding of exploratory epistemic goals.

3.3.1 Research Question of Study 1c

Do 4- and younger 5-year-old children selectively choose objects in line with the exploratory epistemic goals and practical goals?

3.3.2 Method

3.3.2.1 Participants

A total of 62 children participated in the study. Eight children were excluded due to experimenter error. The final sample included 54 children (30 females, *M* = 57 months, range: 45 - 66 months). Children were randomly assigned to hypothesis testing (*n* = 28) and effect production conditions (*n* = 26). There was no significant difference between the distribution of gender and months of age in the two conditions. All children were typically developing children of lower- to upper-middle class background from a larger German city. Parents signed written consent for their participation and children were asked for their verbal consent before the study.

3.3.2.2 Materials

The materials used in Study 1b were used in Study 1c.

3.3.2.3 Procedure

The general procedure was similar to Study 1a. This study consists of a learning phase and an experimental phase.

Learning phase. This phase was identical to the learning phase in Study 1a.

Experimental phase. This phase was identical to the experimental phase in Study 1a, and 1b except no protagonist was introduced. In the hypothesis testing phase, children were told that the aim of the game was to find out which cubes make the box work whereas in the effect production phase the aim was to make sure that the box lights up.

3.3.2.4 Coding

Children's choice of object (effect-known vs. effect-unknown) in the experimental phase was coded as in Study 1 and 2. Choosing the effect-unknown object

was coded as the correct choice in the hypothesis testing condition whereas choosing the effect-known object was coded as the correct choice in the effect production condition. Children's responses to the memory question were coded for exploratory analyses. In the hypothesis testing condition, the utterances meaning finding out which objects make the box work were coded as correct, in the effect production condition, the utterances meaning making the light box work were coded as the correct response. The absence of response and incorrect responses were coded as no response/incorrect.

3.3.3 Results and Discussion

There was no effect of the order of color pairs and the order of the effect of color pairs in the familiarization phase on children's choice of the object in the experimental. There was no effect of the color-pair and the character photo on children's choice of object in the experimental phase. No difference between males and females in terms of choice of object in the experimental phase.

A chi-square test of association was conducted between condition (hypothesis testing vs. effect production) and choice of object (effect known vs. effect-unknown). All expected cell frequencies were greater than five. There was a significant association between condition and choice of object, $\chi^2(1)= 4.676$, $p = .031$, Cramer's V = .294. Figure 3.8 displays the percentages of children choosing the effect-known and the effect-unknown objects in each condition.

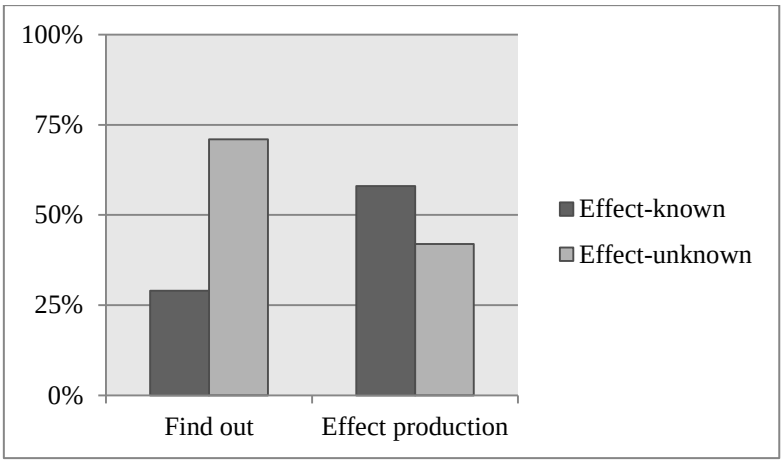

Figure 3.8. Percentages of object choice by condition in Study 1c

Exploratory analyses were conducted to investigate whether children's answers to the memory question informed their choice of object in the experimental phase. A chi-square test of association was conducted in order to investigate whether there was an association between memory performance and choosing the correct object in the two conditions. The score of one child was missing due to experimenter error. One of the expected cell frequencies was five. Therefore Fisher's exact test was conducted. There was no significant association between memory score and children's object choice in the test phase, p = .375, two-sided. Table 3.5 displays the percentage and proportion of the children's correct object and their memory score. Children's memory scores did not inform their performance in the test phase.

Table 3.5

Percentage and Proportion of Choice of Correct Object by Memory Score in Study 1c

	Choice of Object	
Memory Score	Incorrect	Correct
Incorrect/No response	41% (14/34)	59% (20/34)
Correct	26% (5/19)	74% (14/19)

Note. N = 53

3.4 General Discussion

In three studies, Study 1 investigated 4- to 6-year-old preschoolers' ability to differentiate epistemic goals from practical goals. Study 1a showed that preschool age is a critical period for the development of hypothesis testing skills. The present findings suggest that the ability to test hypotheses, which requires generating relevant evidence in line with specific epistemic goals, develops in the preschool years. Study 1a revealed that children differentially chose the correct object in response to epistemic and practical goals. In the hypothesis testing condition, children chose the object which would reveal the necessary information to test the given hypothesis. This demonstrates that children can pinpoint the critical piece of evidence which would provide information regarding the truth or falsity of the hypothesis. On the other hand, in the effect production condition, children chose the object which they had already observed as an efficacious object. Moreover, the results of Study 1a demonstrated that there is a critical change in performance from 4 to 6 years of age in terms of differential responding to hypothesis testing and effect production. The older the children were, the better their performance was.

The results of Study 1b further affirmed that the observed age-wise differences were indeed the result of a developmental effect by showing that poor performance of the younger preschoolers was not due to the subtle presentation of evidence characteristics. Providing a longer learning phase, perceptually simpler materials, and explicit evidence summaries did not facilitate younger children's hypothesis testing performance. There was no significant difference between children's responses in the case of hypothesis testing and effect production. This suggests that younger children's poor performance in our paradigm cannot be fully accounted for low-level explanations such as the short familiarization phase or the complexity of the experiment materials.

Different from Study 1b, Study 1c demonstrated that younger preschoolers chose the informative object more when they were presented with simpler, exploratory epistemic goals. The results of the three studies together suggest that there is a developmental change in hypothesis testing abilities from 4- to 6-year-old children. While 4- and younger 5-year-olds did not differentially make interventions in response to hypothesis testing and effect production goals; they did differentially make interventions in the case of exploratory epistemic goals and practical goals.

Research on theory of mind has demonstrated that around 4 years of age children show a developmental change in the ability to understand that people may hold different beliefs based on information available to them (Doherty, 2009). It has been argued that this development is foundational for scientific reasoning because the understanding that there might be alternative beliefs about a state is crucial for scientific reasoning (e.g., Kuhn, 2010). The developmental pattern that we found between 4- and 6-years does not contradict with the hypothesis that children who develop theory of mind skills have the ability to differentiate epistemic goals from practical goals. In this respect, future studies should investigate within-participant performance in our task together with the performance in a theory of mind tasks.

Contrary to the early conceptual deficit claims (Kuhn, 1989; Kuhn & Franklin, 2006), our results showed that older preschoolers do not have a conceptual deficit in differentiating hypothesis testing from effect production. When presented with simple hypotheses and engaging tasks, preschoolers (i.e., older 5-year-olds and 6-year-olds) understand what it takes to have an epistemic goal, and they generate relevant evidence necessary to test the truth of a given hypothesis. However, when the epistemic goal requires deciding on the truth or falsity of a claim by generating evidence, younger children did not show competence. On the other hand, when the epistemic goal is just

about information gain, the younger group selectively chose the most informative object. Put together, our findings suggest a change from interventions directed towards exploratory goals to interventions aimed at hypothesis testing.

It is interesting that 4- and younger 5-year-olds failed at testing a particular hypothesis, while they chose the informative object when faced with an exploratory epistemic goal. It is an open question what develops in this age period that leads to this developmental change from 4 to 6 years of age. One possibility is that at this age, the ability to make inferential relations from evidence regarding the truth or falsity of hypotheses develops. Although younger preschoolers have a basic understanding of epistemic goals and they are directed at maximizing information gain in the case of ambiguity, they may lack the ability to make purposeful interventions to test a hypothesis. Another possible explanation for the difference between younger and older children might be due to the development of metacognitive abilities. Rohwer et al. (2012) have demonstrated that around 5.5 years of age, children can reflect on their epistemic states in the case of partial information. Children younger than 5.5 years of age mistakenly reported knowledge in the case of partial information when they were indeed ignorant. In contrast, children older than 5.5 years were able to report that they were ignorant in the case of partial information. It is a possibility that younger children in the present study did not perceive the given hypothesis as a statement that needs to be tested. One requirement for hypothesis testing is to be able to reflect on one's epistemic states and acknowledge that one is ignorant. Only after this, one can design a correct test to gain further knowledge. It might be that younger children did not have the understanding that they were ignorant, and this might be the reason of their poor performance in testing a given hypothesis.

Different from exploration studies, Study 1 also provides information regarding children`s strategic choice of object. In our study, children were instructed by an experimenter to test a hypothesis or to produce an effect. They were restricted to choose one object. Therefore, they should have made a decision before acting on the objects, and they needed to inhibit other motivations if they had any. To illustrate, a child might want to try the pink cube just because he likes the color pink. Therefore, intrinsic motivations that contradict with the goal of the task should be inhibited, and the children should be able to orient themselves to successfully achieve the goal presented to them. Our study shows that older preschoolers are able to make different interventions in response to different goals provided to them. On the other hand, it is worth noting that there were not many action possibilities in our test trial since there were only two object choices and we restricted children to choose only one. Therefore, although our task required planned decision making and inhibition of contradictory intrinsic motivations, it is an open question how children of this age would perform in hypothesis tasks, when there are more action possibilities.

Older 5-year-olds' and 6-year-olds' selective choice of objects suggests that they differentiate epistemic goals from practical goals and that they have a preliminary understanding of the relation between hypotheses and evidence. However, one open question is whether children have a metacognitive awareness of this differentiation. The ability to reflect upon one's own knowledge acquisition processes is fundamental for the development of scientific reasoning skills. However, Study 1 does not answer this question. During the experimental sessions, a few children spontaneously provided justifications for their object choices. For instance, in the hypothesis testing condition, they said that they chose the effect-unknown object, because they already knew the effect-known object, so they needed information regarding the other object to find out

whether the protagonist was right. Such verbal justifications suggest that those children, indeed, have a metacognitive awareness of their own epistemic state, and of what it means to generate evidence in order to test a claim. They understand that more information is required to test the specific hypothesis and they can justify why a piece of evidence is relevant for testing the truth of a given hypotheses. Yet, we do not know whether this ability can be generalized to the whole population. Further studies might investigate children's metacognitive understanding by explicitly asking them to provide justifications for their object choices.

Another open question is how within-subject performance would be in the hypothesis testing and effect production conditions. Our pilot study with within-subject design suggested that there were carry-over effects across two conditions: in the second condition, children perseverated on choosing the object that they had chosen in the first condition. This revealed a pattern that children generally performed correctly in the first condition but wrongly in the second condition independent of the order of the conditions. Research on young children's executive functioning shows that inflexibility in switching across different task goals is typical for children of this age (Zelazo, Muller, Frye, & Markovitch, 2003). In our within-subjects pilot study, the task materials and procedure were all identical except for the goal of the study. It is possible that the high resemblance across the materials and the procedure contributed to the observed trend of choosing the same object across conditions. In future studies, using perceptually different materials might be useful for preventing such a carry-over effect.

Studying the development of young children's differentiation of hypothesis testing from effect production is crucial because it informs us about whether young children are capable of representing epistemic goals in the first place. Scientific reasoning studies with older children provide evidence that children often confuse

hypothesis testing with effect production. Yet, at least for older preschoolers, the ability to generate relevant evidence in line with epistemic goals seems to be present in the case of simple hypotheses. Our results show an important distinction in how children's ability of hypothesis testing manifests under conditions of exploratory epistemic goals and hypothesis testing. There are several important questions that future research can address: (a) Is this pattern of change from exploration to experimentation that we found in our three studies generalizable?, (b) Is there a continuity from exploration to experimentation during development?, (c) Which cognitive skills (The theory of mind, language, metacognition, general intelligence) play a role in this change in this age group?

In sum, the results of Study 1 showed that preschool children have a preliminary understanding of hypothesis–evidence links as demonstrated by their differentiation of epistemic goals from practical goals and their selective interventions in response to different task goals. Older preschoolers, but not younger ones, selectively chose the correct objects in line with epistemic and practical goals, which demonstrates that they are able to choose informative evidence in order to test hypotheses. On the other hand, younger preschoolers differentiated epistemic goals from practical goals only in the case of exploratory goals, but we did not find any evidence showing that they can generate a relevant piece of evidence to test a particular hypothesis. Taken together, Study 1 suggests that while some skills for hypothesis testing are present already at the age of four, they continue to develop throughout the preschool years.

4 Study 2: Hypothesis Testing and Argumentation from Evidence in Young Children

Study 1 demonstrated that preschoolers, especially the older 5- and 6-year-olds, are able to differentiate epistemic goals from practical ones and they generate informative evidence in response to epistemic goals of hypothesis testing. These findings suggest that preschoolers have a basic understanding of the relation between hypotheses and evidence which is considered as the key competence for scientific reasoning. Although these findings expand our knowledge on early competences, much is still unknown about whether and to what extent young children are similar to scientists in terms of intentionally guided knowledge seeking processes that characterize scientific reasoning (Kuhn, 2010).

Contrary to the views that scientific reasoning skills do not appear until adolescence (Inhelder & Piaget, 1958; Kuhn et al., 1988), recent research has shown that young children are far more skilled in scientific reasoning tasks than early studies claimed (Zimmerman, 2007). Elementary school age children differentiate hypotheses from evidence in preferring a conclusive over an inconclusive test for a hypothesis (Sodian et al., 1991), and a controlled over a confounded experiment (Bullock & Ziegler, 1999). They possess basic evidence evaluation skills (Masnick, Klahr, & Morris, 2007; Masnick & Morris, 2008; Piekny & Maehler, 2013, Saffran et al., 2016), and some understanding of the nature of science (Sodian, Thoermer, Kircher, Grygier, & Günther, 2002). Furthermore, there is evidence that the different components of scientific reasoning form a unitary construct and show regular progression during the elementary school years (Koerber et al., 2015).

There is far less evidence on scientific reasoning skills in early childhood. Croker and Buchanan (2011) demonstrated that even 4-year-olds chose an appropriate test strategy (manipulate one variable) when the evidence was consistent with their prior belief and the outcome was good or when the evidence was inconsistent with the prior belief and the outcome was bad. Ruffman et al. (1993) showed that 6-year-olds can predict that a character who saw a different pattern of evidence from *reality* would form a belief in accordance with the evidence they saw. Koerber et al. (2005) found competence in a similar task even in 4-year-olds, when the patterns of evidence were perfect or near-perfect covariation data. Thus, young children seem to possess a basic understanding of the role of data in belief formation, but it is unclear to what extent they can judge the quality and relevance of a specific piece of evidence.

The findings from causal learning studies (e.g., Gopnik & Sobel, 2000; Kushnir & Gopnik, 2007; Schulz & Gopnik, 2004) impressively document the systematic *use* of covariation evidence in causal learning in making predictions about the efficacy of objects. The *evaluation* of evidence, however, requires a *judgment* about the relevance or informativeness of a piece of evidence with respect to a hypothesis. Schulz and Bonawitz (2007, see Section 2.3.2.3 for details) found that preschoolers preferentially explored toys for which they had received confounded evidence about the causal structure of the toy, rather than matched toys for which they had received unconfounded evidence, and they spontaneously disambiguated confounded variables in exploration, indicating that children maximize the new information to be gained from exploration. While these findings do not necessarily require explicit judgments of informativeness, they indicate some degree of *awareness* of the informativeness of the evidence with respect to an epistemic goal. Thus, an implicit awareness of the quality of evidence may developmentally precede explicit judgments. However, children may learn from

informative evidence and selectively explore ambiguous evidence to maximize informativeness without understanding *what it is about evidence* that makes it informative or uninformative. The study by Cook et al. (2011, see Section 2.3.2.3 for details) describes better young children's understanding of the quality of evidence. Findings of the study demonstrated that young children make informative interventions by isolating variables to identify the cause of an effect. These findings suggest that preschool children do not merely passively process covariation evidence to learn about cause–effect relations but that they begin to use efficient strategies of active experimentation. This conclusion is also supported by recent studies by Legare (2012) and Bonawitz et al. (2012), which demonstrated that preschoolers are sensitive to inconsistent evidence (See Section 2.3.2.3 for details). Especially the findings of Legare (2012) are informative for hypothesis testing because they show that some children provide hypothesis-like explanations for inconsistent evidence and their explanations predict the way children explore the objects later.

The causal learning literature has generally studied exploratory play, rather than experimentation, which is typically investigated in the scientific reasoning literature. Conclusions from the observation of unconstrained exploration processes about the formation and testing of specific hypotheses are indirect and limited. To gain more direct evidence about young children's hypothesis testing skills, it is necessary to ensure that children form a specific hypothesis about a cause–effect relation during an exploratory phase, which they can subsequently test on new materials. Therefore, in Study 2, children were led to form a specific hypothesis, and were presented with test objects (i.e., objects with x and without x) that allowed them to disambiguate the evidence by generating contrastive tests.

We designed a multiphase exploration task in a blicket detector format (Gopnik & Sobel, 2000) to maximize children's awareness of the relation between hypotheses and evidence. We assumed that children's awareness of the fact that hypotheses may be false (or preliminary) might be heightened by experiencing that a causal hypothesis of their own turned out to be false. Therefore, we first induced the belief that the weight of objects was the deterministic cause of the light effect, and subsequently confronted them with inconsistent evidence and evidence for the alternative hypothesis that a sticker on the bottom of the box was the causal factor. This experience should elicit systematic hypothesis testing behavior more than an unconstrained causal exploration, since children experience the falsification of an insufficiently tested initial belief.

We were interested in whether young children would show any consistent strategies under such conditions (as opposed to random exploratory behavior). We examined which objects children engaged with to test "the sticker hypothesis" (i.e., objects with a sticker turn the machine on). We first looked at whether children would follow a two-variable test strategy during exploratory play. This requires incorporating both features (i.e., heaviness and sticker) and testing unambiguous objects where the two variables appear in isolation rather than ambiguous cases where two variables were either present or absent together. If children test more unambiguous than ambiguous cases, this could be taken to indicate an attempt to isolate variables. Even if children do not systematically isolate variables, they may still produce a contrastive testing strategy, in this case a one-variable test strategy, contrasting positive cases where the hypothesized feature was present (i.e., sticker present) and the negative cases (i.e., sticker absent) where the hypothesized feature was absent.

Testing behaviors will not provide us with conclusive information on children's metacognitive understanding of the significance of specific pieces of evidence regarding

the truth or falsity of a causal claim. Therefore, we studied children's protest behaviors against false causal claims as an additional and more explicit indicator of their understanding of the hypothesis–evidence relation. Findings in other areas of developmental research using *protest paradigms* (e.g., Rakoczy, Warneken, & Tomasello, 2008; Schmidt, Hardecker, & Tomasello, 2016) show that young children enjoy correcting others' mistakes. Children who have undergone a two-phase exploration in which they formed and revised a causal belief themselves should therefore be eager to correct a person who utters a wrong causal hypothesis, either the one the child previously held herself and successfully revised when presented with new evidence, or an entirely new one that the child had not held before. This can make it possible to study children's ability to produce evidence-based arguments, by producing valid counterevidence against an experimenter's false claims.

In two counterargumentation tasks (a) novel hypothesis (i.e., Blue objects turn the light on) and (b) familiar wrong hypothesis (i.e., Heavy objects turn the light on.), an experimenter presented a false hypothesis and generated confounded evidence to support it. We were interested in whether children could diagnose that this evidence was confounded and counterargue by means of both evidence generation and verbal counterargumentation. In terms of evidence generation, we examined whether children showed a preference for evidence that disconfirms experimenters` false hypotheses (e.g., a heavy object that did not turn the light on) over objects that confirmed the false hypothesis (e.g., a heavy object that turned the light on). Preference for disconfirming objects would suggest children not only *explore* causally ambiguous cases, which have been mostly measured so far with duration of play or variability in play (e.g., Legare, 2012; Schulz & Bonawitz, 2007), but also can diagnose the conclusive pieces of evidence that is relevant to falsification of a hypothesis. We also examined children's

verbal counterarguments. The presence of evidence-based verbal counterarguments would yield evidence for the presence of, at least, basic metaconceptual understanding of the epistemic relation between hypotheses and evidence. If children showed such a preference for disconfirming objects and provided valid verbal counterarguments, this would suggest an understanding that inferences can be drawn from evidence regarding the truth or falsity of a hypothesis and that they can use correct disconfirming evidence to falsify a claim.

In sum, Study 2 investigated young children's understanding of the relation between causal hypotheses/claims and empirical evidence by studying hypothesis testing strategies and counterargumentation skills. Although there has been evidence for the implicit understanding of the relation between hypotheses/claims and evidence in early childhood, neither testing strategies, nor argumentation skills have been studied systematically.

4.1 Research Questions of Study 2

1. Do children selectively interact with more unambiguous objects (i.e., put on the light box) than ambiguous objects during hypothesis testing phase?
2. What is the frequency of different hypothesis testing patterns (i.e., contrastive testing, positive testing) during hypothesis testing phase?
3. Do children selectively interact with more unambiguous objects in the novel and familiar counterargumentation phases?[11]

[11] The aim of the Study 2 is not to compare children's counterargumentation in response to novel vs. familiar claims; but to investigate whether children can use evidence as a means to falsify claims at all. Considering even small changes in the characteristics of false claims (i.e., the sentence structure, the content of the claim, counterarguing to adults) might trigger different responses, we employed two

4. Do young children verbally counterargue in response to a false causal claims? What is the content of their verbal counterarguments?
5. Is there a developmental difference between younger and older preschoolers with respect to Research Questions 1, 2, 3 and 4?
6. Is there an association between children's hypothesis testing performance and counterargumentation performance?

4.2 Method

4.2.1 Participants

A total of 67 children participated in the study (M_{age} = 64 months; age range: 44–81 months, 37 females and 30 males). Six additional children were tested but we excluded them due to experimenter error (n = 5) or unwillingness to continue the study (n = 1). All participants were typically developing children of lower- to upper-middle class background from a larger German city. The sample was divided into two age groups (younger and older) via median split (*Median* = 63 months). There were 34 children in the younger group (M_{age}= 55 months, age range = 44 – 63 months) and 33 children in the older group (M_{age} = 73 months, age range = 64 – 81 months). Parents signed a written consent for their children's participation and children were asked for their verbal consent before the study.

counterargumentation tasks to increase the possibility of capturing the ability to generate disconfirming evidence. We set liberal criterion for deciding whether children has this ability. If there is success at least in one of the two tasks, it was considered as evidence for children's ability for using correct piece of evidence in order to falsify claims.

4.2.2 Materials

The light box which was used in Study 1 has been used. Children were told that certain objects activated the light box when they were put on it and some other objects did not. In reality, a second experimenter who sat behind children controlled the light box via a remote control. Nineteen paper boxes were used as objects. They varied in shape, color, weight and in whether or not they had a black sticker on the bottom. The deterministic causal factor was the sticker on the bottom; therefore, the objects with a sticker always activated the light box and objects without a sticker never activated it. (See Figure 4.1 for the exemplars of the materials). Weight (heavy or light) was the causal distractor. In total, six out of ten heavy objects activated the light box. The objects also varied in color (red, yellow, blue, green) and in shape (square, hexagon, round) and these two features were never a systematic cause for the light in any phase of the study. The objects' lid was glued to their bottom so they could not be opened. Two A4 paper sheets were used for sorting the objects: one of them with a picture of an active light box and the other with a picture of an inactive light box.

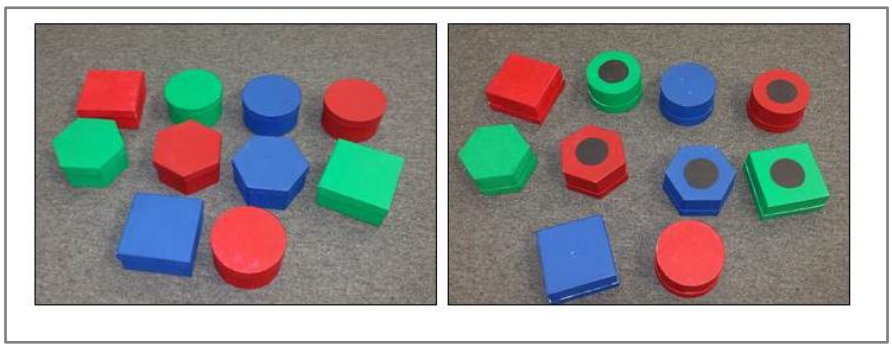

Figure 4.1. Exemplars of materials of Study 2. On the left, picture of the objects (sticker-hidden). On the right, picture of the objects upside down (sticker-salient).

4.2.3 Procedure

All sessions were carried out in separate quiet rooms of kindergartens and recorded by a video camera. Each child was tested individually in a session lasting approximately 15 minutes. The procedure took place on a mat on the floor to maximize children's free movement. Two experimenters were involved: one of the experimenters interacted with the children and the second experimenter controlled the light box. At the beginning of each session, both experimenters introduced themselves and played a warm up game with the children. At the end of the warm up game, the second experimenter said she was going to read a book and told them to continue playing. The second experimenter sat behind the children so that children could not see her during the study. The procedure of the study consisted of five successive phases. Table 4.1 presents the aim of each phase, the salience of the causal features in that phase and the evidence provided in each phase.

Table 4.1

Evidence Characteristics in terms of the Salience of the Causal Factors and Objects in Each Phase of the Study

Phase	**Saliency of the Factor**	**Object Properties**
1. Distractor belief Evidence supports the weight hypothesis	Salient factor: Heaviness Hidden factor: Sticker	Heavy (with sticker) Heavy (with sticker) Light (without sticker) Light (without sticker)
2. Inconsistent evidence Evidence inconsistent with weight hypothesis	Salient factor: Heaviness Hidden factor: Sticker	Heavy (without sticker) Light (with sticker)
3. Hypothesis testing Evidence supports sticker hypothesis	Salient factor: Heaviness Salient factor: Sticker	Heavy with sticker Heavy without sticker Light with sticker Light without sticker

4. Belief check	Salient factor: Heaviness	Heavy with sticker
	Salient factor: Sticker	Light without sticker
		Heavy without sticker
		Light with sticker
5. Counterargumentation	Salient factor: Heaviness	Example: Heavy with sticker
	Salient factor: Sticker	Light without sticker
		Heavy without sticker
		Light with sticker
		Heavy with sticker

Distractor belief. The experimenter introduced the light box and the objects and demonstrated how the light box worked. The aim of this phase was to induce a prior belief: "the weight belief." Children were given four objects (two heavy activators and two light nonactivators) and were encouraged to explore and find out what makes the box light up. After children had tried all of the objects and showed correctly the objects which activated the light box, they were asked whether they had an idea regarding the cause of the light. In this phase, if children explained that heaviness was the causal factor, the experimenter skipped to the next phase. If children did not refer to heaviness as the cause or if they suggested another cause, the experimenter encouraged children to explore one more time and prompted them by asking for the differences between the objects that make the box light up and the objects that did not.

Inconsistent evidence. The experimenter revealed two novel objects, handed them to the child, and asked the child to predict which one would make the box light up. After children made a prediction, the experimenter put two of the objects on the light box one-by-one and children observed their effect. The evidence was inconsistent with the weight belief: the heavy object did not activate the light box, while the light object did.

Hypothesis testing. The experimenter introduced four novel objects and put them in front of the child in a 2 x 2 (front, back) order. The arrangement of the objects was counterbalanced. There were four different presentation orders. Before making them available to the children, the experimenter pretended to accidentally show the bottom of two objects to make the sticker perceptually available to the children. The order of objects was counterbalanced in a way that one of the objects that the experimenter showed always had a sticker on the bottom and the other one did not have a sticker. If children did not show any verbal or behavioral sign that they had noticed the sticker, the experimenter explicitly showed the bottom of the two objects and said, "Look, there is something here". Afterwards, the children were encouraged to explore until they stopped engaging with the objects. The experimenter prompted children once ("You can try and find out what makes the box light up") if they did not spontaneously interact with the objects.

Belief check. The objective of this phase was to assess children's causal beliefs after the hypothesis testing phase and before the counterargumentation phase. The experimenter put two sheets of paper in front of the child, one with a picture of an active light box and one with a picture of an inactive light box and instructed children to sort the objects according to their predicted effects. After explaining how to sort the objects, the experimenter asked a memory question to make sure that children understood how the task worked. If children gave an incorrect answer, the experimenter explained the task one more time and asked again. All children answered the memory question correctly. Afterwards, the experimenter introduced four novel objects one at a time in a counterbalanced order and asked whether they would make the box light up or not.

Counterargumentation. In this last phase, the two experimenters each provided a false hypothesis and confounded evidence supporting it; subsequently the

experimenter asked the child whether her hypothesis was right. Only the children who had passed the belief check in phase 4 were included in the counterargumentation. The first experimenter brought up five novel objects and encouraged children to try the objects. All children tried all five objects. Subsequently, the experimenter put four objects in front of the child in a 2 x 2 (front, back) arrangement before each counterargumentation task. The arrangement of the objects was counterbalanced. There were four different presentation orders. The experimenter used one object as an example object while presenting the false claim and put away this object out of children's reach afterwards. The order of the "heavy counterargumentation" and "blue counterargumentation" phases was counterbalanced.

Heavy counterargumentation. In the phase of heavy counterargumentation, the first experimenter invited the second experimenter to play, explained the aim of the game (i.e., finding out what makes the box light up) and asked whether she knew what makes the box light up. The second experimenter said, "I think the heavy ones make the box light up," then she put a heavy activator (with sticker) on the light box and said, "Look the box lights up because this [the object] is heavy. Right?" The experimenter prompted the children who did not respond spontaneously once ("Am I right? Can you show me how you know?").

Blue counterargumentation. The first experimenter said, "We tried so many objects, I am confused. The blue ones make it light up." Then, she put a blue activator (with sticker) on the light box and said, "Look, the box lights it up because this [the object] is blue. Right?" If children did not give an immediate response the experimenter prompted them in the same way as in the heavy counterargumentation.

4.2.4 Coding

Coding belief check. Children's predictions in the inconsistent evidence phase and in the belief check phase were coded. In the inconsistent evidence phase, we were interested in whether children formed the weight belief (i.e., Heavy objects activate the light box). The responses of children who predicted that the heavy object would activate the light box were coded as having formed the weight belief and only those children were included in the hypothesis testing phase.

In the belief check phase, we were interested in whether children formed the "sticker belief" (i.e., Objects with sticker will activate the light box). Children's predictions for whether each object would activate the light box were coded. This was a binary code representing children's predictions for the four objects. The individual responses for the four objects were combined into a final conservative belief score for each participant. Children who predicted that only the objects with the sticker would activate the light box were coded as having a sticker belief and only those children were included in the counterargumentation phase.

Coding hypothesis testing. Coding hypothesis testing started when the experimenter made the sticker salient to the children at the beginning of hypothesis testing and ended when children stopped engaging with the objects in this phase. Our main interest was to investigate whether children's exploration behavior indicated a higher level two-variable testing strategy (i.e., isolating the variables weight and sticker) or a one-variable contrastive testing strategy for the novel feature (i.e., contrasting objects with and without a sticker). A behavior was coded as a *test* if children checked the bottom of an object for the sticker immediately before or after putting it on the light box or if children verbally indicated that they knew whether an object had a sticker or

not, and put it on the light box. The children who put the objects on the light box without checking their bottoms were coded as *untargeted exploration.*

To examine two-variable testing, the sum scores of the instances of testing ambiguous objects and the sum score of instances of testing unambiguous objects were computed. The comparison of both scores indicates whether there was any preference for unambiguous objects in evidence generation. To examine one-variable testing, we analyzed whether children did a contrastive test or a positive test. Children who tested at least one object with sticker and at least one object without sticker were coded as doing a *contrastive test*. Children who only tested objects with sticker were coded as doing a *positive test*. Some children were not testing the sticker hypothesis because they did not look at the bottom of the objects to see the sticker when they were trying them. This suggests they did not recognize the novel salient feature as the guiding hypothesis. These children were coded as doing untargeted exploration.

Coding counterargumentation. Children's spontaneous verbal comments and evidence generation behaviors in response to the experimenters' false hypotheses were coded separately for the two counterargumentation tasks.

Verbal counterarguments. Each child received a verbal counterargumentation score from two possible scores; namely (1) *explicit verbal counterargumentation* or (2) *no verbal counterargumentation*. Explicit verbal counterargumentation included utterances of explicit disagreements ("No, not all the heavy ones"), disconfirming statements ("Sometimes light ones light up, too"), or suggesting an alternative cause for the effect ("It makes the box light up because the black thing [the sticker] on the bottom"). A few children provided gestures as head shakes accompanied with pointing to the sticker and they were also coded as explicit verbal counterargumentation. There were also some children who took a teaching attitude. They said "Try this one" and offered the

experimenter a disconfirming object. These few cases were also coded as explicit counterargumentation responses. Children who only provided descriptive statements, children who agreed with the experimenters' false hypothesis and children who did not respond or said they did not know were coded as no verbal counterargumentation response. Descriptive statements were the cases when children tried the objects and only described their effects without providing any disagreement (e.g., trying a light activator and saying: "This makes it light up", trying a light nonactivator and saying: "This doesn't make the box light up"). Children who agreed by saying yes or head nod were coded as agreement. Children who only said "No" or made a head shake response without making any other elaborations were also coded as no verbal counterargumentation.

The contents of children's verbal counterarguments were also coded. Children referred to at least one causal feature or both as a justification of their argument and this resulted in three codes for the type of content: (1) the comments that referred to the actual causal factor; the sticker on the bottom ("Only the ones with the stickers make the box light up"), (2) the comments that referred to the causal factor mentioned in the false hypothesis ("Not all blue ones make it work. Yellow ones sometimes work, too") and (3) the comments that referred to both features ("Sometimes blue ones and sometimes the other colors work. It is because of the magnet on the bottom").

Evidence generation. The type and the frequency of the objects children put on the light box after the experimenters presented the false hypotheses were coded. For both counterargumentation phases, two of the objects available to the children were confirming the experimenters' false hypothesis and two of the objects were disconfirming it. In the blue counterargumentation (i.e., "Blue objects make the box light up"), a blue nonactivator and a nonblue activator were disconfirming evidence

whereas a blue activator and a nonblue nonactivator were confirming evidence. In the weight counterargumentation (i.e., "Heavy objects make the box light up"), a light activator and a heavy nonactivator were disconfirming evidence; and a heavy activator and light nonactivator were confirming evidence. First, children's frequency of putting each object on the light box were counted, and later the sum of two-disconfirming objects was calculated for a final disconfirming evidence score and the sum of confirming objects calculated for a final confirming evidence score.

We investigated children's individual competence level of counterargumentation by combining their individual scores of verbal counterarguments and evidence generation. Children who provided an explicit disagreement and, at the same time, who only provided distinctive disconfirming evidence were coded as displaying *full competence*, children who either only provided explicit disagreement or who only generated distinctive disconfirming evidence were coded as displaying *partial competence* and children who did not generate distinctive disconfirming evidence and did not provide any explanations were coded as showing *no competence*. Thus, children received final scores for each of the two counterargumentation tasks.

Twenty percent of the cases were coded by a second rater. The interrater reliability for the inconsistent evidence and belief check codes was perfect (Kappa = 1.00). The interrater reliability for evidence generation (ICC ranging between .87 and .96), one-variable hypothesis testing (Kappa = .87) and verbal counterarguments was high (ICC ranging from .81 and .89).

4.3 Results

We first present children's responses in the inconsistent evidence and belief check phases, followed by hypothesis testing performance. This is followed by the

analyses of counterargumentation in the two counterargumentation phases (heavy counterargumentation and blue counterargumentation). There was no significant gender difference in any of the measures.

We had two measures to test children's evidence generation competence in the counterargumentation phase. Since our aim was to investigate whether there is any competence for argumentation from evidence in early childhood, we set liberal criteria and considered competence in at least one of the two counterargumentation tasks as an indicator of the presence of evidence generation competence. As we ran two significance tests to investigate evidence generation competence in argumentation, to control the family wise error rate, we applied Bonferroni correction with alpha level of .025 ($p < .05/2$) for evidence generation scores of the argumentation. Similarly, we ran two significance tests to investigate whether there is any relation between children's hypothesis testing performance and their performance in two counterargumentation tasks (heavy, blue). For those analyses, we set the alpha level to .025 ($p < .05/2$).

Inconsistent evidence and belief check. Preliminary analyses revealed no effect of order of object presentation on children's belief check scores. Among the 67 children, 62 (93%) predicted that the heavy object would activate the light box and the light object would not in the inconsistent evidence phase and only those children were included in the hypothesis testing. Five children who did not respond in line with the weight hypothesis were excluded from further analyses (two 4-year-olds, two 5-year-olds, and one 6-year-old). In the belief check phase, 76% (47 out of 62) of the children predicted that only the objects with a sticker would activate the light box. Only children who made this correct prediction were included in the counterargumentation analysis. Fifteen children who did not respond in line with the sticker hypothesis were excluded

from the counterargumentation analyses (one 3-year-old, nine 4-year-olds, two 5-year-olds, and three 6-year-olds).

Hypothesis testing. No effect for the presentation order of the objects was found. The first research question was whether children would perform tests considering the two variables (distractor: heaviness and novel: sticker) and try more unambiguous objects (heavy without sticker and light with sticker) in which the distractor feature and the novel salient feature appeared in isolation, compared to ambiguous objects in which the two features were either present together (heavy with sticker) or absent (light without sticker). An Age Group (younger, older) & Ambiguity (unambiguous, ambiguous) mixed-design ANOVA, Ambiguity as the within-subject factor and the number of objects tested as the dependent variable was conducted. The main effects of the variables Ambiguity (p = .593) and Age Group (p = .807); and the interaction effect between Age Group x Ambiguity (p = .986) were nonsignificant. This means children did not show any preference for unambiguous objects (M = .58 out of 2, SD = .80) over ambiguous objects (M = .65 out of 2, SD = .77). This means children did not attempt to isolate variables, showing no preference for unambiguous objects over ambiguous objects.

To examine one-variable testing, children were classified into three categories for their hypothesis testing pattern for the sticker (contrastive test = 3, positive test = 2, untargeted exploration = 1). Table 4.2 displays the proportion and percentages of one-variable hypothesis testing in hypothesis testing phase. Fisher exact test revealed a significant association between hypothesis testing patterns and age group, N = 62, p = .002, Cramer's V = .44. Post hoc tests using Bonferroni correction revealed that the number of children who did contrastive tests and who did untargeted exploration was significantly different in the two age groups, Fisher exact test, N = 40, one-sided, p =

.015, Cramer's V = .39. Similarly, the number of children who did positive test and untargeted exploration was significantly different in the two age groups, Fisher exact test, N = 33, p = .001, Cramer's V = .60. On the other hand, there was no significant difference between the number of children who did contrastive test and who did positive tests, Fisher exact test, N = 51, one-sided, p = .109.

Table 4.2

Proportion and Percentages of One-Variable Hypothesis Testing in Phase 3

	Hypothesis Testing Patterns		
Age Group	**Contrastive test**	**Positive test**	**Untargeted exploration**
Younger Group	47% (14/30)	20% (6/30)	33% (10/30)
Older Group	47% (15/32)	50% (16/32)	3% (1/32)
Total	47% (29/62)	35% (22/62)	18% (11/62)

Counterargumentation. Since children's belief states might influence their counterargumentation responses, children who did not pass the belief check (n = 15; one 3-year-old, nine 4-year-olds, two 5-year-olds, three six-year-olds) were excluded and this resulted in a total sample size of 47 children in the counterargumentation analyses. The order of the heavy and blue counterargumentation did not affect children's argumentation performance.

Heavy counterargumentation. Sixty-eight percent (32 out of 47) of the children provided spontaneous verbal counterarguments after the experimenter's false hypothesis. Thirty-two percent (15 out of 47) of the children either did not provide any verbal counterargument or provided an incorrect one. Twenty-three percent (11 out of 47) of the children referred to the weight feature, 21% (10 out of 47) referred to the sticker feature, and 23% (11 out of 47) referred to both the sticker and the weight

features in their verbal counterarguments. There was no significant difference between the two age groups in terms of the number of children providing verbal counterarguments (Fisher's exact test, one-sided, p = .39). Table 4.3 shows the descriptive results for content of children's verbal counterarguments.

Table 4.3

Proportion and Percentages of Children's Verbal Counter Arguments in the Heavy Counter Argumentation Phase by Age Group and Content

Age Group	Incorrect/Absent	Color	Sticker	Both
Younger Group	45% (9/20)	15% (3/20)	25% (5/20)	15% (3/20)
Older Group	22% (6/27)	30% (8/27)	18% (5/27)	30% (8/27)
Total	32% (15/47)	23% (11/47)	21% (10/47)	23% (11/47)

Children's evidence generation performance was investigated by comparing the frequencies of the disconfirming and confirming evidence they put on the light box after the experimenter's false hypothesis. An Age Group (younger, older) x Type of Evidence (disconfirming, confirming) mixed-design ANOVA with Type of Evidence as the within-subject variable and the number of objects interacted as the dependent variable revealed a significant main effect Type of Evidence, $F(1,45) = 18.94$, $p < .001$, $\eta_p^2 = .30$[12]. Children interacted with more objects that disconfirmed the false hypothesis of the experimenter (M = 0.68 out of 2, SD = .66) than objects which confirmed the

[12] The data was not normally distributed. In order to inspect that the significant findings are robust, we additionally ran nonparametric analyses. An Age Group (younger, older) x Type of Evidence (disconfirming, confirming) mixed-design ATS test, Type of Evidence as the within-subject variable and the number of objects interacted as the dependent variable revealed a significant main effect Type of Evidence, $F(1, \infty) = 16.57$, $p < .001$, relative treatment effects (RTE) suggested a small effect size (see Figure 1 for the relative treatment effects). Children interacted with more objects that disconfirmed the false hypothesis of the experimenter than objects which confirmed the experimenter's false hypothesis. The main effect of age (p = .938) and the interaction effect (p = .436) were not significant.

experimenter's false hypothesis (M = 0.26 out of 2, SD = .44). The main effect of age (p = .686) and the interaction effect (p = .820) were not significant.

We also investigated children's individual counterargumentation performance in the heavy counterargumentation task. Thirty percent (14 out of 47) of the children showed no competence, 34% (16 out of 47) of the children showed partial competence, and 36% (17 out of 47) of the children showed full competence. Table 4.4 shows the frequency of children in each competence level. We also investigated children's individual counter argumentation performance to see whether there were any developmental differences in children's competence in the counter argumentation tasks. A Mann-Whitney U test revealed no significant differences between the younger (M_{Rank} = 19.80) and the older group (M_{Rank} = 27.11) in terms of their heavy counter argumentation performance, U = 354, z = 1.92, p = .055 using an exact sampling distribution for U.

Table 4.4

Proportion and Percentage of the Competency Level in Heavy Counterargumentation Phase by Age Group

Age Group	No Competence	Partial Competence	Full Competence
Younger Group	40% (8/20)	40% (8/20)	20% (4/20)
Older Group	22% (6/27)	30% (8/27)	48% (13/27)
Total	30% (14/47)	34% (16/47)	36% (17/47)

Blue counterargumentation. Sixty-eight percent (32 out of 47) of the children provided a verbal counterargument and 32% (15 out of 47) of the children either did not provide any or provided incorrect verbal comments. In their verbal counterarguments, 17% (8 out of 47) children referred to the color feature, 19% (9 out of 47) of the children referred to the sticker feature, and 32% (15 out of 47) of the children referred to both the

sticker and the color features. There was no significant difference between the frequency of younger children (55%, 13 out of 20) and the frequency of the older children (70%, 19 out of 27) in terms of providing verbal counter arguments (Fisher's exact test, p = .468). Table 4.5 shows the proportion and percentages of the content of children's counterarguments in the blue counterargumentation for the younger and the older group separately.

Table 4.5

Proportion and Percentages of Children's Verbal Comments for False Color Hypothesis by Age Group

Age Group	Incorrect/Absent	Color	Sticker	Both
Younger Group	35% (7/20)	25% (5/20)	10% (2/20)	30% (6/20)
Older Group	30% (8/27)	11% (3/27)	26% (7/27)	33% (9/27)
Total	32% (15/47)	17% (8/47)	19% (9/47)	32% (15/47)

To test the hypothesis whether children interacted with more disconfirming evidence than confirming evidence, an Age Group (younger, older) x Type of Evidence (disconfirming, confirming) mixed design ANOVA with Type of Evidence as the within-subject variable and number of objects put on the detector as the dependent variable was conducted. There was no main effect of Type of Evidence (p = .785), children equally engaged with disconfirming (M = 0.91 out of 2, SD = .72) and confirming objects (M = 0.87 out of 2, SD = .90); and no difference (p = .581) between younger group (M = 0.83 out of 2, SE = .16) and older group (M = 0.94 out of 2, SE = .14) and no interaction of Type of Evidence and Age Group (p = .473)[13].

[13] The data was not normally distributed therefore we repeated the results with nonparametric test.. To test the hypothesis whether children interacted with more disconfirming evidence than confirming evidence an Age Group (younger, older) x Type of Evidence (disconfirming, confirming) mixed design ATS, Type of Evidence as the within-subject variable and number of objects put on the

Children's individual competence scores were calculated. Thirty percent (14 out of 47) of the children showed no competence, 53% (25 out of 47) of the children showed partial competence and 17% (8 out of 47) of the children showed full competence. Table 4.6 shows the percentages of the individual competency level in the blue counterargumentation for the two age groups. A Mann-Whitney U Test was conducted to examine the differences between the two age groups in terms of their blue counter argumentation performance. No significant difference was found between the younger (M_{Rank} = 24.30) and the older group (M_{Rank} = 23.78), U = 264, z = -.143, p = .887).

Table 4.6

Proportion and Percentage of the Competency Level in Blue Counter Argumentation Phase by Age Group

Age Group	No Competency	Partial Competency	Full Competency
Younger Group	7/20 (35%)	8/20 (40%)	5/20 (25%)
Older Group	7/27 (26%)	17/27 (63%)	3/27 (11%)
Total	14/47 (30%)	25/47 (53%)	8/47 (17%)

We also computed an aggregated score across the two counterargumentation tasks by liberal criteria: children who showed at least partial competence (i.e., provided valid disconfirming evidence or/and verbal counterargument) in at least one of the tasks were coded as partially competent and children who showed no competence in either one of the tasks were considered as incompetent. Eighty-three percent (39 out of 47) of the children were partially competent by these criteria.

detector as the dependent variable was conducted; yet, no significant main effects for Age Group (p = .623) and Type of Evidence (p = .530); or interaction effect (p = .410) were found.

Finally, we investigated whether there was an association between children's hypothesis testing performance (contrastive test = 3, positive test = 2, untargeted exploration =1) and their competence in each of the two counterargumentation tasks (full competence = 3, partial competence = 2, no competence = 1). A marginally significant association was found between performance in hypothesis testing and heavy counterargumentation, Somers' d = .303, p = .037 (See Table 4.7) and no significant association between hypothesis testing and blue counterargumentation performance, Somers' d = .072, p = .629 (See Table 3).

Table 4.7

Proportion and Percentages of Children's Individual Level of Competence in Heavy Counterargumentation by Hypothesis Testing Pattern

Hypothesis testing	**Heavy counterargumentation**		
	No competence	Partial competence	Full competence
Contrastive test	16% (4/24)	42% (10/24)	42% (10/24)
Positive test	32% (6/19)	32% (6/19)	37% (7/19)
Untargeted exploration	100% (4/4)	0% (0/4)	0% (0/4)

Table 4.8

Proportion and Percentages of Children's Individual Level of Competence in Blue Counterargumentation by Hypothesis Testing Pattern

Hypothesis testing	**Blue counterargumentation**		
	No competence	Partial competence	Full competence
Contrastive test	33% (8/24)	46% (11/24)	21% (5/24)
Positive test	16% (3/19)	68% (13/19)	16% (3/19)
Untargeted exploration	75% (3/4)	25% (1/4)	0% (0/4)

4.4 Discussion

Study 2 investigated two components of scientific reasoning in early childhood: hypothesis testing and evidence-based argumentation skills. The findings indicate that there is an emerging competence in early childhood in understanding the epistemic relation between claims and evidence and making explicit connections between the two.

With respect to hypothesis testing, there was evidence for both a beginning competence and development in appropriate strategic behaviors. When children had a directed hypothesis about the cause of a light effect, more than two-thirds of the children showed systematic exploratory behaviors. About half of the children followed a contrastive test strategy by comparing cases with and without the hypothesized variable (i.e., sticker on the bottom). The other half of the children in the older group and 20% of the children in the younger group pursued a positive test strategy where they only tested the objects with the hypothesized causal variable. Different from the older group, approximately one third of the children in the younger group did untargeted exploration without a clear strategy. Thus, strategic behaviors that are functional in either testing positive cases or discriminating between the conditions of the hypothesized variable emerge over the preschool years. While contrastive testing clearly serves an epistemic goal, it is less clear whether children who did only positive tests followed an epistemic goal. Still, a positive test strategy may be functional in starting an exploratory process; and it can even be a more informative strategy in other task formats (e.g., probabilistic evidence) (Klayman & Ha, 1987). Note that children who adopted a contrastive test strategy did not differ from children who used a positive test strategy in their counterargumentation performance, while children who did not follow a systematic test strategy clearly performed worse than the other groups, a finding that makes it unlikely

that the contrastive test group was superior to the positive test group in their understanding of the epistemic goals of the task.

The present findings go beyond previous studies of hypothesis testing behaviors in young children's exploratory play (Bonawitz et al., 2012, Cook et al., 2011; Legare, 2012, Schulz & Bonawitz, 2007) as it demonstrates for the first time that young children not only show variable and longer exploratory play when motivated by a causal belief (Legare, 2012), but they also follow systematic testing strategies (i.e., contrastive testing, positive testing). Legare (2012) found that children engaged more in exploratory activities when they had formed a causal belief about the critical phenomena. However, this link between explanation and exploration could be due to individual differences in more general cognitive abilities. Importantly, in the present study, all children were led to form the same causal belief and their causal understanding was controlled for. Under these conditions, more than two-thirds of the children showed systematic testing strategies. These findings indicate that, when given a causal hypothesis, young children clearly pursued a systematic strategy which is consistent with Legare`s (2012) claimed link between explanation and exploration.

Despite the clear evidence for systematic strategic behaviors, the testing strategies that the children used were limited. We did not find evidence for two-variable testing (i.e., for isolating the two variables: sticker on the bottom and heaviness). We hypothesized that a preference for unambiguous objects (i.e., a heavy object without a sticker, a light object with a sticker) would indicate that children understand that unambiguous objects are informative about the individual effects of the two variables in isolation. Yet, children did not preferentially interact with unambiguous objects more than ambiguous objects. However, this does not necessarily indicate that children failed to recognize the difference between ambiguous and unambiguous test objects. The

preference for disconfirming objects in the heavy counterargumentation phase clearly demonstrates that children can differentiate between the ambiguous/confirming and unambiguous/disconfirming cases.

The reason why children did not interact with the unambiguous objects than the ambiguous objects might be due to several reasons. Firstly, even though the evidence pattern is very similar, hypothesis testing and counterargumentation tasks differ from each other in terms of the goal and uncertainty of the tasks. In the hypothesis testing, the truth of the sticker hypothesis was uncertain. Children had to generate evidence in order to evaluate the truth of the hypothesis. Hypothesis testing requires thinking about several alternative future possibilities. To illustrate, one should be able to imagine what it would mean if a light object with sticker turns the machine on or if it does not turn the machine on. On the other hand, in the counterargumentation there is no uncertainty from children's perspective regarding the truth of the claim. This makes choosing correct evidence pieces in the hypothesis testing harder than the counterargumentation. A second related reason is that although the ambiguous objects are not as informative as the unambiguous objects, trying them might still provide information. For instance, if a heavy object with sticker activates the box, this would be contradictory to the sticker hypothesis. Therefore, trying ambiguous objects is also informative. Thirdly, the result might be due to the exploratory nature of the task. In this task, children were free to explore the objects as they like. We did not limit their choices. One hypothesis would be that if children were restricted in their choices they might be more conservative and only try the unambiguous objects[14]. Lastly, at the beginning of the hypothesis testing phase, the likelihood of the two hypotheses (weight and sticker) were not equal since children

[14] This might be a possible explanation for the children who interacted with 3 or more objects. However, this was not the case, at least for some of the children, because 35% of the children did only test the objects with sticker (2 objects out of 4) and did not test the objects without sticker. Therefore, even though we did not limit them, some of the children limited their exploration only to the positive cases.

first saw consistent and then inconsistent evidence for the weight hypothesis. On the other hand, children did not see any evidence pro or against the sticker hypothesis. It is possible that children entirely abandoned the weight hypotheses and only took into account the sticker hypothesis. Presenting children two equal hypotheses and then looking into whether they employ two-variable hypothesis testing would shed light into this possibility. Considering that all these factors may play a role in the poor performance in hypothesis testing, further studies are essential to reach a conclusion regarding this matter.

The finding that young children could counterargue to a false causal claim by providing both valid verbal claims and generating valid evidence to refute the false claim clearly indicates their ability to differentiate claims from evidence and to coordinate the two; an ability that has not been previously shown in this age group. The *protest paradigm* proved to be a useful method for eliciting young children's evidence-based reasoning. Approximately 60% of children provided valid verbal counterarguments both in the heavy and blue counterargumentation tasks and this indicates an explicit understanding of how evidence is related to hypotheses. With respect to evidence generation, children showed competence only in one of the two counterargumentation tasks (the heavy task) with selectively generating more disconfirming than confirming evidence. This difference in evidence generation between the heavy and the blue counterargumentation may be because it is easier for children to counterargue a belief from which they had previously experienced recovery themselves; alternatively, it may also be easier to refute a more plausible than an implausible false claim. Taken together, the findings for the verbal and the nonverbal reasoning clearly indicate ability for explicit and valid evidence-based counterargumentation in early childhood. Thus, the child as scientist metaphor (Gopnik, 2012) appears to be

appropriate not only for describing theory formation through causal learning, but also theory evaluation through scientific reasoning in young children.

This novel finding is surprising given deficits demonstrated in argumentation, even in adults, who appear to form a script-like representation of the way things are, rather than arguing from evidence (Hahn & Oaksfard, 2012; Kuhn, 1991). The differences in difficulty between the tasks used in previous research on scientific argumentation and the present one can be attributed to several factors. Generally, the scientific reasoning literature with older children and adults has used the task to argue for a claim that the participants believe to be true, rather than against one they know to be false. Moreover, the theories in those studies are often complex and the access to evidence is difficult, rather than being laid out in front of the participants in a simple causal learning task.

A key difference between previous research and this study appears to be that we used a deterministic, rather than a probabilistic relation of theory and evidence (Kuhn, 1991; Kuhn & Udell, 2003). In a deterministic environment, one instance of counterevidence is sufficient to refute a claim which is not the case for tasks with probabilistic evidence. Thus, it might be that children's and adults' bad performance in evidence-based argumentation is not a fundamental inability to recognize hypotheses and evidence as different epistemic categories; rather, it may be due to a limited understanding of probabilistic evidence. Although recent research (see Gopnik & Wellman, 2012 for a review) demonstrates that young children are good at learning from probabilistic evidence and making generalizations, these studies did not look into children's explicit understanding of the probabilistic evidence. Future research is needed to compare children's performance in deterministic and probabilistic tasks of the same format.

Study 2 provides mixed findings with respect to developmental differences taking place in scientific reasoning skills. In the hypothesis testing task, older preschoolers performed better than younger preschoolers. This difference was due to the fact that almost all children who did untargeted exploration were from the younger group. This finding suggests that preschoolers who are younger than 5-years of age might have some problems in following testing strategies, while almost all older preschoolers either followed a positive or contrastive test strategy. On the other hand, we did not find any differences between younger and older group in terms of evidence generation (heavy counterargumentation) and verbal comments in the counterargumentation tasks. It is important to note that the majority of the children who were excluded before the counterargumentation phase were from the younger group. Therefore, when we only looked at the children who did form the required belief, there were no age differences in evidence generation or verbal counterargumentation. The question is whether the differences in developmental patterns we found in the two tasks are due to different developmental patterns arising from the very nature of hypothesis testing and counterargumentation skills or due to the exclusion of the younger group that might have resulted in a misrepresentation of the actual differences between younger and older children in the counterargumentation task. In other words, in a hypothetical scenario, if we only measured children's counterargumentation abilities, how would children who were excluded due to their belief formation counterargumentation?

Our following counterargumentation study (Köksal Tuncer, Sodian, & Saffran, 2017) can address this question. In this new study we encourage 4-year-olds to form a belief (sticker matters), then present a false causal claim (size matters). Therefore, children are only expected to form a belief and not expected to revise a prior belief as in Study 2. The results demonstrated that almost all 4-year-olds formed the sticker belief

and they were able to present evidence that refutes the false claim. Therefore, 4-year-olds are able to counterargue which suggest that the difference in the developmental pattern we found between hypothesis testing and counterargumentation tasks in Study 2 is arising from the very nature of the task demands themselves.

The correlation between hypothesis testing and counterargumentation competence was weak which may be because children's competencies were masked in spontaneous exploration task due to information processing demands. Further research needs to address hypothesis testing strategies across a wider range of task conditions.

The exploratory nature of the hypothesis testing and counterargumentation tasks prevents us from making conclusions regarding incompetence when there was no selective preference for unconfounded pieces of evidence. This is because we did not restrict children in their interactions with the objects and interacting with all of the objects is informative, too. As a result, it is unknown whether children would show selectivity for unconfounded evidence in the hypothesis testing and blue counterargumentation if we restricted them in the number of objects they chose. This is especially unclear for the children who interacted with three or four objects (13 children in the hypothesis testing and 18 children in the blue counterargumentation). In order to know whether children would selectively choose the unconfounded evidence if they were given limited choice options, further studies may restrict children in the number of objects they choose, and then look into whether there is a selective preference for the unconfounded objects.

Two different individuals presented the heavy and the blue false claims in the two counterargumentation tasks and this prevents us from comparing performances between the two tasks and making conclusions on whether children's performance difference in the two tasks is due to familiarity or novelty of the false claim. The

blue/novel claim was presented by the main experimenter who engaged with children during whole testing whereas the heavy/familiar claim was presented by the second experimenter who was sitting behind the children without participating in the earlier phases of the testing. There are several findings on how young children and even infants selectively share information based on the knowledge states of agents (e.g., Dunham, Dunham, & O'Keefe, 2000; Liszkowski et al., 2008; Moore & D'Entremont, 2001). Thus, children in the present study might be less likely to selectively generate disconfirming evidence when the false claim was presented by the knowledgeable first experimenter because they might have thought that there was a rationale behind the false claim. This might have led them to question their belief and interact with all of the evidence again (both confirming and disconfirming). This is in line with the present findings since children interacted with significantly more objects in the blue counterargumentation compared to the heavy counterargumentation. On the other hand, in the heavy counterargumentation the children selectively generated more disconfirming evidence since they may not need to question the truth of their own belief in response to the false claim of an ignorant agent. It is unknown whether the different evidence generation performance in the blue and heavy counterargumentation tasks is due to agents' different knowledge states or due to the familiarity and the novelty of the two claims. In order to answer this question, future studies are necessary in which agents with same knowledge states present novel and familiar claims.

In sum, the present results confirm the view that scientific reasoning competencies are present as early as early childhood (Bonawitz et al., 2012; Koerber et al., 2005; Legare, 2012; Ruffman et al., 1993). While previous causal reasoning studies have used indirect and implicit measures, the present study demonstrated explicit verbal argumentation competence in a scientific reasoning task that required children to refute

a false claim. Moreover, children showed spontaneous strategic behaviors when evaluating a specific causal hypothesis. These early competencies may not generalize to other domains and task formats. Yet, the results of the present study show that when task demands are kept low, young children can skillfully differentiate and coordinate claims and evidence.

5 Study 3: Young Children's Understanding of Evidence as an Epistemic Category

One of the foundational metacognitive abilities for scientific reasoning is the ability to reflect on the empirical relation between hypotheses and evidence. Children`s selective exploration in the case of ambiguous and inconsistent evidence (e.g., Legare, 2012; Schulz & Bonawitz, 2007) or their tendency to isolate variables in response to confounded evidence (Cook et al., 2011) suggests that they have some form of implicit understanding of the relation between epistemic states and evidence. Studies investigating children's implicit metacognition of knowledge and ignorance also support the presence of such implicit metacognitive skills in the early years of life (Bartz, Rowe, & Harris, 2017, as cited in Harris et al., 2017; Harris et al., 2017; Liszkowski et al., 2008; Kim et al., 2016). Although these implicit skills are critical, mature scientific reasoning necessitates a metacognitive understanding of how knowledge is constructed as a function of evidence. Yet, little is known about young children's abilities to reflect on their epistemic states and their metalevel understanding of the relation between evidence and their knowledge states.

Studies on evidence evaluation often require participants to give verbal judgments. Although their performance is far from perfect, elementary school children provide verbal judgments and justifications for the hypothesis–evidence relation (e.g., Saffran et al., 2016; Saffran et al., submitted; Sodian et al., 1991). However, only a few studies on preschoolers` evidence evaluation investigated preschoolers` judgments (Koerber et al., 2005; Piekny & Maehler, 2013) and even less examined preschoolers` justifications for the hypothesis–evidence relation (Ruffman et al., 1993; Saffran et al., 2017). Study 2 of this thesis yielded evidence on preschoolers' basic ability to reflect on

the epistemic relation between beliefs and evidence shown by their verbal counterarguments in response to false causal claims showing that approximately 70% of the 4- to 6-year-old preschoolers provided evidence-based counterarguments. These studies document that preschoolers can indeed provide explicit judgments; yet although some of them provide relevant evidence-based justifications, the frequency of the children who provided elaborated justifications was low.

Two points are worth noting with respect to the studies on preschoolers evidence evaluation. Firstly, both Study 2 of this thesis and Ruffman et al. (1993) required children to provide reflections with respect to another person`s beliefs. In Study 2, children provided judgments about the veracity of an experimenter`s belief and justifications for why the experimenter`s belief was wrong; and in Ruffman et al., children provided judgments about the false belief of a protagonist and justifications why the protagonist would have a false belief. In this respect, there may be differences between reflecting on beliefs when those are another agent's (false) beliefs and reflecting on one`s own beliefs. Secondly, and more importantly, none of the studies (Ruffman et al., 1993; Saffran et al., 2017; Study 2 of this thesis) investigated children`s evidence evaluation in the case of uncertainty. The concept of hypothesis necessitates an understanding of uncertainty. Therefore, children need the mental capacities in order to be able to reason in the case of uncertainty, and this is critical for the development of mature forms of scientific reasoning.

Studies on diagnostic reasoning suggest that understanding causal uncertainty is developing between the preschool years and around 5 years children can make correct predictions in the case of uncertainty (Erb & Sobel, 2014; Sobel et al., 2017, see Section 2.3.2.2 for study details). Rohwer et al. (2012, see Section 2.4 for study details) demonstrated that while children older than 5 years report ignorance in the case of

partial information, children younger than 5 years do not have an awareness of their ignorance due to partial evidence. When 3- to 7-year-olds were shown two objects, and later the experimenter put one of the objects in a closed box without showing the child which of the objects it was, children younger than 5 years claimed that they knew what was in the box although they could not know because they did not see. These results suggest that the ability to evaluate one`s own epistemic states in the case of uncertainty emerges around 5 years. Taken together, the findings demonstrate that around 5 years, there is an emerging understanding of causal uncertainty: children are aware of their ignorance as a result of partial information (Rohwer et al., 2012), and they can diagnostically reason in the case of uncertainty. Moreover, they are sensitive to confounded nature of evidence shown by causal reasoning studies (Cook et al., 2011; Schulz & Bonawitz, 2007). Based on these findings, the goal of Study 3 was to investigate the development of the metacognitive abilities, particularly the ability to reflect on the epistemic states and the relation between epistemic states and evidence. In order to investigate this, Study 3 employed a traditional measure frequently used in metacognition and scientific reasoning research: children's explicit judgments for their epistemic states and their evidence-based justifications for their knowledge/ignorance judgments.

Cook et al.'s (2011, see Section 2.3.1.3 for study details) paradigm, in which children were presented with two attached beads (confounded evidence) activating the blicket detector was suitable for our study purposes since it has already been found that preschoolers attempt to isolate beads when prior evidence suggested that the evidence is confounded. We familiarized children with baseline evidence suggesting that some of the objects turned the light box on while some others did not. Relying on the evidence that children tend to make category membership decisions based on the causal functions

of objects (Gopnik & Sobel, 2000; Legare, 2012; Legare et al., 2010), the objects that activated the light were labelled as toma (or baffe) and the objects that did not activate light were labelled as not-a-toma (not-a-baffe). Later, in the Confounded Condition, children were presented with evidence that two novel objects were placed on the light box and they activated the light. We asked children whether they knew that one of the objects (target object) is a toma or not-a-toma. In contrast, in the Unconfounded Condition, children were presented that the target object put on the light box in isolation and the light was activated. We investigated children's awareness of their epistemic states via their knowledge judgments and their understanding of the relation between epistemic states and evidence via their evidence-based justifications. If children have a reflective awareness of the relation between their epistemic states and evidence, they would selectively report ignorance in the Confounded Condition and knowledge in the Unconfounded Condition.

Children's responses to interview questions may be considerably influenced by the way that questions are asked (e.g., Fritzley, 2006). Therefore, we formulated the questions carefully and asked follow-up questions to make sure that children clearly understood and openly responded. Even adults are reluctant to admit ignorance (e.g., Bishop, Tuchfarber, & Oldendick, 1986) and one possibility is that young children may be less likely to admit their ignorance in response to question "Do you know or do you not know?"[15] because *knowing* is usually socially desirable while *not knowing* is not. In order to prevent responses motivated from social desirability, we formulated our experimental question as, e.g., "Do you know whether the blue cube is a toma or not-a-toma, or do you need to know more?" (adapted from Busch & Legare, 2016). In this question format, there was no big difference between stating knowledge and stating

[15] Children were asked "Do you know or do you not know" questions in the familiarization trials.

ignorance in terms of social desirability. Another possibility was children interpreting our questions as part of a guessing game. Guessing games are frequently played in early childhood, and children of this age already display advanced reasoning skills in these games (Fernie & DeVries, 1990). If children interpret the task as a guessing game, they would reply that the object is a toma or not-a-toma, even though they may have the understanding that they were actually ignorant. In this case, we would not be able to differentiate children who truly think that they know from children who were aware that they do not know but just make a guess due to interpreting the task as a guessing game. In order to prevent this, we asked children whether they really knew or whether they were just guessing after the knowledge judgments.

In addition to knowledge or ignorance judgments, we asked children several questions to gain further knowledge about their metacognitive understanding of the relation between evidence and their epistemic states. Firstly, we asked children to justify their knowledge statements (knowledge vs. need to know more) in order to investigate children's awareness of how evidence bears on their epistemic states. We looked at whether children would provide any evidence-based justifications by referring to evidence characteristics. Children's justifications in the Confounded Condition were our special interest because these justifications particularly show that children understand (a) why confounded evidence is *confounded* in the first place and, (b) how confounded evidence leads to ignorance from their point of view. Secondly, we were interested in whether children explicitly describe the correct test which would yield necessary information in the case of confounded evidence. Therefore, if they reported beforehand that they need to know more, we asked what they would do to know more. The correct answer to this was to suggest isolating the objects—a skill that preschoolers were already found to have behavioral competence for (Cook et al., 2011).

Earlier studies demonstrated that children prefer playing with a familiar object than a novel object when the causal relations of the familiar object are not clear to them (Schulz & Bonawitz, 2007). In the present study, we also investigated preschoolers' behavioral tendencies to prefer causally ambiguous objects. After each condition (Confounded & Unconfounded), we presented two objects; the target object and a novel object, and instructed children that they could choose one of the objects and try it. In the two conditions, the only difference was children's knowledge about the efficacy of the target object. In the Unconfounded Condition, children had already learned the efficacy of the target object while in the Confounded Condition they had not. If children chose the target object more often in the Confounded Condition than in the Unconfounded Condition, this would generate evidence to support that they have implicit sensitivity for uninformative nature of confounded evidence.

Overall, Study 3 aimed to investigate preschoolers' abilities to reflect on their own epistemic states and on the connection between their epistemic states (ignorance vs. knowledge) and evidence (confounded vs. unconfounded). In the case of confounded evidence, there is uncertainty; therefore, there are two potential hypotheses regarding the efficacy of an object. Especially if children would provide justifications referring to the confounded nature of evidence, this would be evidence for their ability to reflect on the potential alternative hypotheses (e.g., the blue cube may or may not be a toma.) due to confounded evidence.

5.1 Research Questions of Study 3

1- Do 5- and 6-year-olds request for information in the Confounded Condition more often than in the Unconfounded Condition?

2- (a) Can young children provide evidence-based justifications for their ignorance in response to confounded evidence? (b) At which frequency do children provide evidence-based justifications?

3- (a) Can young children provide explicit verbal instructions for an informative test?

(b) At which frequency do children who provide explicit verbal instructions for the correct test?

4- Do 5- and 6-year-olds choose the causally ambiguous target object in the Confounded Condition more often than the causally unambiguous target object in the Unconfounded Condition?

5.2 Method

5.2.1 Participants

Sixty 5- and 6-year-old preschoolers (31 females, M_{age} = 70 months; range: 60 months-81 months) participated in the study. All children were typically developing children of lower- to upper-middle class background from a larger German city. Parents signed a written consent for their child`s participation and children were asked for their verbal consent before the study.

5.2.2 Materials

A light box which was a 30 cm x 20 cm x 14 cm custom built wooden box with a LED light strip attached around it was used. The light box had an RFID (radio-frequency identification) reader placed inside, and it was automatically activated when objects with RFID chips were put on the center of the top plate. 3 cm x 3 cm x 3 cm cubes in different colors were used as objects. Originally RFID chips were 2.5 cm

diameter white stickers. They were attached to one side of the activator cubes, were colored with the same color of the cube and could be only identified upon close inspection. To control for the possibility that children might realize the RFID chips on the activator objects, exact looking stickers were attached to the nonactivator cubes. As a result, the activator and nonactivator cubes were perceptually identical. None of the children realized the RFID chips/stickers during the study. In total, 20 cubes in 10 different colors (black, white, blue, pink, green, orange, gray, red, yellow, lilac) were used. Each color had one activator and one nonactivator version. The colors of the cubes used in each phase of the study and the effects of the colors were counterbalanced. Each participant was presented with 10 individual cubes in 10 different colors. Novel labels were used for the activator and nonactivator cubes. Half of the children were told that activators were called tomas and nonactivators were called not-tomas; whereas the other half of the children were told that activators were called baffes, and nonactivators were called not-baffes. A 17 cm x 9 cm tray from cardboard covered with cloth was used to put the cubes on it, and a piece of cloth 22 cm x 12 cm was used to cover the cubes in the forced-choice phase. The experimenter used a 22 cm x 13.5 cm x 6 cm box from cardboard with several compartments and signs to organize the presentation of the cubes during the study.

5.2.3 Design

Two within-subjects conditions were designed in order to investigate children's responses to different evidence characteristics. Each child participated in the two conditions, with the order counterbalanced across participants. In the confounded condition, children were presented with a target object simultaneously placed on the light box together with another object, whereas in the unconfounded condition, children were presented with the isolated effect of the target object. A familiarization trial

preceded each experimental trial in which children were shown confounded evidence consisting of the target object. This phase helped children to get familiarized with the target object in each condition and yielded more chances for us to observe children's knowledge judgments and justifications in the case of confounded evidence. A forced-choice phase followed each of the experimental phases to investigate children's choice of object (target vs. novel) in the two conditions. Figure 5.1 is the schematic display of the main procedure.

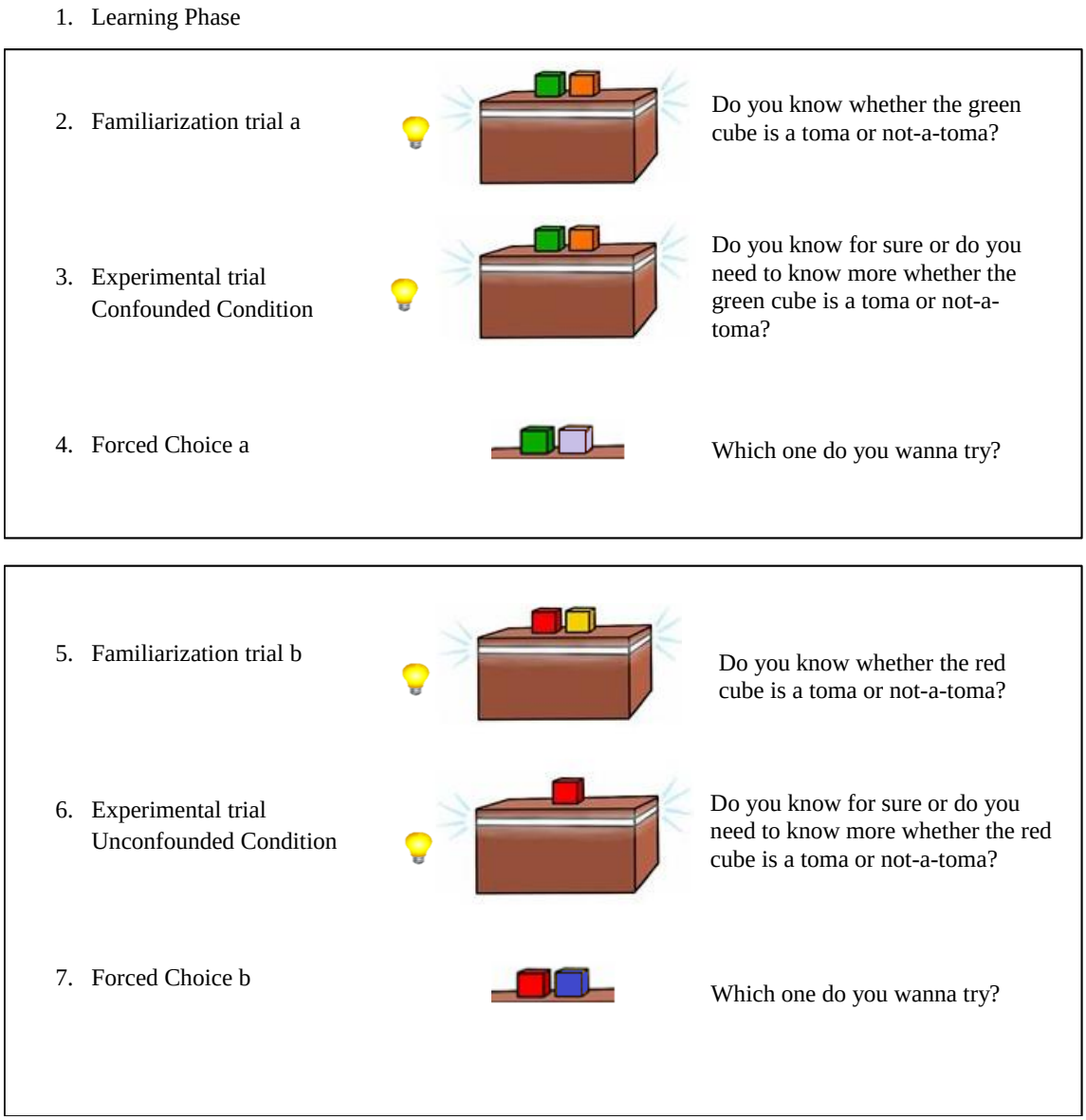

Figure 5.1. Schematic display of the main procedure of Study 3. Yellow bulbs represent that the box lights up.

5.2.4 Procedure

All sessions were carried out in separate rooms of kindergartens and recorded by a video camera. Each child was tested individually in a session lasting approximately 15 minutes. Children were seated across from the experimenter. At the beginning, the experimenter and the child played a warm-up game together (a puzzle matching the animals with their habitats). In all phases of the study, children never interacted with the

objects themselves but observed the experimenter interacting with the objects and the light box. The study consisted of four critical phases: the learning phase, the familiarization phase (two times), experimental phases (confounded vs. unconfounded evidence), and forced-choice phase (two times).

Learning phase. The aims of this phase were to, firstly, familiarize children with the materials and their effects; secondly, to teach them novel category labels for the activator (e.g., a toma) and nonactivator (e.g., not-a-toma) cubes; and thirdly, to demonstrate the effects of the cube pairs when they were placed on the box together. The experimenter placed a cube on the box, the box was activated and the experimenter said "It makes the box light up. The cubes that make the box light up are called toma." Experimenter asked children to repeat the novel word and then placed the cube on the right-side of the light box saying "Let's put tomas here." Subsequently, a cube in a different color was placed on the box and the box was not activated. The experimenter said "The cubes that don't make the box light up are called not-a-toma," similarly as with the first cube, the experimenter asked children to repeat the label and said "Let's put not-tomas here" and put the cube on the left-side of the box. The same procedure was repeated with two novel cubes, one activator and one nonactivator. Subsequently, children were presented with pairs of cubes placed on the light box together. In this phase, experimenter first placed one of the activators alone on the box and asked for the label, then placed one of the nonactivators and asked for the label, and then placed both on them on the box together simultaneously. This procedure was repeated two times: one time with two activators and one time with two nonactivators (See Figure 5.2 for the presentation order of the cubes and their effects in the learning phase).

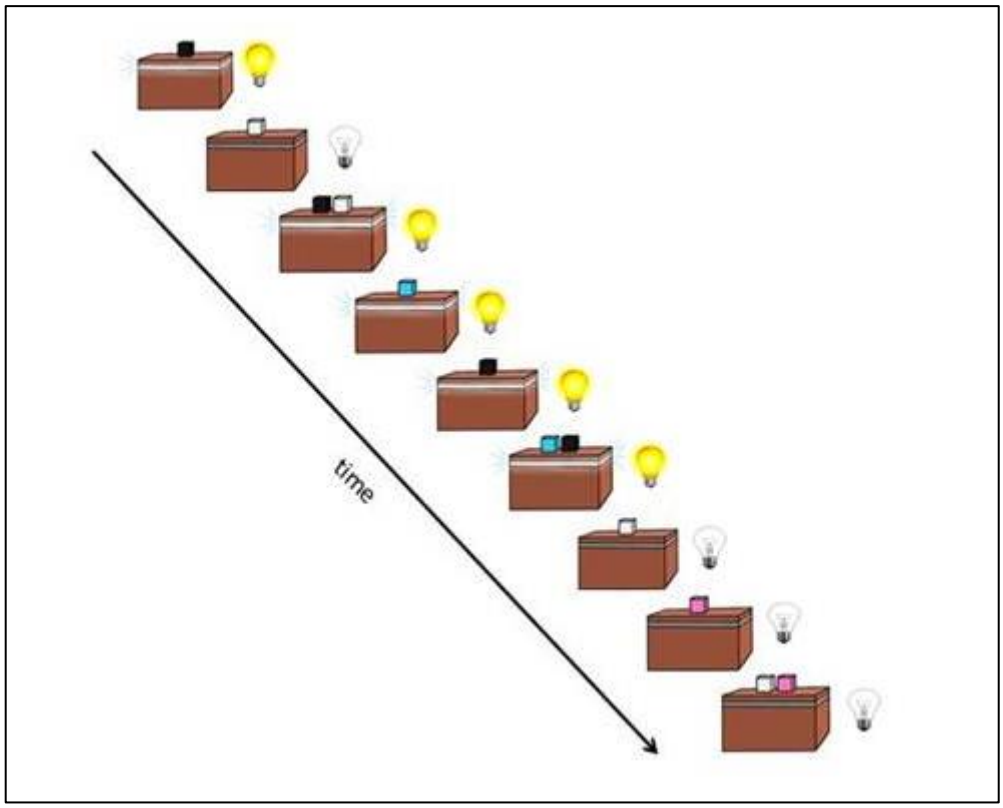

Figure 5.2. The presentation order of the cubes and their effects in the learning phase

Familiarization trials. In this phase, all children were presented with confounded evidence and were asked whether they know that a novel (target) cube was an activator or a nonactivator. The experimenter brought two novel cubes and said, e.g., "I want to find out whether this green cube is a toma or not-a-toma. Let's try." Then, the experimenter simultaneously placed the two cubes on the box together, and the light turned on. The experimenter asked, "Do you know for sure whether the green cube is a toma or not-a-toma?" Depending on their answer, children were asked different sequences of follow-up questions (See Figure 5.3 for the questions). When children claimed that they knew, the experimenter asked a "know vs. guess" question, i.e., "Do you really know are you just guessing?" to differentiate children who guessed and who stated that they really know. After children gave an answer, the experimenter asked "Why?" in order to receive their justifications for their claim. If children gave an unclear

response, the experimenter repeated the question. In the case of another unclear response, the experimenter asked "What do you mean?". In the cases where children remained silent, the experimenter repeated the question one more time.

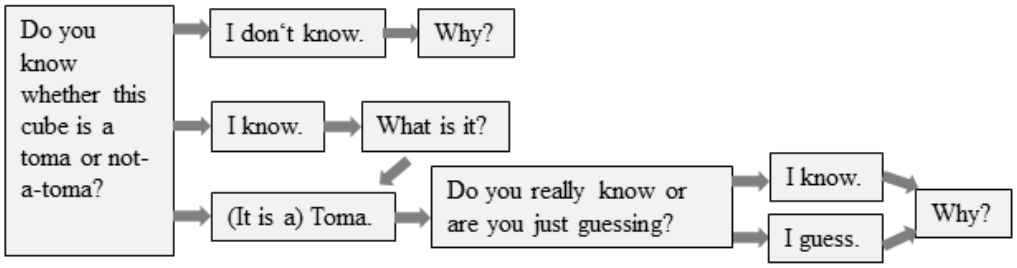

Figure 5.3. Questions in the familiarization trials in response to children's answers

Experimental phase. The experimenter said, "I want to find out whether the green cube is a toma or not-a-toma". The confounded and unconfounded experimental conditions were identical except the evidence presented to children. In the confounded condition, the experimenter placed the same two cubes presented in the familiarization phase simultaneously on the light box and the light turned on. In the unconfounded condition, the experimenter placed the target cube on the box alone and the light turned on. The experimenter asked, "Do you know for sure whether the green cube is a toma or not-a-toma, or do you need to know more about it?" Depending on children's response, children were asked several follow-up questions (See Figure 5.4 for the follow-up questions). If children gave an unclear response, the experimenter repeated the question. In the case of another unclear response, experimenter asked, "What do you mean?". In the case when children remained silent, the experimenter repeated the question one more time.

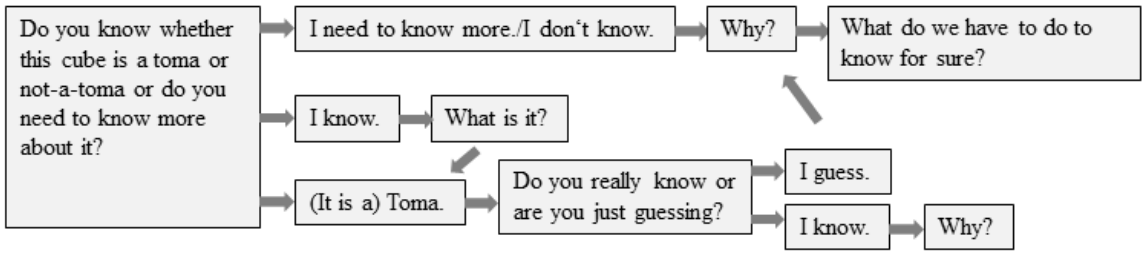

Figure 5.4. Questions in the experimental phase in response to children's answers

Forced-choice. In this phase, the experimenter brought two cubes: one was the target cube and the other was a cube with a novel color, on a tray; cubes were hidden by a cloth piece. Before revealing the cubes, children were told that there were two cubes and they could choose only one of the cubes and try on the box. The experimenter took away the cloth and asked children which cube they wanted to try. After children made a choice, the experimenter asked why they wanted to try that cube. After children's justifications, the experimenter placed the object on the light box. The target object always turned the light on; the novel object turned the light on half of the cases.

5.2.5 Coding

All verbal responses and, additionally the gestures (i.e., head nod, head shake) that pragmatically conveyed critical information in response to interview questions were transcribed. In the familiarization phase, children's responses regarding their (lack of) knowledge were coded into two primary categories: "statement of ignorance" vs. "statement of knowledge". These primary categories were a function of children's responses to two sub-questions: (1) whether they stated that they knew what the object was, and (2) whether they said they really knew or they guessed. Children's immediate utterances indicating that they did not know, such as "I don't know," "No," and head

shake gesture were coded as statements of ignorance. Children who first claimed they knew but said they guessed after the know vs. guess question were also classified as providing statement of ignorance. Verbal utterances that were indicating an object category ("It is a toma"), verbal responses indicating knowledge ("I know.") and head nod gesture which were followed by a statement of knowledge to the know vs. guess question were coded as statements of knowledge. Each child participated in two identical familiarization phases with different objects. We calculated a total score for each child as a function of their knowledge judgments in the two phases. Children received one point for a statement of ignorance in each phase and received zero points for a statement of knowledge. The scores of the two phases were summed up; as a result, each child received a familiarization score ranging from 0-2: score 2 indicating correctly stating ignorance in the two phases, score 1 indicating correctly stating ignorance one time and score 0 indicating incorrectly claiming knowledge in the two phases.

In the experimental phase, responses to the main question (Do you know… or do you need to know more…?) were coded into two primary categories: "information seeking" vs. "statement of knowledge." Utterances indicating that more information was required (e.g., "(I) need to know more") were coded as information seeking. Some of the children suggested trying the cube alone (the isolation of variables strategy) in response to the question. Those cases were also coded as information seeking. Children who first claimed that they knew but answered the know vs. guess question by saying that they guessed were classified in the category of information seeking. Children who said they knew and answered the know vs. guess question by saying that they knew were coded as providing a statement of knowledge.

Observations showed that children sometimes spontaneously uttered the object category (e.g., a toma!) in the confounded evidence phase and experimental phase right

after observing the effect, before the experimenter's question. As these spontaneous responses might be indicative of epistemic states, we transcribed and coded them into two categories: whether children spontaneously uttered an object category (e.g., a toma, not-a-toma) or did not.

In the forced-choice phase, children's object choices were coded in terms of whether they chose the novel or the target object.

In the familiarization and experimental trials, children's justifications for their knowledge judgments were coded into four primary categories: "strict evidence-based", "indistinctive evidence-based," "other/unclear," and "no response." The correctness of the justifications was coded in relation to responses given to the knowledge judgments. Utterances clearly emphasizing the informativeness of evidence were coded as strict evidence-based justifications with two subcategories. Justifications indicating a clear reference to the ambiguous nature of the evidence (e.g., "because you put both of them") were coded as reference to confounded evidence, and justifications indicating a clear reference to the unambiguous nature of the evidence (e.g., "because the box lit up only with that one [target object]") were coded as reference to unconfounded evidence. Justifications that emphasized evidence without commenting on any distinctive reference to evidence characteristics (e.g., the box lit up.) were coded as indistinctive evidence-based justifications. Justifications indicating a knowledge state (e.g., "because I know"), justifications consisting of made-up hypotheses (e.g., "because the green one is not strong as the pink one"), and unclear ones were coded as other/unclear. Children who did not give any justifications were classified as providing no response[16].

[16] The indistinctive evidence-based justifications were wrong in the confounded condition. On the other hand, the indistinctive evidence-based justifications in the unconfounded condition were not necessarily wrong, but missing emphasis on the unambiguous nature of evidence.

In the experimental trials, children's answers to the question "What do we have to do to know for sure?" were coded into two main categories. Children who suggested isolating the cubes (e.g., "have to put that one (target cube) alone") were coded as "isolation of cases". All other responses or missing responses were coded as other/no response.

Children's justifications for their choice in the forced-choice were coded into four main categories: "information gain," "effect production," "other/indistinctive," and "no response". Utterances that indicated that the chosen object would yield new information (e.g., "Because I don't know whether it is a toma or not-a-toma," "Because we haven't tried that one") were coded as information gain. Utterances that indicated the intention to activate the light (e.g., "Because I think that it makes the box light up," "Because it is a toma") were coded as effect production. Utterances irrelevant to any knowledge gain or effect production (e.g., "Because it is my favorite color") or indistinctive utterances (e.g., "Because I don't know") were coded as other/indistinctive, and children who did not respond were classified as providing no response.

All of the data was coded by the author of the thesis. A second rater coded one-third of the data (20 participants). Interrater reliability for all codes was calculated by Cohen's Kappa and all scores were near-perfect (.80 and above) (Landis & Koch, 1977). For the knowledge judgments, interrater reliability was .91, and 1.00 in the two conditions. Interrater reliability was 1.00 for the spontaneous responses and suggestion of correct test responses in the two conditions. The reliability ranged between .83 and .93 for the justifications in different phases.

5.3 Results

First, the comparison of children's information seeking in the Confounded and Unconfounded Condition is reported, followed by children`s knowledge statements in the familiarization trials, children's justifications for their information seeking/knowledge status statements, as well as, the results of the suggestion of correct test, object choice, and justifications for object choice.

Information seeking. The primary interest of the present study was the comparison of children's information seeking decisions in the case of confounded and unconfounded evidence. In the Unconfounded Condition, 95% of the children (57 out of 60) reported that they knew the causal category of the target object and 5% of the children (3 out of 60) reported they did not know. In the Confounded Condition, 47% of the children (28 out of 60) reported that they knew and 53% of them (32 out of 60) reported they did not know the causal category of the target object. Half of the children (30 out of 60) responded correctly in both of the conditions. Mc Nemar's test was run to determine if there was a difference in the proportion of information seeking and knowledge statements in the two conditions. The proportion of information seeking was significantly higher in the Confounded Condition than the proportion of information seeking in the Unconfounded Condition, $\chi^2(1) = 25.290$, two-sided, $p < .001$.

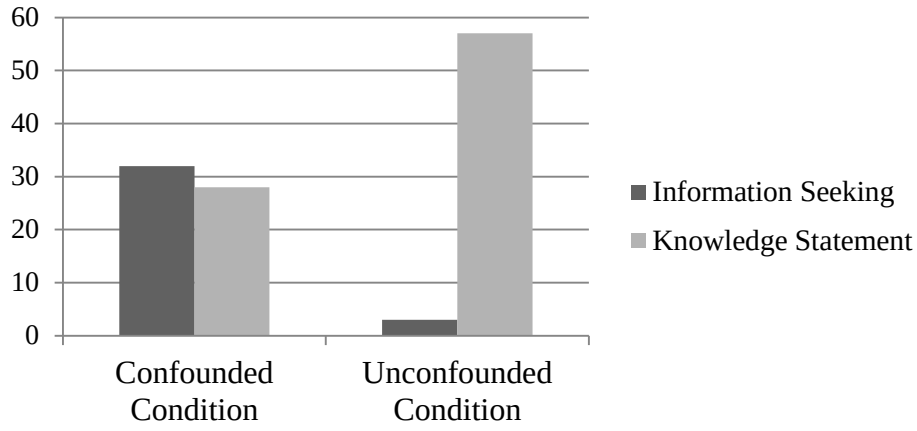

Figure 5.5. Frequencies of information seeking and knowledge statements in the Confounded and Unconfounded Condition

Children's spontaneous reactions after they saw unconfounded vs. confounded evidence might also serve as an implicit indication of their knowledge. Mc Nemar's test was conducted in order to compare children's spontaneous reactions in the Confounded and Unconfounded Condition; it revealed a significant difference between the two conditions, $\chi^2(1) = 10.562$, two-sided, $p = .001$. Seven percent of the children (4 out of 60) spontaneously uttered the object's causal category (e.g., a toma!) in the Confounded Condition whereas 30% of the children (18 out of 60) uttered the causal category in the Unconfounded Condition.

Knowledge judgments in the familiarization trials. The beginning of the task in each condition consisted of a familiarization trial where children always observed a target object placed together with another object (confounded evidence). This means each child was asked two times (one time before each experimental trial) whether they knew or did not know the causal category of the target object. Children got the score 2 if they correctly stated ignorance both times, they got the score 1 if they correctly stated ignorance one time and they got the score 0 if they stated they knew. Table 5.1 displays the distribution of knowledge status scores in the familiarization trials.

Table 5.1

Proportion and Percentage of Children's Knowledge Status Scores in the Familiarization Trials

Score	Frequency	Percentage
2	28	47%
1	14	23%
0	18	30%

Note. $N = 60$

Evidence-based justifications. The two familiarization trials were very similar to the experimental trial of the confounded condition in terms of the evidence children observed and the questions they were asked. In all these three cases, children observed two objects placed together on the light box and were asked either for their knowledge status or whether they required more information and why or why not. As these three scores represent the same theoretical construct, we combined children's justifications for knowledge in the case of confounded evidence and categorized children into four groups: (1) children who gave strict evidence-based justifications in three of the cases, (2) children who gave strict evidence-based justifications in two of the cases, and (3) children who gave strict evidence based justifications only one time and children did not give strict evidence-based justifications at all in the case of confounded evidence. Eleven percent (11 out of 60) provided strict evidence-based justifications three times, 8% of the children (5 out of 60) provided strict evidence-based justifications two of the times, 13% of the children (8 out of 60) provided strict evidence-based justifications once and 60% of the children (36 out of 60) did not provide any strict evidence-based justifications. This indicates that 40% of the children (24 out of 60) provided strict evidence-based justifications at least once. Some example utterances of children are below:

"Because both [the cubes] were put on it [the light box] and I don't know whether that one [target object] was [a toma] or not."

"Because I don't know which one [the cubes] make it light up and which one did not."

"Because both were there and I don't know whether one of the two make it light up or not; or both of them make it light up."

"Because one can't know. Because the other one [other cube] was also there. If one puts a not-a-toma and a toma [on the box], it [the light box] lights up anyways."

"Because when both [cubes] are on it [the light box], then one of the two can be a baffe."

Table 5.2 displays the type of justifications for confounded evidence for the two familiarization trials and confounded experimental trial.

Table 5.2

Percentage and Proportion of Children`s Type of Justification in the Trials of Confounded Evidence

Justifications Type	**Fam Trial a**	**Fam Trial b**	**CC Exp Trial**	**UC Exp Trial**
Strict	27% (16)	32% (19)	27% (16)	17% (10)
Indistinctive	20% (12)	18% (11)	22% (13)	62% (37)
Other/Unclear	30% (18)	37% (22)	35% (21)	13% (8)
No response	23% (14)	13% (8)	17% (10)	8% (5)

Note. *N* = 60. Strict = Strict evidence-based justification, Indistinctive = Indistinctive evidence-based justification, Fam = Familiarizaton, Exp = Experimental, CC = Confounded Condition, UC = Unconfounded Condition

Children observed unconfounded evidence one time in the Unconfounded Condition. Seventeen percent of the children (10 out of 60) provided strict evidence-based justifications emphasizing the unambiguous nature of the evidence. Sixty-two (37 out of 60) of the children provided indistinctive evidence-based justifications (e.g., the

box lit up). Twenty-two percent of the children (13 out of 60) either provided other type justifications or did not provide any justifications. Some example utterances for strict evidence-based justifications are below:

"Because the second one [other cube] was not on it [the light box] and the box lit up when it [the target cube] was alone."

"Because you [the experimenter] put it [the target cube] alone on the box and it made the box light up."

"Because it [the target cube] was on the box alone and box lit up."

"Because it [the light box] lit up only with it [the target cube]."

Suggestion of correct test. Children who stated that they needed more information were asked what should be done in order to have more information. Thirty-two children in the Confounded Condition said they required more information. In two of the cases, the question was not asked due to experimenter error. Out of 30 children who stated that they needed more information, 23 (77%) suggested isolating the objects in order to determine the causal category of the target object. Some example responses are below:

"Only put that one [target cube] on it [the light box]."

"Just put it [the target cube] alone on the box, without the other one."

In the Unconfounded Condition, three children reported that they needed to know more and they did not respond when they were asked what should be done in order to know more.

Object choice. We hypothesized that children would show a stronger preference for the novel object in the Unconfounded Condition in comparison to the Confounded Condition. This forced-object choice would be evidence for a preference to choose the informative object. We conducted a Mc Nemar's test and compared children's choice of

object (target vs. novel) after observing confounded versus unconfounded evidence, and there was no significant difference between the two conditions in terms of object choice, p = .405. Table 5.3 shows children's object choice (novel vs. target) in the two conditions.

Table 5.3

Frequency and Proportion of Children's Object Choice in the Confounded and Unconfounded Condition

	Confounded Condition	**Unconfounded Condition**
Target Object	28 (47%)	23 (38%)
Novel Object	32 (53%)	37 (62%)

Note. N = 60

After children chose an object, they were asked to tell the reason for their choice. In both of the conditions, a high percentage of children provided an "other" type of justifications, e.g., "Because it is my favorite color," "Because I want.").

Table 5.4 shows justifications of children who chose the target object. A qualitative inspection of the results showed that little number of the children (9%) gave information gain justifications for choosing the target object in the Unconfounded Condition. The comparison of the information gain scores in the Confounded and Unconfounded Condition suggests that children did not state there is more to learn from the target object in the Unconfounded Condition. Ideally, choosing the target object with effect production motivation in the Confounded Condition is wrong, because the causal category of the target object was unknown in this condition. However, approximately half of the children stated that they knew the object category in the Confounded Condition. The wrong knowledge statement in the case of confounded evidence might be the reason for why they chose the target object with effect production motivation.

Table 5.4

Percentage and Frequency of Children's Justifications Who Chose the Target Object

	Confounded Condition	**Unconfounded Condition**
Information Gain	29% (8)	9% (2)
Effect Production	29% (8)	39% (9)
Other	43% (12)	52% (12)
Total	28	23

Table 5.5 shows the justifications of children who chose the novel object. It is informative that only a small number of children (3% and 8%) provided effect production justifications for choosing the novel object in both conditions. Information gain and effect production scores for the novel object suggest that about one-third of the children (31% and 43%) were aware of the informative nature of the novel evidence and few children chose the novel object for effect production purposes (3% and 8%).

Table 5.5

The Distribution of Children's Justifications for Choosing the Novel Object

	Confounded Condition	**Unconfounded Condition**
Information Gain	31% (10)	43% (16)
Effect Production	3% (1)	8% (3)
Other	66% (21)	49% (18)
Total	32	37

Exploratory analyses. We investigated whether there was an individual consistency for the object type. In other words, whether there was a tendency that some children always chose the target object and some children always chose the novel object. Table 5.6 shows the distribution of children's choices. Chi-square test was conducted to investigate whether there was a tendency for object choice. The results suggest there

might be a trend for choosing the novel object in both conditions, $\chi^2(3) = 6.800$, $p = .079$.

Table 5.6

Individual Consistency for Object Type (Target vs. Novel) in the Forced Choice

	Frequency	Percentage	Expected	Residual
Both target	14	23%	15	-1.0
Both novel	23	38%	15	8.0
CC Target & UC Novel	14	23%	15	-1.0
CC Novel & UC Target	9	15%	15	-6.0

Note. CC = Confounded Condition, UC = Unconfounded Condition

5.4 Discussion

The goal of Study 3 was to investigate whether preschoolers have a metacognitive understanding of, particularly the ability to reflect upon, the relation between their own epistemic states and evidence. Our first research question was whether preschoolers have an awareness of their epistemic states (knowledgeable vs. ignorant) in response to different evidence patterns (unconfounded vs. confounded). We found that 5- and 6-year-olds differentially stated ignorance and knowledge in the case of confounded and unconfounded evidence. When children were presented with ambiguous evidence regarding the causal category of an object, approximately half of them stated that they required more information to know the category while only two did so when they were presented with causally unambiguous evidence. Our second research question was whether preschoolers have a reflective awareness of the relation between their epistemic states and evidence they observed. We examined their ability to justify their ignorance as a function of evidence they were presented with. Our results demonstrated that 40% of the children provided strict evidence-based justifications at

least one time. Our third research question was whether children explicitly describe an informative test in order to gain knowledge in the case of confounded evidence. Seventy-seven percent of the children who said that they need to know more in the confounded condition (n = 30) correctly described the isolation of variables strategy in order to reveal information. Our final research question was whether children would also show a behavioral preference for the causally ambiguous objects. Our results showed that there was no differential preference for ambiguous object between the Confounded and Unconfounded Condition.

Study 3 demonstrated that preschoolers not only have an implicit understanding of confounded evidence, which has been frequently shown by exploration studies (Bonawitz et al., 2015; Cook et al., 2011; Schulz & Bonawitz, 2007) but that they also have an explicit understanding of their own epistemic states and how their epistemic states are constructed as a function of evidence. Our findings are contrary to the argument that young children lack metacognitive understanding of their knowledge formation processes and they demonstrate that 5- and 6-year-olds can reflect on the relation between their epistemic states and evidence. In this respect, the claim that "[y]oung children think *with* theories, rather than about them" (Kuhn, 2010, p. 499) is an underestimation of young children`s early abilities.

Children's ignorance statements and evidence-based justifications in the present study suggest that they can represent alternative hypotheses. In the case of confounded evidence, there is an uncertainty regarding the efficacy of the target object; evidence is uninformative regarding the hypothesis, and there are two alternative hypotheses (the target cube is a toma vs. the target cube is not-a-toma). Children's ignorance statements are an indirect indicator of their ability to implicitly represent alternative hypotheses. Considering that children are motivated to provide spontaneous judgments regarding the

category of the objects when they (think that they) know it, claiming ignorance suggest that they are aware that they lack relevant evidence with regard to whether the object is a toma or not-a-toma. However, an argument contrary to this conclusion may be that a claim of ignorance does not necessarily entail being able to think about the alternative possibilities. Children may say "I don't know" without proper representation of the alternative hypotheses (it is a toma vs. it is not-a-toma). Therefore, the link between providing ignorance statements and representing alternative hypotheses is less clear. Besides children's evidence-based justifications are especially informative to address the question whether they can represent the alternative hypotheses. Children's justifications in the present study demonstrate that they can, indeed, represent alternative hypotheses. In the case of confounded evidence, many children explicitly addressed that the target cube may or may not be a toma. They have an explicit understanding of the uncertain nature of the hypotheses, and they can represent two mutually exclusive hypotheses in the case of confounded evidence.

The ability to report ignorance after observing confounded evidence, however, was not found for the whole sample. Half of the children claimed that they knew although they had seen confounded evidence. What leads to this difference between the children who correctly reported ignorance and who mistakenly reported knowledge after observing confounded evidence is an open question. Studies investigating young children's implicit understanding of ambiguity often demonstrate that not all children show sensitivity to ambiguity, but only some children do (e.g., Legare, 2012; Schulz & Bonawitz, 2007). In Cook et al. (2011), for instance, half of the children isolated the variables during exploration. Based on these results, it may be that it is not the metacognitive understanding that is missing, but the implicit understanding of ambiguity in the case of confounded evidence. There have been no studies on the relation between

implicit sensitivity to ambiguity and metacognitive ability to understand the relation between epistemic states and ambiguity. Further studies bringing together the measures of implicit exploration and metacognitive abilities and using the same sample of children are necessary to address this question.

Children's wrong knowledge statements in the case of confounded evidence might be due to an immature *feeling of competence*—a cognitive heuristic that Rohwer et al. (2012) addressed in explaining young children`s knowledge statements in the so-called "partial exposure task." In their study, younger children`s poor performance in acknowledging their ignorance was specific to the partial exposure task, while they were successful at reporting ignorance when they were totally ignorant. Rohwer et al. (2012) argued that it may be challenging for younger children to inhibit the feeling of knowing, because their partial knowledge is very salient. In Study 3, the feeling of knowing heuristic may be a possible mechanism for wrong knowledge statements. The saliency of the association between the cubes and the light may be too salient, which may have led some children to a feeling of knowing.

Study 3 did not find any behavioral preference for the causally ambiguous object; there was no difference between children's preference for the target object in the confounded and unconfounded condition. Based on earlier findings (Schulz & Bonawitz, 2007[17]), we expected that children would show strong novelty preference in the unconfounded condition. However, in the present study, children did not show any

[17] The procedure of Schulz and Bonawitz (2007) is similar to Study 3 Object Choice Phase. In the unconfounded condition, children interacted with a causally unambiguous toy; whereas in the confounded condition, they interacted with a causally ambiguous toy. Later, children were given chance to explore the familiar and the novel toy. Researchers measured children's average time playing, play time preference, and first reach. Within these measures, the dependent variable first reach was the most similar one to Study 3's object choice measure. Close inspection of their results show that there was a strong preference for the novel object in the unconfounded condition whereas the preference for the novel object and the ambiguous object was half-half in the confounded condition.

preference for the novel object in the unconfounded condition. Without strong default preference for the novel object in the unconfounded condition, it is hard to experimentally show the preference for the target object in the confounded condition. Therefore, it may be that the lack of novelty preference in the unconfounded condition hindered revealing the sensitivity for ambiguity in the behavioral choice task. This may be because we did not let children interact with the objects. It has been shown that it makes a difference for children whether they do interventions themselves or whether they watch someone doing those interventions (Sobel & Sommerville, 2010). It may be that, although children knew the target objects' efficacy in the unconfounded condition, the motivation to interact with the target object itself was high; and this lowered the preference for the novel object. A follow up study in which children are let to interact with the objects themselves would help us to test this hypothesis.

The present study is one of the first to show preschoolers` ability to reflect on the formation of their epistemic states as a result of evidence. Five- and 6-year-olds selectively reported ignorance and knowledge after observing confounded and unconfounded evidence. Forty-percent of the children were able to provide evidence-based justifications by explicitly commenting on the confounded nature of evidence as a cause of their ignorance and the majority of the children who stated being ignorant were able to describe the correct test—the isolation of variables strategy. Children's justifications provided strong evidence for their ability to represent alternative hypotheses when the evidence is uninformative. Put together, the findings of Study 3 show that the metacognitive understanding of the relation between epistemic states and evidence and, particularly, of the relation between hypotheses and evidence is already present in the late preschool years.

6 General Discussion

The development of early scientific reasoning skills has been largely unexplored. Recent empirical findings on early causal reasoning abilities show that young children possess powerful learning mechanisms that are similar to epistemic practices in science such as hypothesis testing and evidence evaluation. Theoretically, scientific reasoning requires an understanding of hypothesis–evidence distinction and metacognitive understanding of knowledge seeking and formation processes (Kuhn, 1988); however, little was known about the development of these abilities in preschool years. Three studies in this thesis investigated preschoolers` abilities in the epistemic activities of hypothesis testing, evidence evaluation, and argumentation from evidence with the focus on the understanding of the hypothesis–evidence relation and metacognitive understanding of this relation. By bringing together the recent empirical findings on causal reasoning and scientific reasoning, this thesis aimed to enhance our understanding of the development of scientific reasoning in early childhood. This final chapter of the thesis will provide a discussion of the present findings in regard to theories on the development of scientific reasoning, recommend practical implications for applied fields, and propose directions for further research.

6.1 Summary of the Three Studies

Scientific reasoning has been considered a late developing skill. Early studies on scientific reasoning suggested that young children do not have a differentiated understanding of epistemic categories of hypotheses and evidence; in fact, they have a script-like representation that they confuse the epistemic categories (Kuhn & Pearsall, 2000). One argument was that this confused representation is due to a conceptual deficit

in young children in representing epistemic categories. On the other hand, recent research has shown that young children have complex mechanisms for learning from evidence: they are sensitive to evidence characteristics and even make interventions to reveal causal relations (Gopnik & Wellman, 2012). This suggests that young children at least have an implicit understanding of the informativeness of evidence. However, little is known regarding children's understanding of the hypothesis–evidence relation and their metacognitive understanding. The aim of this thesis was to investigate young children's ability to understand the inferential relation between hypothesis and evidence; as well as their ability to reflect on this relation. It is important to note that the aim was not to investigate the developmental trajectory from early forms of causal reasoning to scientific reasoning but to examine young children's basic abilities for scientific reasoning. This line of research can shed light on the early competences for scientific reasoning and inform further research investigating developmental progression from early competences of learning from evidence to mature forms of scientific reasoning.

The first study showed that older 5- and 6-year-olds can differentiate epistemic goals of hypothesis testing from practical goals of effect production which was shown by their selective interventions in the case of different goals. Four-year-olds, on the other hand, did not discriminate hypothesis testing from effect production; yet they showed selective interventions in the case of an exploratory epistemic goal. Older 5- and 6-year-olds' selective interventions in the hypothesis testing and effect production conditions demonstrates that the notion of empirical test is present in this age. These children could assess the missing piece of information in order to test a hypothesis and to choose the correct piece of evidence to test the truth of a hypothesis. These findings add to the literature that, at around the end of 5 years, preschoolers do understand what it means to make conclusive tests and they differentiate between the goals of hypothesis

testing and effect production (Piekny & Maehler, 2013; Piekny et al., 2014). Since preschoolers in our study were able to differentiate epistemic goals from practical goals; this suggests that older children's poor performance in differentiating epistemic goals from practical goals shown in earlier studies (e.g., Perner & Klahr, 1996) is not due to domain-general conceptual deficit in hypothesis–evidence relation, but potentially due to task complexity, or domain-specific content knowledge requirements.

While 4-year-olds and young 5-year-olds did not choose the correct piece of evidence in the Hypothesis Testing condition of Study 1a and 1b, they were able to choose the informative object in the case of exploratory epistemic goal in Study 1c. These findings suggest that the understanding of testing a hypothesis—generating evidence in order to gain information to judge the truth or falsity of a statement—may not be yet present at this age but develops around 5-years. In the case of exploratory epistemic goals, when the main goal is just to maximize information gain without any requirements to test a hypothesis, 4-year-olds did choose the informative object. Our findings are in line with former studies which have shown that preschoolers make informative interventions in the case of ambiguous or inconsistent evidence (Bonawitz et al., 2012; Cook et al., 2011; Legare, 2012). Taken together, these findings may be indicative of a transition from exploratory, information gain-oriented epistemic understanding to a later developing understanding of hypothesis–evidence relation. It is also important to note that changes in several related cognitive changes (e.g., language, executive function, working memory) take place at this age (see Goswami, 2014). Therefore, the difference between 4- and young 5-year-olds might have appeared due to changes in other cognitive skills rather than epistemological development.

Study 1 showed that children around 5-years begin to differentiate hypothesis testing from effect production. However, it does not show that whether children of this

age have an explicit understanding of the empirical relation between hypotheses or beliefs and evidence. Study 2 investigated preschoolers' understanding of hypothesis–evidence relation in the epistemic activities of hypothesis testing and argumentation. Argumentation from evidence especially provided critical information about children's understanding that judgments can be made about the truth or falsity of beliefs based on evidence. In the Hypothesis Testing condition, preschoolers followed systematic strategies (i.e., contrastive testing, positive testing), and there was a developmental difference; while almost all children in the older group followed systematic testing strategies, more children in the younger group did untargeted exploration without any systematic testing patterns. In the argumentation task, around 70% of the children provided verbal counterarguments (e.g., "No, some light ones make the box work, too.) in order to refute the false causal claims, which suggests that children understand the epistemic relation between beliefs and evidence—that judgments can be made based on evidence. Children's explicit verbal comments demonstrated that children can reflect on the relation between beliefs and evidence which suggest that there is at least a beginning metacognitive understanding of belief–evidence relation. Moreover, children did interact with more evidence that refutes the false claims which show that they are able to evaluate relevant evidence to falsify beliefs.

Study 2 demonstrated that children have, at least, a beginning understanding of the relation between claims and evidence shown by their systematic hypothesis testing and argumentation from evidence. Study 3 aimed to further investigate preschoolers' metacognitive understanding of hypothesis–evidence relation in evidence evaluation. We found that 5- and 6-year-olds have the ability to reflect on the epistemic relation between evidence and their knowledge states. Half of the 5- and 6-year-olds differentially (and correctly) reported ignorance and knowledge in the case of

confounded and unconfounded evidence. The reflective ability on one's own epistemological processes seems to be present at 5 years. Forty percent of the children provided, at least once, justifications for why they were ignorant by referring to the confounded nature of evidence as a reason. Approximately 70% of the children who said that they were ignorant in the case of confounded evidence were able to describe the correct test—the isolation of variables—as a way to gain information. Taken together, Study 3 showed that there is a beginning metacognitive understanding of the epistemic relation between evidence and knowledge states. Although it was not general to the whole sample, half of the 5- and 6-year-olds were able to explicitly report ignorance due to confounded evidence.

6.2 Synthesis of the three studies

Young children have powerful mechanisms in learning from evidence: they learn from statistical patterns, they are motivated to learn more in the case of ambiguous evidence, and they even make interventions to reveal causal structure (Gopnik & Wellman, 2012). These similarities between learning processes and mechanisms in young children and the core epistemic practices of science support the child-as-scientist metaphor. However, the question, whether children can reason scientifically, was unclear; because scientific reasoning theoretically requires a differentiated understanding of hypotheses and evidence; and an ability to reflect on the epistemic processes. The present thesis aimed to investigate these aspects of scientific reasoning in epistemic activities of hypothesis testing, evidence evaluation, and argumentation from evidence in order gain more knowledge on the question whether young children can reason scientifically, at least at a basic level.

Findings of the three studies of this thesis yielded evidence that preschoolers have basic abilities for hypothesis testing, evidence evaluation, and argumentation from evidence. Study 1 demonstrated that the ability to differentiate epistemic goals of hypothesis testing and practical goals of effect production is present in preschool age. Study 2 demonstrated preschoolers follow systematic hypothesis testing strategies and can refute false causal claims by means of verbal counterarguments and correct pieces of evidence. Study 3 revealed that about half of the sample showed reflective awareness about their ignorance because of confounded evidence. The three studies together provided evidence that the basic competence for the understanding of hypothesis–evidence relation and metacognitive understanding of knowledge states is developing in preschool years.

Differentiating mental concepts of beliefs, hypotheses from evidence is critical for scientific reasoning. It has been argued that preschoolers have a conceptual deficit in representing the epistemic categories of hypotheses, beliefs, and evidence distinctively; and they have a script-like representation in which hypotheses and evidence are represented as one. This thesis presented evidence contrary to this view. Our findings suggest that preschoolers are able to differentiate these epistemic categories, at least in the simple problems. Preschoolers' informative interventions in Study 1 and Study 2 suggest that children understand the relation of how a piece of evidence is relevant in order to gain knowledge about the truth of a statement. In the Hypothesis Testing condition, especially older preschoolers were able to generate the informative evidence in the case of hypothesis testing; whereas both age groups were able to generate disconfirming evidence in order to refute the claim. The relevant evidence generation behaviors show that children understand the epistemic role of evidence in relation to hypotheses. Although there has been already evidence that preschoolers generate

evidence in order to gain information; little was known about their generation of evidence in relation to the veracity of a proposition. Moreover, children's evidence-based verbal counterarguments in Study 2 and referrals to confounded nature of evidence as a justification for ignorance also suggest that they understand the empirical relation between hypotheses (or beliefs) and evidence.

Metacognitive understanding has been designated as one of the fundamental reasons why young children's knowledge seeking and formation are different from scientific reasoning (Kuhn, O'Loughlin, & Amsel, 1988; Kuhn, 2010). However, little has been known about the metacognitive awareness of epistemic concepts in early childhood. Although there have been many studies on causal reasoning, to our knowledge, none of the studies investigated metacognitive abilities. Yet, this ability is critical for the development of scientific reasoning. Children's evidence-based verbal counterarguments in Study 2 and their knowledge judgments and evidence-based justifications in Study 3 do not support the view that preschoolers lack metacognitive abilities for reflecting on their epistemic states. Our findings suggest that preschoolers have a metacognitive understanding of their epistemic states, at least when the task demands are low. It is critical to note that these findings do not suggest that this is a general skill that one can easily apply to other contexts and domains. It has been again and again shown empirically that the reflective awareness and deliberate control over one's epistemic states are very challenging; even adolescents and adults have problems in doing so (Sodian & Bullock, 2008). This ability is probably a continuously developing skill throughout development as a consequence of experience with successive problems with increasing complexity. Developmental research on metacognition of other cognitive processes, such as metacognition of memory, also suggests a protracted development (O'Leary & Sloutsky, 2017), with basic abilities

manifesting in preschool years (Sodian, Schneier, & Perlmutter, 1988). Our findings are in line with such a cognitive developmental pattern.

Critical developmental changes in several cognitive abilities take place during the preschool years, and one of the research questions of this thesis was to investigate the developmental changes in scientific reasoning from 4- to 6-years of age. Study 1 and Study 2 are especially informative with respect to developmental differences since their samples included 4-, 5-, and 6-year-olds. We found developmental differences in children's hypothesis testing skills in both Study 1 and Study 2. In Study 1 there was a developmental increase in differentiating hypothesis testing from effect production. A close look at the results suggested that 5-years of age is a critical point in development. Older 5- and 6-year-olds differentiated epistemic goals of hypothesis testing from practical goals of effect production; however, 4-year-olds and younger 5-year-olds only showed differential responding only when the epistemic goals were exploratory. In Study 2, whereas almost all older 5- and 6-year-olds followed some form of hypothesis testing strategy (i.e., contrastive or positive), one-third of the 4- and younger 5-year-olds were not able to follow any systematic testing strategies. These two findings suggest a developmental change in hypothesis testing skills from 4 to 6 years of age.

Different from the developmental difference in hypothesis testing, we did not find any differences in verbal counterarguments and evidence-based argumentation. We cannot directly compare the hypothesis testing and argumentation performance in Study 2 because the argumentation results were based on a group of 4- and young 5-year-olds who were able to revise their prior weight belief to the later sticker belief. Therefore, it was unclear in Study 2, how children who were excluded from the argumentation phase would perform in the counterargumentation task. However, the findings of a following study (Köksal Tuncer et al., 2017) which showed that majority of the younger children

were able to provide evidence-based counterarguments (For a discussion see section 4.4). In this regard, at 4 years of age, preschoolers already have an understanding of the epistemic relation between causal beliefs and evidence.

One hypothesis for the different developmental pattern in the hypothesis testing and argumentation tasks may be due to the differences between the representations of hypotheses and beliefs in terms of uncertainty. Understanding of uncertainty is critical for hypothesis testing since the concept of hypothesis is about acknowledging the alternative possibilities, and hypothesis testing is conducting tests which would produce information to show which of the possibilities are indeed true. On the other hand, children already have a belief during the argumentation. They did not seem to question the truth of this belief (and they did not need to); therefore, there was no uncertainty about the truth of the belief they counterargue for. Studies on the development of understanding uncertainty demonstrated that there is a critical change in early childhood in terms of understanding uncertainty. Recent evidence suggests that 4-year-olds show some understanding but their performance was far from perfect (Fernbach et al., 2012). Around 5 years of age children begin to show an understanding of uncertainty in simpler causal models and the ability continues to develop at 6 and 7 years of age (Sobel et al., 2017). This developmental pattern of uncertainty between 4 to 6 years may be one of the underlying reasons for the different developmental patterns we found in hypothesis testing and argumentation tasks. Argumentation task performance suggests that children have an understanding of the belief–evidence relation—that their beliefs are constructed as a result of evidence. This is in line with the children's false belief understanding that develops around 4 years (for a review see Sodian, 2005). However, belief–evidence understanding may not entail understanding of hypothesis–evidence relation which seems to be developing around 5 years of age. We can only speculate about the later

development of the ability to represent hypotheses because there are several other differences between the hypothesis testing tasks in Study 1 and Study 2; and argumentation task in Study 2 in terms of the task characteristics (i.e., belief states, the nature of the hypothesis). This prevents us from making direct comparisons between hypothesis testing and argumentation tasks.

All in all, the present thesis showed that young children have basic abilities for scientific reasoning. They have a differentiated understanding of epistemic categories. The understanding of how evidence is a means for making judgments about the veracity of beliefs is already present around 4 years of age. The advanced understanding of the concept of hypothesis and testing hypotheses found to be developing around 5 years. Four-year-olds have the ability to reflect on the epistemic relation between beliefs and evidence, and half of the 5- and 6-year-olds reflect on the relation between alternative hypotheses as a result of confounded of evidence.

6.3 Implications

Fostering scientific thinking is one of the 21st century education goals since citizens with evidence-based and critical thinking skills are the essential constituents of functioning democratic societies (e.g., Trilling, & Fadel, 2009). Formal and informal ways of fostering scientific reasoning should start from early childhood since the quality, and the characteristics of early environments have critical effects on later development (Heckman, 2016). In this respect, evidence on reasoning skills in early childhood is critical for shaping educational programs and curricula. Contrary to early accounts claiming that young children are concrete, irrational, noncausal thinkers, recent findings have shown that young children's reasoning skills were highly underestimated (Institute of Medicine and National Research Council, 2015). Findings of the present

thesis add to our knowledge regarding young children's early reasoning competencies; and this has critical implications for early childhood education. Bearing in mind that this thesis did not investigate implications to real-life contexts—future studies are necessary to replicate the present findings and further investigate the implementation of the findings into early learning environments—this part of the thesis proposes possible implications of the present findings for applied fields.

Present findings showed that preschoolers have distinct epistemic categories of hypotheses and evidence. Acknowledging this competence and fostering it in different content domains starting in early childhood would be beneficial for developing mature forms of scientific reasoning later in life. Science activities are getting more and more common in kindergartens; however, it is unclear whether children approach these tasks as epistemic activities or whether they only have an engineering approach. In this respect, building on the knowledge that preschoolers already have a basic understanding for the epistemic categories of hypothesis and evidence, early education may foster early competences by engaging epistemic activities, which may subsequently help children when they are dealing with more complex problems. For instance, preschool classroom activities might emphasize the existence of alternative possibilities regarding phenomena or stress uncertainty due to lack of evidence. In this process, children are encouraged to make (informed) predictions and further explore the phenomenon by making interventions. Depending on the task complexity, children may be encouraged to draw their own conclusions or teachers might help children draw conclusions from evidence. Design of such learning opportunities for children would be beneficial in advancing the core skills for scientific reasoning in simpler, child-friendly tasks.

Children's counterarguments in the Study 2 and their explicit judgments in Study 3 demonstrated that metacognitive understanding of the epistemic states and empirical

relations between hypotheses (or beliefs) and evidence is present in preschool years. Fostering metacognitive abilities of knowledge seeking and formation processes in early childhood education would be helpful for lifelong learning. Metacognitive skills are related to reasoning (e.g., Amsel et al., 2008) and learning in general (e.g., Veenman, Wilhelm, & Beishuizen, 2004; Veenman & Spaans, 2005); and interventions on metacognition advance both students' metacognitive skills and their learning outcomes (e.g., Veenman, Elshout & Busato, 1994). One of the fundamental principles of successful metacognitive instruction is "… prolonged training to guarantee the smooth and maintained application of metacognitive activity" (Veenman, van Hout-Wolters, & Afflerbach, 2006, p. 9). Considering the effectiveness of early interventions (Heckman, 2006), fostering metacognitive skills in preschool classrooms with child-appropriate activities would positively influence the later development of metacognitive skills for scientific reasoning.

Informal learning environments are critical for learning, and it would be valuable if sciencing contents are used to foster children's natural curiosity and exploration not only in formal education contexts but also in informal learning contexts. Firstly, the use of technology by young children in daily life is increasing tremendously and carefully developed sciencing games might be beneficial for fostering children's scientific reasoning skills. There is evidence that young children easily adapt to interacting with such tools and these tools have positive influences on learning environments (e.g., Beschorner & Hutchison, 2013; Couse & Chen, 2010). Computer-based tutoring systems have been developed in science for older elementary school and middle school children and have been found to be successful at teaching experimentation skills (Siler, Mowery, Magaro, Willows, & Klahr, 2010). There are ongoing studies to implement such tools for preschool age children (Moeller, Sodian, & Hussman, in progress).

Engaging digital games that bring into play the core epistemic practices such as exploration, experimentation, and evidence evaluation as means to gain knowledge as an end goal may foster young children's reasoning skills. Secondly, considering that early literacy opportunities at home play a critical role in later literacy skills (e.g., Baker, Mackler, Sonnenschein, & Serpell, 2001; Bus, van IJzendoorn, & Pellegrini, 1995) children's books may encourage children's scientific thinking by introducing them to the epistemic categories of hypothesis and evidence. There are many science related books for preschool children. Many of them present factual knowledge (e.g., space, animals, plants), whereas a few of these books emphasize the nature of science as a process of knowledge seeking. The core epistemic practices of experimentation, exploration, and evidence evaluation can be implemented in the picture based story books. To illustrate, a narrative may introduce protagonists with difference ideas (hypotheses) about a phenomenon, and then explain how conducting experiments could inform protagonists about the truth of their ideas. Emphasizing the uncertain nature of hypotheses and the role of evidence in the process of knowledge seeking may foster children's understanding of hypothesis and conducting empirical tests. Together with parent or teacher support, such books may also encourage children to make predictions and brainstorm about ways to plan experiments.

6.4 Future Research Directions

Present studies yielded empirical evidence that a basic understanding of hypothesis–evidence relation and metacognitive understanding of knowledge seeking and formation processes are already present in early childhood years. The present thesis is one of the first investigating these abilities in early childhood and much is unknown regarding the characteristics and cognitive mechanisms of early development of

scientific reasoning. In this chapter, open research questions for future research will be summarized.

Recently, the possible relation between early causal and scientific reasoning skills has captured researchers' interest in the field (e.g., Gopnik, 2012; Sobel et al., 2017). The main question, in this respect, is whether there is developmental continuity from early causal reasoning abilities to mature scientific reasoning. To date there is no empirical evidence to answer this question, but the similarities between the epistemic practices suggest that the continuity hypothesis is a legitimate one and an important direction for further research. Considering children's intrinsic motivation for causal learning (Alvarez & Booth, 2015) if there is really continuity from early forms of learning and scientific reasoning, this would be both critical for increasing our knowledge on scientific reasoning and very informative in developing better approaches to foster scientific reasoning beginning from early years of life.

If there is a developmental continuity, the few investigations in both areas signal to a complex relation. Studies investigating the development of scientific reasoning suggest that there is a developmental increase in scientific reasoning activities such as designing experiments and evaluating evidence (Bullock et al., 2009; Klahr et al., 1993; Kuhn et al., 1992; Penner & Klahr, 1996). On the other hand, older participants' performance in scientific reasoning tasks are far from perfect; this is interesting considering young children employ such strategies (e.g., isolating variables, performing unconfounded tests) during their exploratory play (e.g., Cook et al., 2011; van Schijndel et al., 2015). One important aspect of scientific reasoning (and learning in general) is the ability to acknowledge the inconsistencies in evidence and learn from them (Chin & Brewer, 1993). In the two research areas, studies investigating this ability suggest different developmental patterns. Investigation of causal reasoning abilities later in life

showed that the older the participants are, harder it is to learn unlikely causal models (Gopnik et al., 2017; Gopnik, Griffiths, & Lucas, 2015; Lucas, Bridgers, Griffiths, & Gopnik, 2014) whereas several scientific reasoning studies showed that adults are better at evaluating implausible hypotheses than children (Amsel & Brock, 1996; Klahr et al., 1993; Kuhn et al., 1988). However, this evidence is far from being sufficient to make general conclusions about the continuity hypothesis; firstly, because scientific reasoning is a complex set of abilities and cognitive flexibility is only one component of it. Secondly, it is highly likely that scientific reasoning introduces other cognitive demands (e.g., working memory, executive function, familiarity with content domain) which may make it challenging to show the relation between early and later abilities. Altogether, we believe the continuity hypothesis is a plausible hypothesis due to the similarity in the mechanisms for knowledge acquisition; yet there are no empirical findings investigating this hypothesis. Future studies are necessary in order to shed light on the developmental trajectory of scientific reasoning and its relation to early causal reasoning abilities.

Another pressing question is whether there are consistent individual differences and if so, what underlying factors and psychological mechanisms give rise to such differences. There is evidence for individual differences in children's curiosity and motivation to learn causal information (Alvarez & Booth, 2016). In the causal reasoning studies, not all, but a subgroup of children follow isolation of variables strategies in the case of ambiguous evidence (Cook et al., 2011) and provide causal explanations and differentially explore in the case of inconsistent evidence (Legare, 2012). Similarly, in Study 3, only half of the children were aware of their ignorance due to confounded evidence. It is unknown whether the differences that appear in experimental studies are due to authentic and stable individual differences. Investigating whether these individual differences are stable over time and, if so, which underlying cognitive mechanisms and

processes pave the way for these differences is critical for understanding the development of scientific reasoning.

Recent research findings on young children's powerful learning skills are valuable for fostering scientific reasoning in formal and informal learning environments because they provide evidence on children's strengths and weaknesses for reasoning scientifically. As mentioned in the implications section, the main findings of the present studies—core abilities for scientific reasoning are already present in preschool age—could be used in order to promote early learning environments for young children. In this respect, it is critical that the design of early learning environments is informed by research, because research shows there is a critical balance between learning from exploration and direct instruction. On the one side, children are intrinsically motivated to explore, and this is an important source of learning. There is evidence that direct instruction limits the scope of their exploration and restricts learning outcomes (Bonawitz et al., 2011). On the other side, scientific reasoning is difficult and requires some form of instruction (e.g., Chen & Klahr, 1999)—it is certain that children would not develop mature forms of scientific reasoning if we simply let them learn from their exploration. In this respect, future research is necessary in order to be able to design learning environments that would promote children's already present early skills and help them develop mature forms of scientific reasoning.

All three studies reported in this thesis were conducted with typically developing children of lower- to upper-middle class background from a larger German city. Since the data collection took place in kindergartens, all of the participants had been participating in early education for some time. Therefore, the findings of the studies in this thesis came from a narrow sample that has been called WEIRD populations: Western, educated, industrialized, rich, and democratic (Henrich, Heine, & Norenzayan,

2010). Cross-cultural comparisons of cognitive processes demonstrated that researchers should be careful in making generalizations based on studies conducted with WEIRD samples. Comparisons across different cultures and socioeconomic background showed that WEIRD samples found to be outlier groups in some domains such as fairness and collaboration, or folk biological reasoning. There are also other cognitive skills which show a universal pattern, although the age of the acquisition varies across cultures, such as false belief understanding, emotional expression, or psychological essentialism. Little is known empirically about the similarities or differences in scientific reasoning and causal reasoning across different cultures and socioeconomic status. Most of the studies on these topics have been conducted with children from middle- to upper class in North America and Europe (Wente et al., 2017). Wente et al. demonstrated that there were no differences between low-income Peruvian children, low-income U.S. children, and middle-class U.S. children learning about causal structure from patterns of evidence. Yet, there is much to learn regarding the differences to make conclusive judgments about the development of scientific reasoning and causal reasoning. Our findings only represent a narrow sample of WEIRD children. Future studies are necessary to investigate whether the epistemological understanding of hypotheses and evidence that we found in our studies are also present in other cultures and socioeconomic status.

6.5 Conclusion

This thesis was motivated by the recent findings on young children's powerful causal learning abilities and their potential relation to early scientific reasoning skills. Three empirical studies investigated the ability of the hypothesis–evidence coordination and the ability of reflecting on the hypothesis–evidence relation in three epistemic practices; namely, hypothesis testing, evidence evaluation, and argumentation from

evidence. In particular, the first study investigated 4- to 6-year-old preschoolers' ability to differentiate epistemic goals of hypothesis testing from practical goals of effect production. While younger preschoolers were able to differentiate epistemic goals from practical ones in the case of exploratory epistemic goals, older preschoolers properly chose relevant pieces of evidence in order to test a given hypothesis. These findings suggest that the understanding of hypotheses and providing relevant evidence in order to test hypotheses is developing between 4 to 6 years and older preschoolers can test simple hypotheses. The second study investigated 4- to 6-year-old preschoolers' hypothesis testing and argumentation from evidence in an exploration setting. Most of the preschoolers were able to follow some form of testing strategy (i.e., contrastive or positive testing) and there was a developmental improvement in the ability to follow systematic testing strategies. When presented with a false causal claim, approximately 70% children did provide evidence-based counterarguments, and they spontaneously generated more disconfirming evidence than confirming evidence. Interestingly, there was no developmental change in evidence-based counterargumentation. These findings suggest that the understanding of the belief–evidence relation precedes the understanding of the hypothesis–evidence relation. Moreover, children's evidence-based counterarguments demonstrate that children of this age can reflect on the inferential relation between beliefs and evidence. The third study investigated 5- and 6-year olds' ability to reflect on the relation between their epistemic states and evidence in an evidence evaluation paradigm. The findings revealed that half of the 5- and 6-year-olds were able to acknowledge their ignorance due to confounded evidence. Furthermore, 40% of the children provided justifications referring to the uncertainty due to two potential hypotheses at least once out of three times. In this respect, the basic ability to reflect on the hypothesis–evidence relation is present in 5 and 6 years of age.

This thesis contributes to the existing literature by bringing together two related research lines; namely, scientific reasoning and causal reasoning. The findings of this thesis showed that preschoolers' not only have an implicit understanding of the informativeness of evidence, but they already have basic foundational abilities for scientific reasoning. Future studies should continue to investigate in more detail the development of early scientific reasoning abilities in preschoolers across different epistemic activities and different domains together with examining underlying psychological mechanisms and related cognitive abilities in order to increase our understanding of the development of early scientific reasoning abilities.

References

Alvarez, A. L., & Booth, A. E. (2015). Preschoolers prefer to learn causal information. *Frontiers in Psychology, 6*, 1–5. doi: 10.3389/fpsyg.2015.00060

Alvarez, A., & Booth, A. E. (2016). Exploring individual differences in preschoolers' causal stance. *Developmental Psychology, 52*(3), 411–422. doi: 10.1037/dev0000085

Amsel, E., & Brock, S. (1996). The development of evidence evaluation skills. *Cognitive Development, 11*, 523–550. doi: 10.1016/S0885-2014(96)90016-7

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. doi: 10.1016/j.cogdev.2008.09.002

Ash, A., Torrance, N., Lee, E., & Olson, D. R. (1993). The development of children's understanding of the evidence for beliefs. *Educational Psychology, 13,* 371-384. doi: 10.1080/0144341930130313

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*, 131–144. doi: 10.1016/S0732-118X(02)00005-3

Baker, L., Mackler, K., Sonnenschein, S., & Serpell, R. (2001). Parents' interactions with their first-grade children during storybook reading and relations with subsequent home reading activity and reading achievement. *Journal of School Psychology, 39*(5), 415–438. doi: 10.1016/S0022-4405(01)00082-6

Beschorner, B., & Hutchison, A. (2013). iPads as a literacy teaching tool in early childhood. *International Journal of Education in Mathematics, Science and Technology, 1*(1), 16–24.

Bindra, D., Clarke, K. A., & Shultz, T. R. (1980). Understanding predictive relations of necessity and sufficiency in formally equivalent "causal" and "logical" problems. *Journal of Experimental Psychology: General*, *109*(4), 422–443. doi: 10.1037/0096-3445.109.4.422

Bishop, G. F., Tuchfarber, A. J., & Oldendick, R. W. (1986). Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly*, *50*(2), 240–250. doi: 10.1086/268978

Bonawitz, E. B., Griffiths, T. L., Schulz, L., Sun, R., & Miyake, N. (2006). Modeling cross-domain causal learning in preschoolers as Bayesian inference. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, *28*(28), 89–94.

Bonawitz, E. B., van Schijndel, T. J., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*, 215–234. doi: 10.1016/j.cogpsych.2011.12.002

Bonawitz, E., Fisher, A., & Schulz, L. (2012). Teaching 3.5-year-olds to revise their beliefs given ambiguous evidence. *Journal of Cognition and Development, 13*(2), 266–280. doi: 10.1080/15248372.2011.577701

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: NY Science Editions.

Bruner, J. S., Jolly, A., & Sylva, K. (Eds.). (1976). Play: Its role in development and education. Harmondsworth, Middx: Penguin.

Buchsbaum, D., Bridgers, S., Weisberg, D. S., & Gopnik, A. (2012). The power of possibility: Causal learning, counterfactual reasoning, and pretend play.

Philosophical Transactions of the Royal Society of London B: Biological Sciences, *367*, 2202–2212. doi:10.1098/rstb.2012.0122

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331–340. doi: 10.1016/j.cognition.2010.12.001

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). New York: Academic Press.

Bullock, M., Sodian, B., & Koerber, S. (2009). Doing experiments and understanding science: Development of scientific reasoning from childhood to adulthood. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood: Findings from a 20 year longitudinal study* (pp. 173–197). New York, NY: Psychology Press.

Bullock, M., & Ziegler, A. (1999). Scientific reasoning: Developmental and individual differences. In F. E. Weinert & W. Schneider (Eds.), *Individual development from 3 to 12: Findings from the Munich Longitudinal Study* (pp. 38–54). Cambridge: Cambridge University Press.

Bus, A. G., Van Ijzendoorn, M. H., & Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research*, *65*(1), 1–21. doi: 10.3102/00346543065001001

Carey, S. (1985). Conceptual change in childhood. Cambridge, MA: MIT Press/Bradford Books.

Carey, S., & Spelke, E. S. (1994). Domain-specific knowledge and conceptual change. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity*

in cognition and culture (pp. 169–200). New York: Cambridge University Press. doi: 10.1017/CBO9780511752902.008

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. doi: 10.1111/1467-8624.00081

Chen, S. (2009). Shadows: Young Taiwanese children's views and understanding. *International Journal of Science Education, 31*(1), 59–79. doi: 10.1080/09500690701633145

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*(2), 367–405. doi: 10.1037/0033-295X.104.2.367

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*(4), 391–416. doi: 10.1016/0010-0285(85)90014-3

Cheng, P. W., Holyoak, K. J., Nisbett, R., & Oliver, L. (1989). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18*(3), 293–331. doi: 10.1016/0010-0285(86)90002-2

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58*(4), 545–567. doi: 10.1037/0022-3514.58.4.545

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*(1), 1–49. doi: 10.3102/00346543063001001

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition, 120*, 341–349. doi:10.1016/j.cognition.2011.03.003

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, 43*(1), 75–96. doi: 10.1080/15391523.2010.10782562

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. *British Journal of Developmental Psychology*, *29*(3), 409–424. doi: 10.1348/026151010X496906

Cultice, J. C., Somerville, S. C., & Wellman, H. M. (1983). Preschoolers' memory monitoring: Feeling-of-knowing judgments. *Child Development, 54*(6), 1480–1486. doi: 10.2307/1129810

Dean Jr., D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education*, *91*(3), 384–397. doi: 10.1002/sce.20194

Doherty, M. (2009). *Theory of mind*. Hove: Psychology Press.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, *17*(3), 397–434. doi: 10.1207/s15516709cog1703_3

Dunbar, K., & Fugelsang, J. (2005). Scientific thinking and reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 705–725). Cambridge University Press.

Dunbar, K., & Klahr, D. (1989). Developmental differences in scientific discovery processes. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon*, 109–143. New Jersey: Lawrence Erlbaum Associates, Inc.

Dunham, P., Dunham, F., & O'Keefe, C. (2000). Two- year- olds' sensitivity to a parent's knowledge state: Mind reading or contextual cues? *British Journal of Developmental Psychology*, *18*(4), 519–532. doi: 10.1348/026151000165832

Ebersbach, M., & Resing, W. C. M. (2007). Shedding new light on an old problem: The estimation of shadow sizes in children and adults. *Journal of Experimental Child Psychology, 97*(4), 265–285. doi: 10.1016/j.jecp.2007.02.002

Erb, C. D., & Sobel, D. M. (2014). The development of diagnostic reasoning about uncertain events between ages 4–7. *PloS ONE*, *9*(3), e92285. doi: 10.1371/journal.pone.0092285

Estes, D., Wellman, H. M., & Woolley, J. D. (1989). Children's understanding of mental phenomena. *Advances in Child Development and Behavior*, *22*, 41–87. doi: 10.1016/S0065-2407(08)60412-7

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. doi:10.1016/j.cogdev.2011.10.002

Fernie, D. E., & DeVries, R. (1990). Young children's reasoning in games of nonsocial and social logic:"Tic Tac Toe" and a "Guessing Game". *Early Childhood Research Quarterly*, *5*(4), 445–459. doi: 10.1016/0885-2006(90)90013-Q

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, 2(3),* 28–45. doi: 10.14786/flr.v2i2.96

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822–15826. doi: 10.1073/pnas.232472899

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist, 34*(10), 906–911. doi: 10.1037/0003-066X.34.10.906

Flavell , J. H., & Wellman, H. M. (1977). Metamemory . In R. V. Kail & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 3–33). Hillsdale, NJ: Erlbaum.

Fritzley, V. H. (2006). *Questioning the Questioning of Children: The Effect of Questions on Children's Verbal Responses* (Unpublished doctoral dissertation). Queen's University, Canada.

Galyer, K. T., & Evans, I. M. (2001). Pretend play and the development of emotion regulation in preschool children. *Early Child Development and Care, 166*(1), 93–108. doi: 10.1080/0300443011660108

Ghanem, C., Kollar, I., Fischer, F., Lawson, T. R., & Pankofer, S. (2017). How do social work novices and experts solve professional problems? A micro-analysis of epistemic activities and the use of evidence. *European Journal of Social Work, 21*(1), 3–19. doi: 10.1080/13691457.2016.1255931

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*(5), 431–436. doi: 10.1111/1467-9280.00476

Gooch, D., Thompson, P., Nash, H. M., Snowling, M. J., & Hulme, C. (2016). The development of executive function and language skills in the early school years. *Journal of Child Psychology and Psychiatry, 57*(2), 180–187. doi: 10.1111/jcpp.12458

Gopnik, A. (1996). The scientist as child. *Philosophy of Science, 63*(4), 485–514. doi: 10.1086/289970

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science, 337*(6102), 1623–1627. doi: 10.1126/science.1223416

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*, *111*(1), 3–32. doi: 10.1037/0033-295X.111.1.3

Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, *24*(2), 87–92. doi: 10.1177/0963721414556653

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, *8*(8), 371–377. doi: 10.1016/j.tics.2004.06.005

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *71*(5), 1205-1222. doi: 10.1111/1467-8624.00224

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. doi: 10.1037/0012-1649.37.5.620

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*(1-2), 145–171. doi: 10.1111/j.1468-0017.1992.tb00202.x

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). New York: Cambridge University Press. doi: 10.1017/CBO9780511752902.011

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 1*38*(6), 1085–1108. doi: 10.1037/a0028044

Gopnik, A., Meltzoff, A. N., & Bryant, P. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. *Child Development*, *62*(1), 1–22. doi: 10.2307/1130701

Gweon, H., & Schulz, L. (2008). Stretching to learn: Ambiguous evidence and variability in preschoolers' exploratory play. In B. C. Love, K. McRae, & M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 570–574). *Austin, TX: Cognitive Science Society*.

Hahn, U., & Oaksford, M., (2012). Rational argument. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 277–300). Oxford, UK: Oxford University Press. doi: 10.1093/oxfordhb/9780199734689.013.0015

Harris, P. L., Ronfard, S., & Bartz, D. (2017). Young children's developing conception of knowledge and ignorance: work in progress. *European Journal of Developmental Psychology*, *14*(2), 221–232. doi: 10.1080/17405629.2016.1190267

Hatano, G., & Inagaki, K. (1994). Young children's naïve theory of biology. *Cognition*, *50*(1–3), 171–188. doi: 10.1016/0010-0277(94)90027-2

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900–1902. doi: 10.1126/science.1128898

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–135. doi: 10.1017/S0140525X0999152X

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

Institute of Medicine and National Research Council, (2012). Transforming the Workforce for Children Birth through Age 8: Unifying Foundation. Washington, DC: The National Academic Press.

Karmiloff-Smith, A., & Inhelder, B. (1974). If you want to get ahead, get a theory. *Cognition, 3*(3), 195–212. doi: 10.1016/0010-0277(74)90008-0

Keil, F. C. (1995). The growth of causal understandings of natural kinds. In D. Sperber & D. Premack, & A. J. Premack (Eds.), *Symposia of the Fyssen Foundation. Causal cognition: A multidisciplinary debate* (pp. 234–267). New York: Clarendon Press/Oxford University Press.

Khishfe, R., & Abd- El- Khalick, F. (2002). Influence of explicit and reflective versus implicit inquiry- oriented instruction on sixth graders' views of nature of science. *Journal of Research in Science Teaching, 39*(7), 551–578. doi: 10.1002/tea.10036

Kim, S., Paulus, M., Sodian, B., & Proust, J. (2016). Young children's sensitivity to their own ignorance in informing others. *PloS ONE*, *11*(3), e0152595. doi: 10.1371/journal.pone.0152595

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

Klahr, D., Dunbar, K., & Fay, A. L. (1990). Designing good experiments to test bad hypotheses. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation* (pp. 355–402). San Mateo, CA: Morgan Kaufmann.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*(1), 111–146. doi: 10.1006/cogp.1993.1003

Klayman, J. & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–228. doi: 10.1037/0033-295X.94.2.211

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. doi: 10.1111/cdev.12298

Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss Journal of Psychology/Schweizerische Zeitschrift für Psychologie/Revue Suisse de Psychologie*, *64*(3), 141–152. doi: 10.1024/1421-0185.64.3.141

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge: MIT Press.

Köksal Tuncer, Ö., Sodian, B., & Saffran, A. (2017). [Preschoolers's evidence generation and verbal counterarguments in evidence-based argumentation task]. Unpublished raw data.

Kuhn, D. (1988). Introduction. In H. Beilin (Ed.), *The development of scientific thinking skills* (pp. 3–12). San Diego: Academic Press, Inc.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, *96*(4), 674–689. doi: 10.1037/0033-295X.96.4.674

Kuhn, D. (1991). *The skills of argument.* New York: Cambridge University Press. doi: 10.1017/CBO9780511571350

Kuhn, D. (1992). Piaget's child as scientist. In H. Beilin & P. Pufall (Eds.), *Piaget's theory: Prospects and possibilities* (pp. 185–208). Hillsdale, N.J.: Erlbaum.

Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science*, *9*(5), 178–181. doi: 10.1111/1467-8721.00088

Kuhn, D. (2010). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (pp. 497–523). doi: 10.1002/9781444325485.ch19

Kuhn, D., Amsel, E., O'Loughlin, M., Schauble, L., Leadbeater, B., & Yotive, W. (1988). *The development of scientific thinking skills*. San Diego, CA: Academic Press.

Kuhn, D., & Franklin, S. (2006). The second decade: What develops (and how). In W. Damon & R. M. Lerner (Eds.). *Child and adolescent development. An advanced course* (pp. 517–550). Hoboken, NJ: John Wiley & Sons, Inc.

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., White, S. H., Klahr, D., & Carver, S. M. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development*, *60*(4), 1–157. doi: 10.2307/1166059

Kuhn, D., Iordanou, 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. doi: 10.1016/j.cogdev.2008.09.006

Kuhn, D., & Pearsall, S. (2000). Developmental origins of scientific thinking. *Journal of Cognition and Development*, *1*, 113–129. doi: 10.1207/S15327647JCD0101N_11

Kuhn, D., Pease, M., & Wirkala, C. (2009). Coordinating the effects of multiple variables: A skill fundamental to scientific thinking. *Journal of Experimental Child Psychology*, *103*(3), 268–284. doi: 10.1016/j.jecp.2009.01.009

Kuhn, D., Pennington, N., & Leadbeater, B. (1983). Adult reasoning in developmental perspective: The sample case of juror reasoning. In P. Baltes & O. Brim (Eds.), *Life span development and behavior* (pp. 157–195). New York: Academic Press.

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction*, *9*(4), 285–327. doi: 10.1207/s1532690xci0904_1

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development, 74*(5), 1245–1260. doi: 10.1111/1467-8624.00605

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science*, *16*(9), 678–683. doi: 10.1111/j.1467-9280.2005.01595.x

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. doi: 10.1037/0012-1649.43.1.186

Kushnir, T., Gopnik, A., Lucas, C., & Schulz, L. (2010). Inferring hidden causal structure. *Cognitive Science*, *34*(1), 148–160. doi: 10.1111/j.1551-6709.2009.01072.x

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*(8), 1134–1140. doi: 10.1177/0956797610376652

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*, *37*(1), 44–62. doi: 10.1002/(SICI)1098-2736(200001)37:1<44::AID-TEA4>3.0.CO;2-J

Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development*, *83*(1), 173–185. doi: 10.1111/j.1467-8624.2011.01691.x

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. doi: 10.1111/j.1467-8624.2010.01443.x

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology, 126,* 198–212. doi: 10.1016/j.jecp.2014.03.001

Leslie, A. M. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 119–148). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511752902.006

Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition, 108*(3), 732–739. doi: 10.1016/j.cognition.2008.06.013

Lorch Jr., R. F., 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. *Journal of Educational Psychology, 102*(1), 90–101. doi: 10.1037/a0017972

Lyons, K. E., & Ghetti, S. (2013). I don't want to pick! Introspection on uncertainty supports early strategic behavior. *Child Development, 84*(2), 726–736. doi: 10.1111/cdev.12004

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. doi: 10.1016/j.cognition.2013.12.010

Masnick, A. M., Klahr, D., & Morris, B. J. (2007). Separating signal from noise: Children's understanding of error and variability in experimental outcomes. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 3–26). Mawah, NJ: Erlbaum.

Masnick, A. M., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development, 79*(4), 1032–1048. doi: 10.1111/j.1467-8624.2008.01174.x

Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning and Instruction, 29*, 43-55. doi: 10.1016/j.learninstruc.2013.07.005

Michalsky, T., Mevarech, Z. R., & Haibi, L. (2009). Elementary school children reading scientific texts: Effects of metacognitive instruction. *The Journal of Educational Research, 102*(5), 363–376. doi: 10.3200/JOER.102.5.363-376

Morris, B. J., Croker, S., Masnick, A. M., & Zimmerman, C. (2012). The emergence of scientific reasoning. In H. Kloos, B. J. Morris & J. L. Amaral (Eds.), *Current topics in children's learning and cognition* (pp. 61–82). Croatia: InTech.

Moore, C., & D'Entremont, B. (2001). Developmental changes in pointing as a function of attentional focus. *Journal of Cognition and Development, 2*(2), 109–129. doi: 10.1207/S15327647JCD0202_1

Moeller, A. C., Sodian, B., & Hussman, H. (in progress). Promotion of preschoolers' scientific reasoning abilities with educational technology.

Müller, U., Jacques, S., Brocki, K., & Zelazo, P. D. (2009). The executive functions of language in preschool children. In A. Winsler, C. Fernyhough, & I. Montero (Eds.), *Private speech, executive functioning, and the development of verbal self-regulation.* New York: Cambridge University Press.

National Research Council. (2013). *Next Generation Science Standards: For States, By States.* Washington, DC: The National Academic Press. doi: 10.17226/18290

Notaro, P. C., Gelman, S. A., & Zimmerman, M. A. (2001). Children's understanding of psychogenic bodily reactions. *Child Development, 72*(2), 444–459. doi: 10.1111/1467-8624.00289

Okasha, S. (2016). *Philosophy of science: Very short introduction*. New York: Oxford University Press. doi: 10.1093/actrade/9780198745587.001.0001

O'Leary, A. P., & Sloutsky, V. M. (2017). Carving metacognition at its joints: Protracted development of component processes. *Child Development*, *88*(3), 1015–1032. doi: 10.2307/1131839

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development, 67*(2), 659–677. doi: 10.2307/1131839

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. doi: 10.1037/dev0000260

Paulus, M., Proust, J., & Sodian, B. (2013). Examining implicit metacognition in 3.5-year-old children: an eye-tracking and pupillometric study. *Frontiers in Psychology*, *4*, 1–7. doi: 10.3389/fpsyg.2013.00145

Patel, V. L., Arocha, J. F., & Zhang, J. (2005). Thinking and reasoning in medicine. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 727–750). New York: Cambridge University Press.

Penner, D. E., & Klahr, D. (1996). The interaction of domain-specific knowledge and domain-general discovery strategies: a study with sinking objects. *Child Development*, *67*(6), 2709-2727. doi: 10.2307/1131748

Piaget, J. (1962). *Play, dreams and imitation in children*. New York: Norton.

Piaget, J., & Garcia, R. (1989). *Psychogenesis and the history of science*. Columbia University Press.

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. doi: 10.1007/s11409-013-9097-4

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. doi: 10.1080/09500693.2013.776192

Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology*, *31*(2), 153–179. doi:10.1111/j.2044-835X.2012.02082.x

Rakoczy, H. (2006). Pretend play and the development of collective intentionality. *Cognitive Systems Research*, *7*(2–3), 113–127. doi: 10.1016/j.cogsys.2005.11.008

Rakoczy, H. (2007). Play, games, and the development of collective intentionality. *New Directions for Child and Adolescent Development*, *2007*(115), 53–67. doi: 10.1002/cd.182

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology*, *44*(3), 875–881. doi: 10.1037/0012-1649.44.3.875

Reichenbach, H. (1956). *The direction of time.* Berkeley: University of California Press.

Rohwer, M., Kloo, D., & Perner, J. (2012). Escape from metaignorance: How children develop an understanding of their own lack of knowledge. *Child Development*, *83*(6), 1869–1883. doi: 10.1111/j.1467-8624.2012.01830.x

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. doi: 10.2307/1131459

Rozencwajg, P. (2003). Metacognitive factors in scientific problem-solving strategies. *European Journal of Psychology of Education, 18*(3), 281–294. doi: 10.1007/BF03173249

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928. doi: 10.1126/science.274.5294.1926

Saffran, A., Barchfeld, P., Sodian, B., & Alibali, M.W. (2017, September). Die Interpretation von Kovariationsdaten im Vor- und Grundschulalter – Einfluss der Symmetrie der Variablen [Preschool and elementary school children's interpretation of covariation data – The effect of symmetry of variables]. In B. Sodian & A. Saffran (Chairs), *Wissenschaftliches Denken im Vor- und Grundschulalter.* Symposium conducted at the „Gemeinsame Tagung der Fachgruppen Entwicklungspsychologie und Pädagogische Psychologie, Münster, Germany.

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. doi: 10.1037/dev0000203

Schanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology, 37*B, 1–21. doi: 10.1080/14640748508402082

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49*(1), 31–57. doi: 10.1016/0022-0965(90)90048-D

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*(1), 102–119. doi: 10.1037/0012-1649.32.1.102

Schmidt, M. F., Hardecker, S., & Tomasello, M. (2016). Preschoolers understand the normativity of cooperatively structured competition. *Journal of Experimental Child Psychology*, *143*, 34–47. doi: 10.1016/j.jecp.2015.10.014

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education*, *2*(3), 114–121. doi: 10.1111/j.1751-228X.2008.00041.x

Schneider, W., & Sodian, B. (1988). Metamemory-memory behavior relationships in young children: Evidence from a memory-for-location task. *Journal of Experimental Child Psychology*, *45*(2), 209–233. doi: 10.1016/0022-0965(88)90030-6

Scholl, B. J., & Leslie, A. M. (1999). Modularity, development and "'theory of mind'. *Mind & Language*, *14*(1), 131–153. doi: 10.1111/1468-0017.00106

Schulz, L. (2012). The origins of inquiry: inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*(7), 382–389. doi: 10.1016/j.tics.2012.06.004

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. doi: 10.1037/0012-1649.43.4.1045

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*, *43*(5), 1124–1139. doi: 10.1037/0012-1649.43.5.1124

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, *40*(2), 162–176. doi: 10.1037/0012-1649.40.2.162

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*(3), 322–332. doi: 10.1111/j.1467-7687.2007.00587.x

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development*, *77*(2), 427–442. doi: 10.1111/j.1467-8624.2006.00880.x

Shaklee, H., & Mims, M. (1981). Development of rule use in judgments of covariation between events. *Child Development*, *52*(1), 317–325. doi: 10.2307/1129245

Shaklee, H., & Paszek, D. (1985). Covariation judgment: Systematic rule use in middle childhood. *Child Development*, *56*(5), 1229–1240. doi: 10.2307/1130238

Siler, S., Mowery, D., Magaro, C., Willows, K., & Klahr, D. (2010). Comparison of a computer-based to hands-on lesson in experimental design. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems. ITS 2010. Lecture notes in computer science, 6095* (pp. 408–410). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-13437-1_86

Simon, H. A., & Lea, G. (1974). Problem solving and rule induction.In H. Simon (Ed.), *Models of thought* (pp. 329–346). New Haven, CT: Yale University Press.

Singer, D. G., Golinkoff, R. M., & Hirsh-Pasek, K. (2006). *Play= Learning: How play motivates and enhances children's cognitive and social-emotional growth*. Oxford University Press. doi: 10.1093/acprof:oso/9780195304381.001.0001

Smith, C., Carey, S., & Wiser, M. (1985). On differentiation: A case study of the development of the concepts of size, weight, and density. *Cognition*, *21*(3), 177–237. doi: 10.1016/0010-0277(85)90025-3

Sobel, D. M., & Buchanan, D. W. (2009). Bridging the gap: Causality-at-a-distance in children's categorization and inferences about internal properties. *Cognitive Development*, *24*(3), 274–283. doi: 10.1016/j.cogdev.2009.03.003

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*, *18*(5), 556–576. doi:10.1080/15248372.2017.1387117

Sobel, D. M., & Sommerville, J. A. (2010). The importance of discovery in children's causal learning from interventions. *Frontiers in Psychology, 1*(176), 1–7. doi: 10.3389/fpsyg.2010.00176

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. doi: 10.1207/s15516709cog2803_1

Sodian, B. (2005). Theory of mind. The case for conceptual development. In W. Schneider, R. Schumann-Hengsteler, & B. Sodian (Eds.), *Young children's cognitive development. Interrelationships among executive functioning, working memory, verbal ability, and theory of mind* (pp. 95–130). Hillsdale, NJ: Erlbaum.

Sodian, B. (in press). The development of scientific thinking in preschool and elementary school age. A conceptual model. In F. Fischer, C. Chinn, K. Engelmann & J. Osborne (Eds.), *Interplay of domain-specific and domain-general aspects of scientific reasoning and argumentation skills*. Taylor & Francis.

Sodian, B., & Bullock, M. (2008). Scientific reasoning—Where are we now? *Cognitive Development*, *23*(4), 431–434. doi: 10.1016/j.cogdev.2008.09.003

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*, *2*(3), 111–113. doi: 10.1111/j.1751-228X.2008.00040.x

Sodian, B., Kristen-Antonow, S., & Koerber, S. (2016, July). Theory of mind predicts scientific reasoning. A longitudinal study from preschool to elementary school age. Paper presented at the International Congress of Psychology, Yokohama, Japan.

Sodian, B., Schneider, W., & Perlmutter, M. (1986). Recall, clustering, and metamemory in young children. *Journal of Experimental Child Psychology*, *41*(3), 395–410. doi: 10.1016/0022-0965(86)90001-9

Sodian, B., Thoermer, C., Kircher, E., Grygier, P., & Günther, J. (2002). Vermittlung von Wissenschaftsverständnis in der Grundschule [Teaching nature of science in elementary school]. *Zeitschrift für Pädagogik, Beiheft 45*, 192–206.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*(4), 753–766. doi: 10.2307/1131175

Tamis-Le Monda, C. S., Shannon, J. D., Cabrera, N. J., & Lamb, M. E. (2004). Fathers and mothers at play with their 2- and 3-year-olds: Contributions to language and cognitive development. *Child Development*, *75*(6), 1806–1820. doi: 10.1111/j.1467-8624.2004.00818.x

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, *51*(1), 1–10. doi: 10.2307/1129583

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. John Wiley & Sons.

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. doi: 10.1007/s11251-015-9344-y

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*, 190–200. doi: 10.1016/j.lindif.2016.06.006

Van Reet, J., Green, K. F., & Sobel, D. M. (2015). Preschoolers' theory-of-mind knowledge influences whom they trust about others' theories of mind. *Journal of Cognition and Development*, *16*(3), 471–491. doi: 10.1080/15248372.2014.892875

Veenman, M. V., Elshout, J. J., & Busato, V. V. (1994). Metacognitive mediation in learning with computer-based simulations. *Computers in Human Behavior*, *10*(1), 93–106. doi: 10.1016/0747-5632(94)90031-0

Veenman, M. V., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences*, *15*(2), 159–176. doi: 10.1016/j.lindif.2004.12.001

Veenman, M. V., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction*, *14*(1), 89–109. doi: 10.1016/j.learninstruc.2003.10.004

Veenman, M. V., Van Hout-Wolters, B. H., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*(1), 3–14. doi: 10.1007/s11409-006-6893-0

Vygotsky, L. (1962). *Thought and language*. Cambridge, MA: MIT Press.

Vygotsky, L. S. (1978). The role of play in development. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society*, (pp. 92–104). Cambridge, MA: Harvard University Press.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, *82*(1), 27–58. doi: 10.1016/S0010-0277(01)00141-X

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2013). Explaining to Others Prompts Children to Favor Inductively Rich Properties. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society, Germany*, *35*(35), 1558–1563.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. doi: 10.1080/14640746808400161

Wellman, H. M. (2011). Developing a theory of mind. In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development*, (pp. 258–284). Wiley-Blackwell, Oxford, UK. doi: 10.1002/9781444325485.ch10

Wellman, H. M., & Estes, D. (1986). Early understanding of mental entities: A reexamination of childhood realism. *Child Development*, *57*(4), 910–923. doi: 10.2307/1130367

Wente, A. O., Kimura, K., Walker, C. M., Banerjee, N., Fernández Flecha, M., MacDonald, B., . . . Gopnik, A. (2017). Causal learning across culture and socioeconomic status. *Child Development, online version of record published before inclusion in an issue.* doi: 10.1111/cdev.12943

Wertheimer, M. (1945). *Productive thinking*. New York: Harper & Row.

White, B., & Frederiksen, J. (2000). Metacognitive facilitation: An approach to making scientific inquiry accessible to all. In J. Minstrell & E. van Zee (Eds.), *Inquiring*

into Inquiry Learning and Teaching in Science (pp. 331–370). Washington, DC: American Association for the Advancement of Science.

Whitebread, D., & O'Sullivan, L. (2012). Preschool children's social pretend play: supporting the development of metacommunication, metacognition and self-regulation. *International Journal of Play*, *1*(2), 197–213. doi: 10.1080/21594937.2012.693384

Wu, R., Gopnik, A., Richardson, D. C., & Kirkham, N. Z. (2011). Infants learn about objects from statistics and people. *Developmental Psychology, 47*(5), 1220–1229. doi: 10.1037/a0024023

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012–5015. doi: 10.1073/pnas.0704450105

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science, 10(3),* 288–297. doi: 10.1111/j.1467-7687.2007.00590.x

Zelazo, P. D., Müller, U., Frye, D., Marcovitch, S., Argitis, G., Boseovski, J., ... & Carlson, S. M. (2003). The development of executive function in early childhood. *Monographs of the Society for Research in Child Development*, *68*(3), i–151. doi: 10.1111/j.0037-976X.2003.00261.x

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20(1),* 99–149. doi: 10.1006/drev.1999.0497

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172–223. doi: 10.1016/j.dr.2006.12.001

Zion, M., Michalsky, T., & Mevarech, Z. R. (2005). The effects of metacognitive instruction embedded within an asynchronous learning network on scientific

inquiry skills. *International Journal of Science Education, 27*(8), 957–983. doi: 10.1080/09500690500068626

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. doi: 10.1016/j.learninstruc.2007.07.001