Children's Evaluations of Observable vs. Unobservable Properties During Scientific Reasoning

Taneisha Vilma
taneisha.vilma@gmail.com

Follow this and additional works at: https://huskiecommons.lib.niu.edu/allgraduate-thesesdissertations

Recommended Citation

This Dissertation/Thesis is brought to you for free and open access by the Graduate Research & Artistry at Huskie Commons. It has been accepted for inclusion in Graduate Research Theses & Dissertations by an authorized administrator of Huskie Commons. For more information, please contact jschumacher@niu.edu.
ABSTRACT

CHILDREN’S EVALUATIONS OF OBSERVABLE VS. UNOBSERVABLE PROPERTIES DURING SCIENTIFIC REASONING

Taneisha Vilma, Ph.D.
Department of Psychology
Northern Illinois University, 2023
Dr. Bradford H. Pillow, Director

Science inquiry involves reasoning and drawing conclusions about properties or constructs that may not be directly observed and are not immediately verifiable. Prior scientific reasoning investigations have examined children’s reasoning about observable items and immediately verifiable conclusions. The current investigation examined children’s evaluations of the certainty of conclusions drawn from evidence when reasoning about both immediately observable properties and unobservable properties with the same task. Kindergarten, first grade, third grade, and adult participants (N = 70) were presented with an online experimental procedure. The procedure included two conditions (Observable vs. Unobservable) each involved reasoning about three levels of evidence (conclusive vs. inconclusive vs. guess). The Observable condition assessed participant’s ability to reason about observable properties using animated cues, and the Unobservable condition assessed children’s ability to reason about unobservable properties using animated cues. A 4×2×3 ANOVA (Age × Condition × Evidence Level) was conducted, with Age (kindergarten, first grade, third grade, adults) as a between-subjects factor and Condition (observable vs. unobservable) and Evidence Level (conclusive vs. inconclusive vs. guess) as within-subjects factors. For the Observable condition, the ANOVA revealed a
significant effect of Evidence Level. Kindergarten children gave significantly higher certainty ratings for the conclusive compared to the inconclusive trials. In contrast, first graders, third graders, and adults, reported significantly higher certainty ratings for the conclusive trials compared to the inconclusive and conclusive trials compared to the guess trials. The Unobservable condition revealed a significant effect of Evidence Level and a significant Age × Evidence Level interaction. Kindergarten and first grade children demonstrated difficulty differentiating the three levels of evidence resulting in no significant differences. However, third grade children reported significantly higher certainty ratings for conclusive compared to inconclusive trials and adults reported significantly higher certainty ratings for conclusive compared to inconclusive trials and conclusive compared to guess items. Results indicate a difference in reasoning between observable and unobservable evidence across levels of evidence. By kindergarten, children can begin to distinguish different levels of evidence with certainty ratings when reasoning about observable properties. The comparison of the Unobservable condition demonstrated improved performance with age, it was not until third grade that children could successfully reason about unobservable properties. The participants’ certainty ratings suggest a progressive differentiation between levels of evidence with increased age, especially for the unobservable outcomes. The current investigation promotes greater understanding of the extent of abilities for scientific reasoning across age groups.
CHILDREN’S EVALUATIONS OF OBSERVABLE VS. UNOBSERVABLE PROPERTIES DURING SCIENTIFIC REASONING

BY

TANEISHA VILMA
©2023 Taneisha Vilma

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

Doctoral Director:
Bradford H. Pillow, Ph.D.
ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my dissertation chair, Dr. Brad Pillow for his patience, guidance, and recommendations during the course of this project and throughout my graduate experience. My gratitude extends to my committee members Drs. Anne Britt, Leslie Matuszewich, and Katja Wiemer for their insightful comments and suggestions. Special thanks to the faculty members and classmates for their feedback and assistance at various stages of the research project. I would also like to thank the families and study participants for their cooperation. To my mentors, family members, and friends I truly appreciate your encouraging words and unwavering support.
DEDICATION

I dedicate this dissertation to my mom and dad, thank you.
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td>viii</td>
</tr>
</tbody>
</table>

### Chapter

1. **INTRODUCTION AND LITERATURE REVIEW** ........................................ 1

   - Conceptualizations of scientific reasoning ....................................... 4
   - Scientific Discovery as Dual Search (SDDS) model ................................ 7

2. **Research on Scientific Reasoning Development** ................................ 8

   - Experimentation ....................................................................................... 8
   - Experimentation - Late competence ......................................................... 8
   - Experimentation - Early competence ...................................................... 12

3. **Evidence evaluation investigations** ................................................. 17

   - Evidence evaluation - Late competence .................................................. 18
   - Evidence evaluation - Early competence ................................................. 20

4. **Evidence evaluation during deductive reasoning** ................................. 31

5. **Evaluation of reasoning** ..................................................................... 35

6. **Reasoning about observable and unobservable properties** ...................... 39

   - Reasoning about observable properties ............................................... 39
   - Reasoning about unobservable properties ............................................. 40
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. CURRENT STUDY AIMS</td>
<td>43</td>
</tr>
<tr>
<td>3. METHOD</td>
<td>49</td>
</tr>
<tr>
<td>Participants</td>
<td>49</td>
</tr>
<tr>
<td>Materials</td>
<td>50</td>
</tr>
<tr>
<td>Procedure</td>
<td>50</td>
</tr>
<tr>
<td>Scoring</td>
<td>56</td>
</tr>
<tr>
<td>4. RESULTS</td>
<td>57</td>
</tr>
<tr>
<td>5. DISCUSSION</td>
<td>65</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>89</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>100</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Mean Certainty Ratings (and Standard Deviation) by age and evidence level for the Observable and Unobservable conditions</td>
<td>62</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Mean certainty ratings by age and evidence level for the Observable condition</td>
<td>63</td>
</tr>
<tr>
<td>2. Mean certainty ratings by age and evidence level for the Unobservable condition</td>
<td>63</td>
</tr>
</tbody>
</table>
## LIST OF APPENDICES

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. EVIDENCE EVALUATION TASK</td>
<td>100</td>
</tr>
<tr>
<td>B. CORRESPONDING FEATURES ACROSS TASKS</td>
<td>105</td>
</tr>
</tbody>
</table>
The development of scientific reasoning is of considerable interest to developmental and
cognitive psychologists as well as science educators (Carey & Smith, 1993; Klahr et al., 2011;
Zimmerman, 2000). The general objective of science is to increase knowledge by systematically
gathering and interpreting evidence. Frequently, scientists use evidence and reasoning to draw
conclusions about properties and events that are not yet known and cannot be directly observed.
For example, physicists examine the interactions of matter and energy while at times studying
subatomic particles (Gottfried & Weisskopf, 1986; Tipler & Llewellyn, 2003). Through the
analysis of fossils or in the rare case preservation of soft tissue, paleontologists make inferences
about ancient organisms and draw conclusions about their anatomy, behaviors, and environment
(Allison & Briggs, 1993; Jablonski & Shubin, 2015). Additionally, psychologists employ
unobservable constructs such as attitudes, intelligence, and metacognition to explain human
cognition and behavior (Croizet & Dutrévis, 2004; Durik et al., 2008; Pillow & Pearson, 2015).
In general, scientists apply logical reasoning processes to the evidence available to make
inferences and draw appropriate conclusions. Thinking critically about science involves
evaluating the soundness of research procedures, the conclusiveness of evidence, and the validity
of reasoning in scientific inquiries. The current investigation seeks to examine children’s
evaluations of the certainty of conclusions drawn from evidence during scientific reasoning tasks
when reasoning about both immediately observable properties and unobservable properties.
There is a robust literature on the development of scientific reasoning during childhood. However, empirical studies have revealed inconsistent findings concerning early childhood scientific reasoning competency. Some researchers have reported that successful scientific reasoning appears in late childhood or adolescence. For example, Piaget’s (Inhelder & Piaget, 1958) pioneering investigations revealed that children’s ability to isolate and control variables to conduct systematic experiments did not develop until approximately 11 years old. Likewise, Kuhn and colleagues argued that children exhibit deficiencies in understanding experimentation and correctly evaluating evidence (Kuhn, 2011; Kuhn et al., 1988; Kuhn & Pearsall, 2000; Kuhn & Pease, 2008). Thus, they concluded that children were unable to engage in scientific thinking until about 10 to 11 years old. Alternatively, some investigations have found that younger children can engage in scientific reasoning when given appropriate tasks (Croker & Buchanan, 2011; Piekny et al., 2014; Ruffman et al., 1993; Sodian et al., 1991). As a result, there remain conceptual and empirical discrepancies within the scientific reasoning development literature. These discrepancies will be further discussed in the following sections.

Despite the debate concerning the age of emergence of scientific reasoning skills, there is substantial agreement that experimentation and evidence evaluation are fundamental to scientific reasoning (Dunbar & Klahr, 1989; Kuhn et al., 1988; van der Graaf et al., 2016; Zimmerman, 2005). Scientific reasoning requires evaluating and reasoning about a situation, answering questions, or solving problems based on the present information (Klahr et al., 2019; Kuhn et al., 2008; Kuhn & Franklin, 2006; Zimmerman, 2000). This process of evaluating experiments and evidence is unlike answering questions about science concepts such as the shape of the earth.
(Sneider & Ohadi, 1998; Vosniadou et al., 2005) or tasks where children do not conduct experiments or evaluate evidence to determine their answer; for example, predicting the trajectory of falling objects (e.g., Kaiser et al., 1985). As a result, experimentation and evidence evaluation skills are required for advanced scientific reasoning.

The present investigation will examine children’s evaluation of the certainty of conclusions during scientific reasoning tasks. The goals of the study are to (a) compare children’s evaluations of the certainty of inferences about immediately observable properties with children’s evaluation of the certainty of inferences about unobservable properties, (b) compare children’s evaluation of the certainty of inferences when reasoning about varying levels of evidence, and (c) examine age differences in evaluations of certainty.

The current literature review will begin with an overview of the development of scientific reasoning. The purpose of this section is to define scientific reasoning and describe how scientific reasoning has been conceptualized in psychological research. Empirical investigations used to assess scientific reasoning abilities during childhood will then be discussed using the Scientific Discovery as Dual Search (SDDS) model as a guiding framework. The SDDS model provides an organized framework for discussing and investigating the principal components in science reasoning development. Investigations concerning children’s experimentation skills will then be reviewed, highlighting methodological differences and the contradictory conclusions from empirical studies. Next, research involving children’s evaluating of evidence will be emphasized, including studies that examine children’s ability to evaluate whether the presented information allows for a certain or uncertain conclusion. The purpose of this section is to present different methods used to examine children’s reasoning about evidence with corresponding
empirical results. Additionally, relevant studies that examine children’s evaluation of reasoning in contexts other than scientific inquiry will be discussed. This will specifically include studies that examine children’s ability to reflect on their own reasoning.

The final section of the review will consider children’s reasoning about observable and concrete properties compared to unobservable or non-obvious properties. Research shows that in some situations children can make appropriate inferences about unobservable properties, outcomes, or novel examples (Gelman & Wellman, 1990; Jaswal & Markman, 2019; Pillow, 1993). However, it is still unclear if children’s reasoning is influenced by the observability of the outcome that one is making inferences about (i.e., the conclusion). That is, do children’s evaluations of certainty differ when reasoning about observable vs. unobservable content? The primary interest of the proposed study is to compare children’s evaluations about reasoning about evidence and their ability to draw conclusions about properties that are observable compared to unobservable. Generally, the current investigation seeks to increase the understanding of early childhood logical thinking and scientific reasoning development.

**Conceptualizations of scientific reasoning**

Scientific reasoning is defined as reasoning or problem-solving skills involved in generating, testing, and revising hypotheses or theories, and the ability to reflect on this process (Koslowski, 1996; Kuhn & Franklin, 2006; Morris et al., 2012; Wilkening & Sodian, 2005; Zimmerman, 2000). Piaget’s investigations have informed views of the foundational skills for scientific reasoning (Inhelder & Piaget, 1958). Piaget’s research examined children’s ability to use the control-of-variables strategy (CVS) to identify causal relations. That is, Piaget’s investigations required children to manipulate and isolate variables, identify the causal relations
between variables, design conclusive experiments, and interpret evidence to identify a conclusive outcome from an inconclusive outcome. Young children did not engage in any systematic procedure to solve the problems. For example, during the balance scale task different weighted pegs are placed across two arms of the crossbar and are positioned at varying distances from the fulcrum. The task required children to coordinate the weight of the pegs and distance from the fulcrum on the right and left arms of the bar to remain in balance. This involved understanding the relationship between weight and distance (e.g., holding weight constant while varying distance, or vice versa) to predict the placement of the pegs to balance the scale. Children placed pegs in various positions unsystematically or focused on the weights of the pegs while ignoring the distance from the fulcrum. Children’s responses demonstrated that they did not implement a systematic strategy to select or arrange the weighted pegs on either side of the scale for it to balance. Inhelder and Piaget concluded that young children (aged 5-8) did not have the fundamental skills required for scientific thinking. Specifically, children are unable to coordinate multiple variables and understand the relation between their actions and the objects. It was suggested that reasoning abilities and procedural strategies improved with age. Consequently, scientific thinking did not develop until the formal operational stage of Piaget’s theory of cognitive development (approximately 11-years-old).

Contemporary approaches to examining scientific reasoning abilities have derived from Piaget’s early findings. Subsequent research has continued to examine children’s use of the control-of-variables strategy. It is understood that CVS is a prerequisite for making valid inferences about multivariable contexts, identifying confounded experiments, constructing unconfounded experiments, and interpreting indeterminate evidence (Chen & Klahr, 1999, 2008;
Kuhn, 2007, 2009; Schwichow et al., 2016). Unlike Piaget’s assertion of late CVS competence, others have found that children as young as 5 to 6 years old can implement CVS and evaluate or design unconfounded experiments (Chen & Klahr, 1999; Klahr & Chen, 2003). Kuhn and colleagues define scientific reasoning as the intentional and systematic coordination of theory and evidence (Kuhn, 2005, 2016; Kuhn et al., 1988). They proposed that the coordination of theory and evidence to be the single critical component to accurate scientific thinking. They suggest that scientific thinking entails a deliberate knowledge seeking that involves the ability to generate, test, and revise theories and to reflect on the process of knowledge acquisition and change (Kuhn & Franklin, 2006; Kuhn & Pearsall, 2000). Kuhn et al. suggested that young children (prior to age 10-11) cannot differentiate between evidence-based reasoning (i.e., using the data presented to draw conclusions) opposed to theory-based reasoning (i.e., relying on prior knowledge or background information to draw conclusions) (Kuhn et al., 1988, 1995; Kuhn & Pearsall, 2000). According to Kuhn (1989), the intentional and systematic coordination of theory and evidence is fundamental for accurate scientific thinking. That is, without this consistent awareness of this differentiation scientific reasoning is not feasible.

Kuhn and colleagues reported that children can successfully recognize theory and evidence as distinctive categories around 10 or 11-years-old (Kuhn, 2011; Kuhn & Pease, 2008). This corresponds with Piaget (1958), who determined that scientific reasoning emerges around 11 to 12-years of age. However, ensuing investigations have demonstrated extensive early childhood competency in scientific reasoning (Piekny et al., 2014; Piekny & Maehler, 2013; Ruffman et al., 1993; Sodian et al., 1991). Thus, a prominent debate persists regarding the emergence and extent of children’s scientific reasoning abilities.
Klahr and Dunbar (1988) defined scientific reasoning as problem solving involving a guided search and information-gathering processes. They proposed The Scientific Discovery as Dual Search (SDDS) model to investigate scientific reasoning development. The SDDS model has been applied not only to studying the development of scientific reasoning during childhood, but also to examining how working scientists conduct inquiries. The SDDS model prompted an expansion in empirical research about scientific thinking, predominantly focused on children’s abilities. The following sections will provide an overview of the scientific reasoning literature using the SDDS model as a guiding framework.

**Scientific Discovery as Dual Search (SDDS) model**

Klahr and Dunbar’s (1988) Scientific Discovery as Dual Search model provides a coherent framework for examining the fundamental components in scientific reasoning development. The model addresses the conceptual understanding as well as the cognitive skills involved in scientific reasoning (Joolingen & De Jong, 1997; Zimmerman, 2007). The integrated model consists of two major components: (a) hypothesis space which consists of two parts: hypothesis generation, experimentation, and (b) evidence evaluation (Dunbar & Klahr, 1989; Klahr et al., 2019; Klahr & Dunbar, 1988). Hypothesis generation involves constructing a new hypothesis based on prior beliefs or observed outcomes. Experimentation involves testing a hypothesis or producing a novel hypothesis based on changes in evidence. Evidence evaluation involves deliberating about conflicting hypotheses based on evidence presented. Tasks typically require participants to determine if an outcome is conclusive or inconclusive. Klahr and Dunbar assert that scientific reasoning is the result of integration of all three processes (Dunbar & Klahr, 1989; Klahr et al., 1993; Klahr, 2002; Klahr & Dunbar, 1988). A primary goal of the SDDS
model is to provide a framework to assess the conceptual and procedural skills involved in scientific reasoning. In the following section, the SDDS model will be used as a framework to discuss children’s scientific reasoning literature. This section will consider empirical investigations and highlight conceptual and methodological approaches to studying scientific reasoning. The SDDS model also provides a guide to empirically test specific skills of scientific reasoning independently or simultaneously. This section will also use the SDDS model as a guide to review empirical findings of the key components of the current study: experimentation and evidence evaluation investigations during childhood.

Research on Scientific Reasoning Development

Experimentation

*Experimentation - Late competence*

Experimentation involves testing a hypothesis or producing a novel hypothesis based on changes in evidence. Researchers have designed experimentation tasks that required children to reason about potential outcomes while controlling for a single variable at a time to assess understanding of experimentation (e.g., Kuhn, 1989; Schauble, 1996, 1990; Siegler & Liebert, 1975). Early developmental studies suggested that experimentation poses a conceptual and practical problem across development. The following section will briefly summarize experimentation studies that indicate late competency.

Building on the research tradition of Piaget, many science reasoning assessments concentrated on experimentation and investigation skills. Kuhn defines scientific reasoning as the deliberate coordination of theory and evidence (Kuhn, 1989). Kuhn and colleagues maintain that preadolescent children do not yet have the cognitive control of this coordination to
Successfully engage in scientific reasoning. Several empirical investigations have documented insufficient understanding of the distinction between theory and evidence (Kuhn et al., 1988, 1995; Kuhn & Pearsall, 2000). Young children demonstrated deficiencies across scientific reasoning skills such as the ability to correctly assess causal information, isolate and control variables one at a time, select a conclusive test, or design an effective experiment until approximately 11 years old (Kuhn, 2005; Kuhn et al., 1988, 1988; Kuhn & Pearsall, 2000; Schauble, 1996).

Kuhn et al. (1988) presented 6th and 9th grade students new information that was consistent or inconsistent with their prior beliefs. Students were interviewed to identify student’s prior beliefs about various foods and their association with health outcomes. Study questions were adjusted based on interview responses, resulting in a personalized series of questions for each participant. For example, pictures included a food item (e.g., carrot or chocolate cake) presented with a health outcome of an experiment character (e.g., healthy or has a cold). For each scenario, two characters presented contradictory beliefs (e.g., one believed chocolate cake caused colds and the other who presumed carrot cake caused colds) and participants were asked to decide who was correct. Then participants were to determine if the type of cake can make a difference, does not make a difference, or if they were unable to tell the relation to getting a cold based on the covariation evidence presented. Children had difficulty reasoning about causal variables when presented with the covariation evidence. Responses were often theory-based which relied on prior belief (e.g., chocolate cake has “sugar and a lot of bad stuff in it”) and not evidence-based which considered the patterns of observations presented. Additionally, children often ignored the evidence or modified their theories to align with the evidence, while failing to
acknowledge the discrepancy between their initial judgments and their subsequent responses. Kuhn et al. suggested that young children have difficulty during scientific reasoning because they are unable to differentiate between evidence-based reasoning (i.e., using the data presented to draw conclusions) opposed to theory-based reasoning (i.e., relying on prior knowledge or background information to draw conclusions (Kuhn et al., 1988, 1995; Kuhn & Pearsall, 2000).

In other studies, a hypothesis was presented and children were tasked with determining if there was support for a competing hypothesis and to recognize an alternative possibility associated with an outcome (Simon, 1989). Influenced by the work of Piaget, Siegler and Liebert (1975) examined children’s experimental design abilities. Students in 5th and 8th grade (aged 10-13) were instructed to find the correct configuration of four on/off switches to make an electric toy train run. The task was designed to assess if children could produce all possible combinations of the switches. Specifically, the task required students to generate all 16 possibilities to discover the correct combination that would make the train run. Results illustrated that few students (0% of fifth graders and only 10% of eighth graders) completed all correct switch configurations. Siegler and Liebert suggested that children were unable to engage in experimentation as they have difficulty manipulating and isolating variables without explicit instruction. Additionally, Siegler and Liebert argued that children lacked the ability to appropriately plan and implement the appropriate strategies to execute a controlled experimental design.

Schauble (1996) assessed the scientific reasoning abilities of 5th and 6th grade students (aged 10-12) with a physical system task. Students were to manipulate variables (e.g., object size, weight, start position) during a canal task and a spring task. The primary objective was to
determine variables that would influence travel time (e.g., maximize boat speed during canal task). Children received initial instruction on how to manipulate several variables to achieve the objective. For example, adjusting the depth of the model canal (from shallow to deep), the shape of the boat (circular, square, or diamond shaped), boat size (small or large), and boat weight (light or heavy). Results illustrated that children had difficulty controlling one variable at a time to generate an unconfounded test. Children also had difficulties correctly reasoning about causal mechanisms (e.g., boat size) and interpreting outcomes (e.g., boat speed). Additionally, most justifications reported were inferences linked to prior belief and not based on the patterns of evidence. Schaubles’ results demonstrated that children aged 10- to 12-years-old lack the necessary skills for scientific reasoning.

Kuhn and colleagues concluded that poor performance was predominantly due to the inability for children to discern theory and evidence (Kuhn, 1989, 2005, 2016; Kuhn et al., 1988). They suggested that children are unable to isolate and control variables one at a time, select an appropriate design or a conclusive test, evaluate causal relationships, or design unconfounded tests. Kuhn and her collaborators proposed that children’s errors may be attributed to a limited understanding of the basic objective of experimentation (Kuhn, 2005; Kuhn et al., 2000). Moreover, others claim that children may lack the proper metacognitive competence to identify and use appropriate strategies to achieve task goals during experimentation. Specifically, the differentiation of theory and evidence is reliant on the ability to understand and implement the appropriate strategy while coordinating between different sources of knowledge (Garcia-Mila & Andersen, 2007; Zimmerman, 2000, 2007). Kuhn and colleagues argue that since children lack this active coordination, theory and evidence is joined as a single representation (Kuhn,
Thus, Kuhn and others suggest children may not yet have the relevant skills or are unable to recognize when or how to implement the necessary skills. Consequently, children exhibit poor performance during experimentation assessments. Collectively, the results of early experimentation studies indicate significant limitations in young children’s scientific reasoning abilities. Kuhn reported that it is not until later in development (10 or 11-years-old) that children can successfully recognize theory and evidence as distinctive categories with comprehensive practice and scaffolding (Kuhn, 2011; Kuhn et al., 2000, 2008). These experimentation assessments exhibited a deficiency in children’s scientific reasoning abilities until preadolescence (e.g., Kuhn, 1989; Schauble, 1996, 1990; Siegler & Liebert, 1975). However, alternative empirical findings have resulted in an inconsistent interpretation of the early childhood scientific reasoning abilities. Subsequent studies with modified paradigms reveal that young children can successfully engage in experimentation within the appropriate contexts (Bullock & Ziegler, 1999; Edelsbrunner et al., 2015; Piekny et al., 2014; Piekny & Maehler, 2013; Sodian et al., 1991; Varma, 2014).

**Experimentation - Early competence**

Assessments of children’s experimentation abilities have used distinctive methods, and these different methods at times have yielded contrasting results. Whereas some assessments have involved children designing and performing experiments, others have required children to evaluate the informativeness of potential experiments. Researchers have employed both methods to examine previous scientific reasoning conclusions during childhood. Early studies indicated poor performance on both methods, designing experiments, and evaluating experimental outcomes. Researchers have argued that Kuhn and others underestimate children’s abilities and
have provided evidence for earlier competence using modified or alternative tasks. Subsequent investigations have revealed that young children can engage in experimentation within the appropriate contexts (Bullock & Ziegler, 1999; Edelsbrunner et al., 2015; Sodian et al., 1991; Varma, 2014). Differences in studies that find early vs. later competence will be further discussed in the current section.

Sodian et al. (1991) studied early elementary school children’s ability to discriminate between hypothetical beliefs and evidence. The study was specifically designed to address Kuhn et al.’s (1988) claim that children do not differentiate theory from evidence. Sodian and colleagues argued that young children do demonstrate a basic understanding of scientific reasoning in early elementary. Students in grades 1 and 2 (aged 6-9) were presented with a story problem about experimentation and asked to distinguish between conclusive and inconclusive tests. Students were informed of a disagreement between two story characters about the size of a mouse within their home. One character believed the mouse was big and the other believed it was small. Two boxes were presented as “mouse houses”, with different sized openings and had food inside. Each child was then presented with two conditions. In the ‘Feed’ condition, children were to determine which box should be used, the one with a large or with the small opening, if the characters wanted the mouse to secure the food regardless of the mouse’s size. In the ‘Find Out’ condition, children were asked to select the box that would facilitate the characters to determine the size of the mouse based on whether the mouse was able to access the food in the box. Results revealed high accuracy for each condition. Most first graders (90%) and all second graders (100%) answered the ‘Find Out’ question correctly. The majority of the second graders (86%) and only half of the first graders (55%) answered the ‘Feed’ questions correctly. Children
of both ages identified when a conclusive outcome was possible and when it was not. Results demonstrated that children were able to successfully distinguish conclusive from inconclusive outcomes when presented with simple experiments. Sodian et al. (1991) argued children’s ability to recognize how to generate a desired effect (feeding a mouse) versus how to test a hypothesis (deciding if the mouse was big or small) exhibits understanding of the difference between hypothesis or “theory” and evidence.

Piekny and Maehler (2013) administered Sodian et al. ’s (1991) mouse task across five age groups (aged 4-13) to measure children’s experimentation skills. Similar results indicated early awareness of conclusive vs inconclusive outcomes with general improvement with age. Most students accurately answered the Feed question preschool (70%), first grade (90%), third grade (95%) and fifth grade (88%). Children across age groups demonstrated improved performance with age. Performance for ‘Find Out’ questions were lower for the younger age groups, preschool (33%) and first grade (34%) and above chance for third and fifth graders, 63% and 88% respectively. Most students could identify conclusive experiments. Preschoolers (51%) correctly answered the test questions and subsequent age groups were significantly above chance first (90%) third (93%) and fifth grade (96%). The majority of students successfully identified inconclusive experiments across age groups: preschool (54%), first (74%), third (83%), and fifth (96%). Piekny and Maehler argued that children as young as 4 possess a rudimentary understanding of difference between hypothetical beliefs and evidence thus, successfully engage in science reasoning when presented with a simple test.

Results by Sodian and Piekny demonstrate basic understanding of scientific concepts during early childhood (Piekny et al., 2014; Piekny & Maehler, 2013; Sodian et al., 1991). These
results challenge the claims of previous investigations by demonstrating that children do understand the difference between theory and. The researchers argue that children’s performance may be a function of task demands. Researchers have claimed that the tasks used in many previous studies were too complex to accurately measure experimentation skills among younger children (e.g., Piekny & Maehler, 2013; Sodian et al., 1991). Specifically, task demands may have exceeded the capabilities of young participants. For example, Kuhn et al. (1988) and Siegler and Liebert (1975) required children to control for multiple variables while making causal judgments; however, Sodian et al. (1991) did not require children to control for multiple variables. Piekny et al. (2013, 2014) suggest that selecting an informative experiment may be challenging because it requires predicting potential outcomes. Thus, it is argued during age-appropriate tasks young children can engage in successful scientific reasoning. However, researchers have critiqued the claims made by Sodian et al. regarding early scientific reasoning competence. For example, Zimmerman highlighted that it is unclear if children could distinguish between testing a hypothesis versus producing an effect since the task instructed children to reason about producing an effect (Zimmerman 2000, 2005). Additionally, the “mouse task” procedure did not include the opportunity to compare children’s performance when directed to test a hypothesis from when directed to produce an effect. Furthermore, Kuhn and Pearsall (2000) argue that the task by Sodian et al. required children to select between the more informative outcome of two potential forms of evidence. They determined that children’s ability to choose within this decidedly constrained set of outcomes does not demonstrate metacognitive awareness of the distinction between theory and evidence. Thus, Sodian et al.’s findings and the
subsequent works by Piekny and colleagues (2013, 2014) potentially overestimate children’s scientific reasoning abilities.

Methodological critiques of experimentation studies have focused on issues related to the strength of children’s prior beliefs. Researchers have proposed that prior beliefs impact children’s reasoning during scientific reasoning tasks across development. Kanari and Millar (2004) examined inferencing in students aged 10- to 14-year-olds with a multivariable physical science experimental task. During two self-directed experimentation sessions, students were asked to consider the variables that influence a pendulum period or determine the factors that influence the force produced by the weight and surface area to pull a box along a surface. Both experiments involved reasoning about the relationships between two independent variables and a dependent variable. In one experiment, an independent variable covaried with the dependent variable and in the second experiment it did not. For the physical experiment, students were asked to vary the physical conditions of the task and use the available measuring equipment (e.g., using a stopwatch with the pendulums or using the force meter to measure the force of the weight and surface area of a box). For the video-experiment, students watched a video recording of another student doing a similar physical task and asked to interpret the data collected during an interview. Results indicated that there were not significant differences by age or reasoning when students were completing the physical task versus their video task responses. For example, the majority of students (60%) could identify a causal relation between the target variable and the outcome variable. However, there were significant performance differences across context when reasoning about covariation compared to noncovariation problems. Students reported correct covariation outcomes (88%) however, demonstrated reduced performance during
noncovariation cases (28%). They found that students could design proper experiments, record data, and correctly interpret findings. However, some students distorted findings to preserve expectations. That is, students' prior beliefs specifically, their ideas about data and measurement affected the evaluation process resulting in misinterpretation of experimental outcomes. Kanari and Millar proposed that less ambiguous data requires a revision of prior beliefs to make correct inferences from the experiments.

Evidence suggests that preserving prior beliefs leads students to make inaccurate judgments. Therefore, it was critical to ensure that children did not hold strong prior beliefs concerning study materials if that was not specifically being assessed. Jointly, results summarized in this section on early experimentation competence indicate that children can effectively participate in scientific reasoning activities (e.g., Kanari & Millar, 2004; Piekny & Maehler, 2013; Sodian et al., 1991). Children can test and revise theories and to reflect on the observed evidence in the appropriate context with the proper experimental materials. The next section will highlight empirical findings concerning evidence evaluation.

**Evidence evaluation investigations**

Evidence is defined as information or observation that serves to confirm or disconfirm a hypothesis (Amsel & Brock, 1996; Astington et al., 2002; Kuhn et al., 1995; Zimmerman, 2000). Developmental studies of evidence evaluation have primarily examined children’s ability to evaluate covariation evidence or children’s ability to distinguish determinate from indeterminate evidence. Research procedures have varied in terms of task complexity, questions asked, and response format (e.g., certainty judgments or forced choice, and type of judgment (Fay & Klahr, 1996; Koerber et al., 2005; Miralda-Banda et al., 2019; Tullos & Woolley, 2009). The number of
correct valid inferences is typically used as a measure of performance since inferences involve accurate evaluation of evidence (i.e., knowing whether the evidence is conclusive or not) and determining a conclusion that can be correctly supported (Kuhn et al., 2000; Zimmerman, 2000, 2005). Early investigations suggested children experience significant difficulty when asked to reason about patterns of evidence (Kuhn, 1989). However, ensuing investigations demonstrated that young children possess basic evidence evaluation skills (Masnick & Morris, 2008; Piekny & Maehler, 2013; Ruffman et al., 1993). The current section will summarize empirical research examining children’s evidence evaluation skills. The review will highlight developmental discrepancies between early evidence evaluation tasks and subsequent investigations and discuss the conceptual and methodological differences within the literature.

Evidence evaluation – Late competence

Kuhn and Pearsall found that young children had difficulty drawing conclusions based on patterns of evidence (Kuhn & Pearsall, 1998, 2000). Children ages 4-6 were presented a series of pictures showing two characters winning in a race. Pictorial cues suggested an explanation for why one character won the race; for example, one character had decorative shoes on, and the other character did not. The last picture of the sequence presented evidence of the result of the race, for example one character hoisting a trophy with a smile. For each trial, children were asked to determine who won the race with two questions, “How do you know?” to examine their use of the evidence and “Why is it so?” The first question examined children’s use of the evidence for their claim (the outcome cue) and the second question assessed children’s explanation (the theory-generating cue). Results demonstrated that most 4-year-olds had difficulty providing valid inferences and even when prompted less than a third gave evidence-
based responses for outcome claims. For example, when asked, “How do you know [he/she won]?” 4-year-olds reported theory-based responses (e.g., “Because he has fast sneakers”). There were noticeable improvements among 6-year-olds, who more consistently gave evidence-based responses for task items (e.g., “He’s holding the trophy”). In contrast, all adults demonstrated full understanding of the difference between theory and evidence. Adults successfully reported evidence-based responses for all presented pictorial cues. According to Kuhn and Pearsall children do not grasp that knowledge is founded on evidence and that prior belief can be modified with new evidence. Results indicate that young children have difficulty providing evidence-based responses during simple covariation tasks. Most children used theoretical cues (e.g., “Because he has fast sneakers”) and did not use evidence (e.g., “He’s holding the trophy”) as justification for responses. Kuhn and Pearsall argued that 4-year-olds display a combination of theory-based and evidence-based responses because they confuse a theory to explain the plausibility of an outcome and evidence demonstrating an outcome. Six-year-old children displayed an increased performance of this differentiation however, the researchers suggested this revealed a limited understanding that is present in limited contexts. Additionally, Kuhn et al. suggest that young children are not yet capable of scientific reasoning since they do not yet understand the concepts of testing an idea or proving a claim (Kuhn, 1989; Kuhn et al., 1995; Kuhn & Pearsall, 2000; Schauble, 1996). Kuhn and colleagues maintain that young children are unable to evaluate evidence independently of beliefs, and it was not until adolescence that most children could consistently reason about scientific concepts. However, it is important to note that similar errors do persist into adulthood. In a study by Kuhn (1989) both adolescents (Grades 6 and 9) and adults provided theory-based (invalid) responses rather than evidence-based response
(valid) responses when questions related to patterns of evidence. They suggested that errors during tasks occur because theory and evidence merge into one representation of “the way things are” resulting in invalid responses (Kuhn, 1989, 2005; Zimmerman, 2007). However, studies concerning evidence evaluation have yielded inconsistent findings.

**Evidence evaluation skills – Early competence**

Covariation studies. Many evidence evaluation assessments involve children’s inference of causal relations or interpretation of covariation evidence (Amsel & Brock, 1996; Fay & Klahr, 1996; Koerber et al., 2005; Kuhn et al., 1995, 2000; Ruffman et al., 1993). Customarily, children are presented patterns of evidence and are asked to determine if a target variable is causally related to a specific outcome or not (Amsel & Brock, 1996; Zimmerman, 2000; Zimmerman & Klahr, 2018). The ability to infer causal relationships and accurately interpret observed patterns support the claim that young children can accurately evaluate evidence (Zimmerman, 2005, 2007).

Kuhn and colleagues pioneered investigations on children’s evaluation of covariation evidence during scientific reasoning (Kuhn et al., 1988; Zimmerman, 2007). Their studies examined how children’s prior beliefs influenced reasoning about causal variables (Kuhn, 1989). Young children did not provide evidence-based explanations when asked to reason about a presented event (Kuhn 1988, 1989). Kuhn concluded that young children could not differentiate theory or beliefs from evidence (Kuhn, 1989; Kuhn et al., 1988; Zimmerman, 2000, 2005). However, Kuhn reported a significant improvement in skill between 4 and 6 years. Subsequent investigations challenged Kuhn’s conclusions by demonstrating that children can distinguish between theory and evidence when an age-appropriate task is administered. Other researchers
(e.g., Ruffman et al., 1993; Sodian et al., 1991) suggested that methodological issues, specifically the strength of children’s prior beliefs, and the task complexity, could influence children’s performance. The following section will discuss investigations that demonstrated that children as young as 4 years old can distinguish theory from evidence.

Ruffman and colleagues conducted three experiments that assessed (a) children’s reasoning with varying strengths of covariation evidence, and (b) the influence of prior beliefs on children’s reasoning. Ruffman et al. (1993) examined 4- to 7-year-old children’s ability to form a hypothesis based on perfect covariation, imperfect covariation, and non-covariation evidence. Several tasks were used to evaluate children's understanding of the role of covariation evidence and the ability to reason about the same variables differently depending on the patterns of evidence. For Experiment 1, children were presented stories about fictional food items with different outcomes related to dental health. Children were shown pictures of characters with a type of food (e.g., green food or red food) and corresponding oral health outcome (e.g., lost teeth, maintain all teeth). For example, "Look, this boy ate some green food. And look, he has some teeth missing." Children were presented with either green food or red food being associated with tooth loss. The experimenter asked what the story character would say caused tooth loss as well what the child believed caused the presented outcome. All children answered the perfect covariation task correctly. Next, “food” items were rearranged to indicate that the opposite food color was now responsible for tooth loss. For example, children who were initially presented with green ‘food’ associated with healthy teeth and were then shown pieces of green ‘food’ paired with tooth loss. Children were asked how another character would respond to the new misleading evidence, "Why will Sally say the [red/green] food makes kids' teeth fall out?"
Results indicated that some 4-years-olds (38%) and the majority of 5-year old’s (81%) could correctly assess the covariation evidence. Additionally, 6-year-olds (71%) and 7-year-olds (80%) could successfully reason about the causal relationships during perfect covariation outcomes (e.g., one causal factor (red food) paired with a specific outcome (tooth loss)).

For Experiment 2, children were asked to assess imperfect covariation evidence. In Experiment 1, all the evidence presented was consistent with a single hypothesis and inconsistent with the other hypothesis; however, for Experiment 2 most of the evidence was consistent with a hypothesis, but not all evidence was consistent with that hypothesis. Thus, this task provided an increased level of difficulty for children to reason about patterns of evidence. Children were introduced to a puppet, small cardboard figures that represented “real people” each with a blue hat on their head, and two types of fictitious food (e.g., red and green) pieces depicted as stars. The goal was to determine which type of food the characters liked to eat. Children were told a brief story and watched the story characters go to eat each type of food. When the character bent down to eat their hat fell off and would be left next to the food. The hats would be clues for the number of characters that ate each food. For example, one story suggested that the characters liked both types of food (e.g., three ate green food and three ate red food). During an alternative story there was a noticeable preference of one type of food (e.g., five characters ate red food and only one ate green food). The experimenter then asked, "Which picture best shows which type of food the girls like to eat? Do they like to eat the red food, or the green food, or do they like to eat both types of food?" Next, the experimenter confirmed the child’s response and then rearranged the study variables prior to reintroducing the puppet (e.g., Now you and I know which food the girls ate. We know that three girls ate red food and three ate green food. But let's do something
before John gets back. We'll move some hats.”. When the puppet returned and viewed the manipulated evidence, children were asked what the puppet would say and asked again about their initial hypothesis (e.g., "John comes back and sees things this way. Which picture best shows which type of food John will say the girls like to eat? Will John say they like to eat the red food, or the green food, or will he say they like to eat both types of food?). Then children were asked, “And what about you? Do you say they like to eat the red food, or the green food, or do they like to eat both types of food?” This sequence was then followed by justification questions. The primary focus was children’s response to the second question since it required them to assign a new hypothesis to the puppet while maintaining their initial hypothesis. Results found that 5-year-olds (22%) had difficulty, but the majority of 6-year-olds (71%) and 7-year-olds (80%) provided correct responses and could understand that evidence did not have to be perfect to generate a conclusion. Ruffman and colleagues concluded that by approximately 5- to 6- years old children have a basic understanding about causal relationships when presented with covariation outcomes.

Ruffman and colleagues proposed that children have a basic understanding of the relation between hypothesis or “theory” and evidence. They suggested that results reported by Kuhn et al. 1988, that children could not coordinate theory and evidence until approximately 12 years old was due to methodological limitations of the study. Ruffman et al. suggested that previous low performance was likely related to how evidence evaluation skills were assessed. For example, Kuhn reported that children did not consistently provide evidence-based responses for covariation tasks and did not recognize when their new theory contradicted an earlier theory that they held. Ruffman argued that children may have misunderstood Kuhn’s justification prompts
and believed they were to provide reasoning of their own view rather than based on the evidence presented. Additionally, Ruffman contended the study procedure likely caused errors due to children’s strong prior beliefs about the presented information. Lastly, Ruffman suggested Kuhn’s investigation involved manipulating three or more potential causal variables resulting in poor performance due to high cognitive demands of the task. Ruffman et al. concluded that children as young as 5 years old could successfully interpret covariation evidence and construct a causal hypothesis if prior knowledge was not a factor and the task complexity was age appropriate. However, the main component of the Ruffman et al. task required children to reason about characters engaging in one behavior more than another. Thus, it could be debated that children were comparing frequencies of a behavior outcome rather than actively reasoning about covariation evidence. Metz (2004) argues that increased performance of scientific reasoning during laboratory tasks may be due to the oversimplification of the task demands. Metz suggests that in the attempt to develop age-appropriate procedures the reasoning required may not be demanding enough. As a result, findings by Ruffman et al. and Sodian et al. could overestimate children’s scientific reasoning abilities. Thus, when examining early evidence evaluation studies, it is critical to consider the skills being assessed and the interpretation of study findings.

Koerber et al., (2005) examined children’s ability to interpret patterns of covariation evidence. Four- to six-year-olds were read stories concerning a story character’s beliefs about health outcomes displayed through pictures. Children were asked to interpret patterns of perfect covariation, imperfect covariation, and non-covariation (e.g., a character believes green chewing gum is unhealthy, and 10 pictures of green chewing unhealthy/missing teeth and 10 pictures of red chewing gum with healthy teeth are presented). Koerber et al. demonstrated that preschool
children could successfully identify simple patterns of covariation evidence, specifically
detecting the correct association between chewing gum and teeth health during perfect
covariation outcomes (e.g., the story character believes red chewing gum is healthy and is
presented 10 pictures of red chewing gum with healthy teeth and 10 pictures of green chewing
gum with unhealthy/missing teeth). Performance on perfect covariation tasks expressed robust
understanding: 4-year-olds (90%), 5-year-olds (87%) and 6-year-olds (91%). Additionally,
children performed well during the imperfect covariation task (e.g., 8 pictures displayed the
combination of a blue handkerchief with a red/sick nose and a yellow handkerchief with a
healthy nose and 2 pictures showed the arrangement of a blue handkerchief with a healthy nose
and a yellow handkerchief with a red/sick nose). Children demonstrated understanding during
the imperfect covariation task: 4-year-olds (71%), 5-year-olds (91%) and 6-year-olds (90%).
However, children had notable difficulty evaluating non-covariation information when presented
with an unclear outcome (e.g., 2 pictures of green chewing gum with unhealthy/missing teeth
and 2 pictures of red chewing gum and healthy teeth); 4-year-olds (19%), 5-year-olds (39%) and
6-year-olds (50%). It was suggested that the ambiguous information during non-covariation tasks
provided a level of complexity that increased the level of difficulty for young children.
Moreover, it is potentially more challenging to recognize inclusive evidence or identify the lack
of a relationship between variables. Koerber’s results extend Ruffman’s findings indicating that
young children can understand the causal relation between variables when evaluating
unambiguous covariation data. Koerber and colleagues determined that young children could
reason about patterns of evidence and can successfully differentiate hypotheses or “theory” from
evidence when presented appropriate tasks.
Using a nearly identical task, Piekny and Maehler (2013) demonstrated comparable results to Koerber et al. (2005). Children were told a story and then shown pictures depicting oral health outcomes with a puppet present. For the perfect covariation (e.g., 10 pictures with red chewing gum and unhealthy teeth and 10 with green chewing gum and healthy teeth) children as young as 4 could effectively reason about perfect covariation evidence; preschool (87%), 1st grade (96%), 3rd grade (98%), and 5th grade (100%). Though there was evidence of improvement with age there was a noticeable decline in imperfect covariation trials (e.g., 8 pictures with the combination of yellow chewing gum and unhealthy teeth and blue chewing gum with healthy teeth, and 2 with the grouping of blue chewing gum with unhealthy teeth and yellow chewing gum with healthy teeth). Results demonstrated that difficulty with younger children with some improvement by middle elementary; preschool (59%), 1st grade (33%), 3rd grade (54%), and 5th grade (90%) correctly evaluated the imperfect covariation pattern. Performance on noncovariation tasks (e.g., 2 pictures displaying yellow chewing gum and unhealthy teeth and 2 pictures of blue chewing gum with healthy teeth) revealed to be the most difficult across age groups; preschool (30%), 1st grade (75%), 3rd grade (78%), and 5th grade (73%). Additionally, Piekny and Maehler highlighted that there was not a notable difference between 1st, 3rd, and 5th grade for non-covariation performance. Children’s ability to evaluate perfect covariation appears to be present early in development however, imperfect and noncovariation appears to be more challenging. Piekny and Maehler concluded children can differentiate theory and evidence however, some abilities within evidence evaluation skills such as evaluating noncovariation evidence may develop non-linearly. Findings suggest rudimentary aptitude of evidence evaluation are present as early as preschool however performance
differences are consistently found within covariation assessments. The studies summarized
above examined children’s evaluation of covariation evidence. Other studies have examined
children’s evaluation of whether evidence is indeterminate or determinate.

**Conclusive vs. inconclusive evidence.** To further investigate evidence evaluation skills,
researchers assessed children’s ability to distinguish determinate (i.e., when information or
argument is conclusive) from indeterminate (i.e., when the information presented is inconclusive
or uncertain) evidence. Studies that evaluate determinate or indeterminate outcomes require
participants to recognize whether evidence is consistent with only one possibility or is
ambiguous and is consistent with multiple possibilities. The following section will review
empirical studies that explore children’s ability to reason about conclusive from inconclusive
evidence.

Fay and Klahr presented children aged 4 to 6 (M=5.3) a target object (e.g., a necklace
made from red wooden beads) to a sequence of four boxes. Enclosed in each box were a different
color of beads (e.g., blue, green, red). First, each box was closed and consecutively opened in a
fixed order. After each box was opened children were asked about which of the four boxes was
used to assemble the target item (e.g., necklace made from red beads). The experiment asked,
"Is this a time that you can tell for sure which box I used to make this or is this a time that you
cannot tell?" followed by a justification prompt (e.g., “How can you tell for sure which box I
used” or “Why can't you tell which box I used?”). Children were asked to reason about
observable information (opened boxes) and forthcoming evidence (boxes that were to be
opened). The indeterminate instances included, 1) at least two boxes that included the materials
of the target item (e.g., box 2 and 4 contained red beads) and 2) each of the four box contained
materials of the target item (e.g., box 1-4 contained red beads). A determinate outcome included only one box with the target item pieces (e.g., box 3 contained red beads). Results demonstrated that preschool children can successfully identify determinate from indeterminate evidence using “Can tell” or “Can’t tell” answers within simple investigations. Children showed increased performance when the outcomes were observable (e.g., each box was open) compared to reasoning about prospective situations (e.g., boxes to be opened). Fay and Klahr determined that preschool aged children have a basic understanding of the distinction between determinate and indeterminate evidence. According to Fay and Klahr (1996) children have greater difficulty examining indeterminate evidence compared to determinate evidence due to increased ambiguity.

Klahr and Chen’s procedure was designed to extend Fay and Klahr’s (1996) findings on the ability of children to discern determinate from indeterminate conclusions. Children were randomly assigned either to neutral problems or to situations where the task involved participants giving credit to or assigning blame to a story character. In the Familiarization phase, children were presented six boxes with coloring markers inside: three blue, two red, and one green. Children could see one marker was present in a box however, the same color marker could be assigned to multiple boxes. During the Practice phase, children saw six marker boxes in sets of two (yellow/green, purple/purple, and brown/brown). Children were presented two marker boxes and a picture (e.g., a row of circles drawn with one color marker). During the story problem, children learned that a story character drew the picture using one of the markers in the box. Children were then asked, “With both boxes closed, do you know for sure which box was used to make this design, or do you have to guess?” The experimenter opened the boxes one at a time.
and asked, “Do you know for sure which box was used to make the design, or do you have to guess?” Children were presented with one determinate and one indeterminate example during the Practice phase.

During the Experimental stage children were shown three marker boxes with a picture (e.g., a flower), a scribble or a design (e.g., a series of shapes). A story problem notified the participants that one of the characters drew the picture, scribble, or design using a marker located in one of the boxes. During the context trials, children were shown a picture (credit outcome) or a scribble made directly on the desk (blame outcome). Children were asked, “Do you know for sure who made this picture/mess/design, or do you have to guess?” Justification prompts were asked following each response, (e.g., “How do you know for sure?” or “Why do you have to guess?”). For the context condition, four credit and four blame problems were presented. The neutral condition consisted of four determinate and four indeterminate outcomes. The results demonstrated that 4-year-olds (59%) and 5-years-olds (76%) could correctly distinguish determinate from indeterminate evidence. However, there was no difference between the context and neutral conditions. Similar to Fay and Klahr (1996) children displayed increased difficulty during indeterminate problems particularly when some evidence was still hidden (i.e., not all boxes were open). It was concluded young children have a basic understanding of the difference between determinate from indeterminate information; however, the study procedures and the methods for presenting evidence can influence performance.

In addition, Tullos and Woolley (2009) investigated children’s ability to make an inference about determinant vs. indeterminate outcomes using the presented evidence. Specifically, 4- to 8-year-old children were shown a series of animals and asked them to make
inferences about the reality status of novel animals based on the evidence provided. Children were asked to use “clues” or evidence about various materials (e.g., twigs, feathers, tree leaves, etc.,) to determine if the selected animal was “real” or “pretend”. In the story, children were asked to help a researcher verify the reality status of an animal based on the materials left in evidence boxes by the animals. An experimenter read aloud a description of the animal while children examined the contents of the box. For example, while observing a twig in the evidence box children would hear, “Takins eat twigs and leave twigs behind wherever they go.” Alternatively, when prompted to examine the evidence box children may have heard irrelevant information or an empty evidence box accompanied the description. Follow-up questions were then administered (e.g., “Do you think an animal was in the box, yes or no?” “Is this a time when you can tell if Takins are real or not; can you tell or can you not tell?” Two experiments assessed children’s use of evidence and their ability to recognize supporting information or ignore irrelevant evidence.

Children aged 6 to 8 were above chance for each of the three levels of evidence (e.g., supporting, irrelevant, and no evidence). Older children consistently identified that supporting evidence led to determinate outcomes and when irrelevant or insufficient evidence resulted in indeterminate outcomes. However, younger children’s (4- and 5-year-olds) assumptions about the reality status of the novel objects influenced the strategies they used when evaluating the evidence. For example, children who originally considered the new animal to be real continued to state that it was real despite evidence indicating that it was pretend. Children who initially judged that the novel animal was pretend maintained this conclusion even when presented with evidence and “clues” that suggested the animal was real. The children who did not establish a
prior viewpoint were more likely to use the evidence presented to determine the reality status of the object. Four- and five-year-old children showed the ability to distinguish the novel animals, however, they did not consistently use evidence-based responses. These findings show that children aged 6 to 8 can use and understand inference as a source of knowledge and this skill improves with age. However, prior beliefs influence reasoning which could account for difficulties for younger children.

**Evidence evaluation during deductive reasoning**

In addition to scientific reasoning tasks, researchers have developed alternative approaches to examine children’s evidence evaluation skills. Pillow and colleagues have tangentially assessed children’s evaluation of evidence and reasoning using deduction vs. guessing tasks (Pillow et al., 2000, 2010; Pillow & Pearson, 2012). Research has shown that children can make valid deductive inferences by age 5 or 6 (Pillow & Anderson, 2006; Pillow & Pearson, 2009), and suggest that young children can reason about the basic properties of deductive inference (Galotti et al., 1997; Hawkins et al., 1984; Pillow, 2002). Deductive inferences require evaluating two or more premises to reach a novel conclusion which requires reasoning from the general to the specific.

In a series of studies, Pillow and colleagues assessed children’s abilities to distinguish between conclusive evidence that provides a deductive inference compared to indeterminate evidence resulting in an inductive inference or guess (Pillow, 2002; Pillow et al., 2010; Pillow & Anderson, 2006). Pillow and Anderson (2006) examined first grade ($M = 6.8$) and third grade ($M = 8.10$) children’s concepts of deductive inference and guessing. Children were asked to identify and remember a deductive inference or guess as the source of a belief, to remember the certainty
of a belief, and to evaluate their own deductive inferences and guesses. A warm-up procedure familiarized the children with the rating scale and the key terms used in the main task (e.g., ‘figure out’ and ‘look’). Six trials (2 infer, 2 look, and 2 guess) were presented. In the infer task, children were presented with two toys that differed on an observable property (e.g., color, or size) that were hidden in separate containers. Children were asked questions about the toys and to indicate the certainty of their responses. For example, “I have two pigs. There’s a yellow pig and a green pig. Now I’m going to hide the pigs. There’s one in here and one in here.” In the infer trials, children were told they can look into one container then asked about the color of the other toy. Children were then asked certainty questions (a) “How sure are you that the pig is green/yellow? Why did you say the pig is green/yellow and a source question? Did you figure it out or did you look at it with your eyes?”. For look trials, the procedure was similar, and questions involved the toy the child had directly seen. For guess trials, children were asked about the color of one of the hidden toys but did not look into either container. The certainty scale ranged from 0 to 5 with higher numbers corresponding to greater certainty. Results indicated that for the immediate condition, first grade (Look = 4.84, Infer = 4.94, Guess = 4.17) and third grade (Look = 5.00, Infer = 4.81, Guess = 3.53) and the delayed condition (Look = 4.55, Infer = 4.34, Guess = 4.25) and third grade (Look = 4.72, Infer = 4.72, Guess = 3.28) children rated the direct visual trials and inferences as more certain than the guess examples.

Pillow et al. (2000) investigated children’s understanding of deductive inference from perception and guessing. Children in preschool (M = 4.5), kindergarten (M = 5.8), 1st grade (M = 6.8), and 3rd grade (M = 8.9) were presented with two toys of the same color or of different colors. In the different color examples children were told, “There's a blue marble and an orange
marble." Then the toys are hidden under two separate containers. The puppet would then look under one container and state a color (e.g., “This marble is blue”). Children would then be asked about the puppet’s belief, "How sure is Bob that the marble is blue? Can you show me with the arrow?” and the puppet’s certainty, "How sure is Bob that the marble is blue? Can you show me with the arrow?” The two inference trials had the puppet look into one container and reported the color of the toy still hidden under the other container. For the guess trials, the puppet pointed to one container and reported the color of the hidden toy without looking into either container. Results showed on the own-inference trials all children could make a simple deductive inference by correctly inferring the color of the object they did not directly look into.

Results show that 6- and 8-year-olds could consistently distinguish their own and another person’s deductive inferences as more certain than guesses, however younger children aged 4 and 5 showed inconsistent response patterns. Specifically, the younger children did not demonstrate a difference in certainty scale ratings for inferences and guesses which indicates they did not fully distinguish these inferential activities. Evidence has demonstrated that children demonstrate basic understanding of deduction however, young children may exhibit greater inconsistencies in pattern of performance when presented with outcomes of varying levels of certainty.

Pillow and Anderson (2006) proposed that the understanding that deduction can provide knowledge that is certain, and induction and guesses are uncertain and need additional verification may be an important component of scientific reasoning development. For example, children recognize that valid deductive inferences are conclusive similarly to determinate evidence during scientific reasoning investigations. Additionally, children have greater difficulty
recognizing that guesses are less conclusive and certain which relates to indeterminate or inconclusive evidence evaluation. Consequently, the understanding of the varying properties of deduction and induction may facilitate scientific thinking development (Kuhn & Pearsall, 2000; Pillow, 2002; Pillow et al., 2000; Watters & English, 1995).

**Summary**

Collectively studies examining evidence evaluation in childhood demonstrate that basic abilities to evaluate evidence are present in preschool and continue to improve with age (Koerber et al., 2005; Piekny et al., 2014). Researchers have concluded that young children can accurately reason about evidence in concert with the fundamental principles of scientific inquiry. Specifically, young children can evaluate whether information allows for a certain conclusion or not. Evidence suggests that beginning at approximately age 5 or 6 children can successfully evaluate outcomes such as identifying determinate from indeterminate evidence, reasoning about covariation and have a basic understanding of the difference between theory and evidence (Koerber et al., 2005; Piekny & Maehler, 2013; Sodian et al., 1991). However, performance can be influenced by study methods (e.g., task complexity) and prior knowledge (Croker & Buchanan, 2011; Jirout & Zimmerman, 2015; Ruffman et al., 1993; Tullos & Woolley, 2009). Additionally, studies involving children’s evaluation of deduction vs. guess provide alternative procedures and additional support of children’s evidence evaluation abilities. The debate regarding the emergence and the scope of scientific reasoning abilities of children persists. The dispute between early competence vs. late competence investigations is often reduced to is scientific reasoning an all or nothing ability, such that children either have the ability or they do not. However, it is more likely a progression in which expertise and performance improves
across development. Nonetheless, different approaches to examining early evidence evaluation skills provide the opportunity to examine cognitive development processes during development.

**Evaluation of reasoning**

The following section will discuss empirical studies that consider children’s reflections of their own reasoning. This section will specifically consider children’s ability to evaluate whether the presented information allows for a certain or uncertain conclusion. Additionally, applicable investigations that assess children’s evaluation of reasoning in contexts other than scientific inquiry will be discussed. The purpose of this section is to highlight alternative methods used to examine children’s reasoning about evidence with corresponding empirical results.

Feelings of certainty or uncertainty allow for self-monitoring of cognitive activities. These feelings may serve as a cue for children about their progress during comprehension, reasoning, or problem solving (Flavell et al., 1985; Pillow & Pearson, 2012). Certainty may signify understanding, recognition of complete and conclusive information, or identifying appropriate strategies to complete a goal. Alternatively, uncertainty may indicate errors in thinking, inaccurate, or incomplete information (Lyons & Ghetti, 2011; Metz, 2004; Pillow & Pearson, 2012). By approximately 5 years of age children can successfully evaluate their own reasoning and report accurate levels of certainty (Ghetti et al., 2013; Pillow & Anderson, 2006; Pillow & Pearson, 2012). Many studies of the development of reasoning have assessed children’s evaluation of their own reasoning. This is typically captured via certainty judgments and open-ended justification prompts. Results have indicated that children aged 4 to 6 can reflect on their own reasoning by scoring the certainty of their responses. The following section will include a brief introduction to empirical works that involve children’s reflections during reasoning.
Four-to-eight-year-old children have difficulty identifying, interpreting, or using uncertainty cues (Byrnes & Overton, 1986; Harris et al., 1981; Scholnick & Wing, 1988). However, other studies have demonstrated that young children have an awareness of uncertainty and can recognize their own feelings of uncertainty (Pillow, 2002). For example, preschool children (aged 3-5) can monitor their own feelings of uncertainty to report greater confidence for accurate answers and less confidence for inaccurate responses (Destan et al., 2014; Lyons & Ghetti, 2013). Other studies have demonstrated children’s ability to use certainty rating to differentiate cognitive activities such as deduction, induction, and guess (Pillow & Anderson, 2006; Pillow & Pearson, 2012, 2015). However, results indicate that children monitor successfully during deduction and have increased difficulty during induction and guess.

Galotti et al. (1997) asked children (aged 5-11) to draw an inference from deductive or inductive inference problems and rate their own certainty. Sixteen syllogisms were presented with fictional content and corresponding drawings. Each syllogism had a deductive and an inductive version that was matched for content but differed in level of certainty. For example, a deductive version stated, “All poggops wear blue boots. Tombor is a poggop. Does Tombor wear blue boots?” The corresponding inductive version stated, “Tombor is a poggop. Tombor wears blue boots. Do all poggops wear blue boots?” Children provided “yes” or “no” responses and a rating on a 5-point confidence scale. Results indicated that children rated the deductive syllogisms as more certain and had a faster response time compared to the inductive syllogisms. It was suggested that young children can begin to detect feelings of certainty around 5-years-old. Galotti et al. emphasized that children were more successful when tasks required deductive
inferences compared to inductive inferences. Children were also able to reflect on their reasoning through confidence ratings and relevant justifications explanations.

Further investigations have revealed children’s ability to differentiate between the types of inference via certainty judgments. Pillow and colleagues demonstrated young children’s basic knowledge about the distinct properties of deduction and inductive inference. Specifically, young children can successfully discern valid deductive inferences from invalid inferences or guesses (Pillow et al., 2010; Pillow & Pearson, 2009, 2012). Pillow (2002) found that 6-year-old children reported greater certainty during deductive inference than guessing. Additionally, by approximately 8-years-old children consistently report greater certainty to deductive items compared to inductive items regarding their own judgments. Kindergarten children rated deductions as more certain than weak inductions or guesses and by 8 years of age, children judge deductive inferences as more certain than weak inductive inferences informed guesses or pure guesses (Pillow & Anderson, 2006). Study findings demonstrated that 8-year-olds also made this distinction when reasoning about another person’s inferences. Children aged 8 to 9 reported a greater certainty when observing a puppet made a deductive inference compared to a guess (Pillow et al., 2000; Pillow & Pearson, 2012). Pillow and Pearson (2009) determined that young children beginning at approximately 5-years-old rated deductions as more certain than weak inductions or guesses. Additionally, 8- and 9-year-olds could identify that deductions were more certain than strong inductions and could distinguish strong inductions, weak inductions, and informed guesses from pure guesses.

In science, inductive and deductive inferences are made during exploratory investigations. Scientific thinking requires evaluating hypotheses by drawing conclusions from
evidence (Kuhn et al., 1988; Kuhn & Pearsall, 2000). The level of certainty must be considered when evaluating evidence to formulate conclusions. Pillow and colleagues (2002, 2006, 2012) examined children’s ability to recognize reasoning and guessing as distinct processes that are associated with different degrees of certainty. Pillow and Anderson (2006) demonstrated that 5-to 9-year-old children can make this distinction with certainty ratings. Pillow et al. (2010) found that when provided with different patterns of indeterminate evidence of observable items the strength of the induction influenced children’s certainty judgments. Specifically, they demonstrate that elementary school children (age 6-8) can identify deductions as more certain from guesses. The current investigation will extend the work of Pillow and colleagues to examine if children can successfully make certainty differentiations in a scientific reasoning context.

**Summary**

Together these findings demonstrate that children aged 4 to 9 can reflect on and monitor their levels of certainty. Certainty judgments are evidence that children can identify and respond differently to varying levels of evidence. For example, using certainty ratings children can distinguish deductive and inductive inferences in early childhood (Pillow & Pearson, 2009; Ricco, 2015). The presented studies demonstrate that children can successfully reflect on their own reasoning using certainty judgments ratings. Children as young as 6-years-old report greater certainty during valid deductive inferences and guesses as less certain. Findings demonstrate improved monitoring of certainty and uncertainty with age (Pillow & Anderson, 2006; Pillow & Pearson, 2012). Given that the literature has demonstrated that elementary aged children can successfully reflect on their own reasoning using certainty judgments ratings. It is important to
investigate if they can replicate this reasoning in a context that is more like real world science inquiry. Thus, the present study will examine children’s ability to differentiate the certainty of inference vs. guess during scientific reasoning.

**Reasoning about observable and unobservable properties**

**Reasoning about observable properties**

This section will consider children’s reasoning about readily observable and concrete properties compared to unobservable or non-obvious properties. Part of the investigative process involves scientists using evidence and reasoning to draw conclusions about properties or events that they do not yet know or cannot directly observe. The current investigation seeks to examine potential differences when children reason about observable compared unobservable properties during scientific reasoning tasks.

Studies have demonstrated that preschool children recognize that knowledge can be attained by a direct perceptual encounter (Pillow, 1993; Pratt & Bryant, 1990). The scientific reasoning literature has frequently used observable properties during tasks. During experimental procedures participants are often asked to design an experiment or evaluate evidence that allows for conclusive test or outcomes (Chen & Klahr, 1999; Kuhn & Pearsall, 2000; Piekny et al., 2014; Sodian et al., 1991; van der Graaf et al., 2016). Many experimentation tasks require participants to manipulate the independent variables such as the distance traveled by a toy car, floating or sinking of a toy boat (Kuhn et al., 1995; Siegler & Liebert, 1975). Chen and Klahr (1999) presented children in grades 2, 3, and 4 with physical tasks (e.g., springs task, ramps task and found young children could design unconfounded experiments by manipulating one element at a time (e.g., height of ramp). Other tasks ask children to reason about the presented evidence
that involved immediately verifiable information. Evidence evaluations examine children’s ability to reason about items that covary to determine the outcome. For example, evaluating the quality of the evidence (e.g., color of chewing gum and health of teeth or the relation between eating a type of cake and becoming sick) or reasoning from evidence size of tennis balls and a good or bad serve (Koerber et al., 2005; Kuhn, 1989; Sodian et al., 1991). Additionally, the target box paradigm discussed previously examined preschool children’s (aged 4-5) ability to draw conclusions about immediately observable materials (e.g., beads, plastic curved or square pieces) presented to determine if it is conclusive or inconclusive evidence. Additionally, Pillow and colleagues (2000, 2006, 2012) found that children identified deductive inference answers as more certain than inductive inferences or guesses for their own and another person’s judgments. Each study revealed that children can draw inferences when reasoning about concrete items (e.g., blocks, marble, toy car) and observable properties (e.g., color, shape, size). However, it is critical to note in addition to being concrete, the items were immediately available but were not immediately perceptible as some objects were hidden at various points during the task.

**Reasoning about unobservable properties**

Early investigations proposed that children have difficulty reasoning about properties they cannot directly see (Bruner et al., 1966; Carey, 1985; Piaget, 1929). Piaget (1929) argued that young children were unable to reason about non-obvious properties and attributed natural or mechanical occurrences to people’s actions rather than to natural events or internal processes. Additionally, Carey (1985, 1988) maintained that 4- and 5-year-old children have limited understanding of biological properties of people or animals and often hold inaccurate beliefs about inanimate objects (e.g., reporting “the button is alive”). Carey suggested that at
approximately 10-years-old children could fully grasp knowledge related to the internal mechanism of the human body.

In contrast, other studies show that young children demonstrate the ability to reason about items they cannot directly observe. However, fewer studies include tasks which require children to reason about such non-obvious properties. In scientific reasoning, studies that involve non-observable properties often focus on domain-specific content. For example, Hatano and Inagaki (2008) studied children’s understanding of internal biological concepts. Results indicated that preschool children successfully distinguished between plants and nonliving items (e.g., the ability to grow or repair). The researchers asserted that children recognized that humans share internal biological properties with other living entities and use this understanding to make predictions about novel biological situations beyond the information that is explicitly available. However, these studies focused on the properties and successful classification of the variables rather than providing direct evidence for knowledge of shared internal properties. Contrary to the work of Carey (1985, 1988), these results suggest that by approximately age 5, children have a naïve knowledge of unobservable internal biological mechanisms (Inagaki & Hatano, 2013, 2002). Additional studies have shown that children can reason about unobservable concepts in chemistry such as atoms and molecule (Wiser & Smith, 2008), physics (Baillargeon et al., 1995; Spelke et al., 1995), and evolution (Borgerding & Raven, 2018).

Furthermore, evidence shows that children can reason about unobservable properties to make inferences about other entities or events. For example, Sodian et al. (1991) asked 6- and 7-year-olds to determine if food should be placed in a box with a small or large opening for an unseen mouse (either small or large) to access the food. The majority of children correctly
identified that the food should be placed in the box with the small opening. Children recognized that evidence could help formulate an inference about an event that was not directly observable to decide between multiple hypotheses (see Piekny et al., 2013, 2014). The procedure by Sodian et al. is relevant as it suggests that children do recognize that information that is not directly observed can be acquired through inference which requires more advanced reasoning. However, the conclusion by Sodian et al. concerns the outward appearance of the mouse, which is a readily available observable property and involves a singular element, thus, the knowledge is limited in scope.

Collectively, studies have demonstrated evidence that starting at approximately age 5 to 6 children can successfully reason about observable properties. However, results do suggest that there is potentially a gradation present in the extent or nature of observability vs. nonobservability that influences children’s reasoning. Children can accurately reason about concepts that are not immediately available (e.g., internal biological properties, molecules, physics). Additionally, children as young as 6-years-old can make inference about the conclusiveness of a test based on information that is not immediately observable (Sodian, 1991). The present study seeks to examine how children’s reasoning differs when reasoning about observable and unobservable properties within the same task. This will allow further understanding of the logical processing abilities of young children and potentially enhance scientific reasoning knowledge during development. Additionally, the present investigation will compare children’s evaluation of the certainty of inferences when reasoning about varying levels of evidence and assess children’s ability to reflect on their own reasoning using certainty judgment ratings.
Science is used to learn new things about properties of the natural world. Scientists and researchers reason about constructs and properties that are not immediately observable. Thus, they are required to make inferences based on the conclusions drawn from the evidence that is not often readily verifiable. The purpose of this study was to extend the literature by examining children’s reasoning for observable vs. unobservable properties within the same task. Additionally, it is rare for scientists to reason about a series of established facts. Pillow et al.’s procedure required children to reason about temporary facts about readily observable items (e.g., color of marble, size of toy). The current study included a task where children reasoned about an unknown outcome that was not immediately verifiable. This type of reasoning more closely relates to science inquiry in a real-world context. The current study evaluated children reasoning about information that was specific and readily verifiable compared to reasoning about new knowledge that was not specific to a singular item and not readily verifiable.

Scientific reasoning investigations provide evidence that children as young as 4-years-old can successfully distinguish conclusive (determinant) from inconclusive (indeterminant) evidence (Fay & Klahr, 1996; Klahr & Chen, 2003; Piekny & Maehler, 2013). However, previous studies have predominantly used directly observable objects to examine children’s judgments of certainty (e.g., toy bicycle, dinosaur figures, etc.) or involved reasoning about observable properties (e.g., color or size). Pillow and colleagues (e.g., Pillow, 2002; Pillow et al., 2000; Pillow & Anderson, 2006; Pillow & Pearson, 2012) have provided substantial evidence
that young children can evaluate the certainty of inferences about concrete items when the conclusions are immediate and verifiable (e.g., marbles, toy cars hidden in containers near the child). Together study results demonstrate that (a) during evidence evaluation children can distinguish conclusive from inconclusive outcomes and (b) children can successfully reflect on their own reasoning with certainty judgments. However, previous studies involved children reasoning about immediately observable content and did not examine if children can make the same distinctions if the evidence involves unobservable properties. Thus, the current investigation first considered if children’s certainty judgments differed across observable and unobservable properties. The procedure compared certainty ratings for conclusions concerning observable properties (immediately available) vs. unobservable properties (not immediately available). During the Box Task (observable), children were presented with a series of trials with readily observable items (e.g., different color boxes) and asked to make an inference about what was inside the box based on an observable property (e.g., color of box). During the Animal Task (unobservable), children were asked to make an inference about an unobservable internal property of an animal based on the type of animal food that was found in its nest. Children were asked to report their level of certainty for each trial using a 5-point certainty scale. This procedure examined if children made the same distinctions if the conclusion concerns something that was not as concrete and immediately available within a similar task. The current investigation provided further evidence for scientific reasoning development by directly comparing observable and non-obvious properties.

Previous scientific reasoning investigations have utilized force choice paradigms to collect responses. Sodian et al. (1991) presented conclusive and inconclusive experiments and
asked children if they would know the outcome or not, (e.g., “If they put the food in the house with the small opening will they know”). Using the target box paradigm, after the evidence was presented Fay and Klahr (1996) asked children if they would “Know for sure” or “Not know for sure”. Similarly, Klahr and Chen (2003) asked children if they “Can tell” or “Can’t tell” if the item was made from a specific box. The current investigation used a modified certainty rating scale used by Pillow et al. (2015). The scale included a sequence of five vertical bars increasing in height from left to right (0.76, 2.54, 5.08, 7.62, & 10.16 cm.) and width for each bar was 3.33 cm. Each vertical bar had a corresponding number and description: 1-not sure at all, 2-a little sure, 3-medium, 4-kind of sure, 5-very sure. The rating scale could potentially detect differences in children’s evaluations of certainty for each trial that would not be captured by dichotomous response options.

The study was designed to specifically investigate the following research questions:

Research Question 1: How does children’s evaluation of the certainty of inferences about immediately observable properties differ from children’s evaluation of the certainty of inferences about unobservable properties?

Early investigations maintained that the ability to reason about unobservable properties did not develop until preadolescence (Carey, 1988; Piaget, 1929). However, subsequent investigations demonstrated that children could successfully reason about unobservable properties with age-appropriate tasks (Inagaki & Hatano, 2008; Sodian et al., 1991). Investigations by Pillow and colleagues (2002; 2006; 2009) identified that children can recognize that deduction is more certain than guesses. However, study tasks exclusively asked children to reason about observable properties and all content was immediately available. The proposed
study will make it possible to examine the relative difference of reasoning about observable and unobservable information by comparing certainty judgments directly. The current task is informed by the procedures used by Pillow and colleagues (2010; 2006) and Tullos and Woolley (2009). We anticipate that children will report greater levels of certainty concerning concrete immediately observable tasks compared to unobservable properties.

Research Question 2: How does children’s ability to judge the certainty of information when evaluating evidence about observable vs. unobservable properties change with age?

Reasoning about observable and unobservable properties has been considered separately, thus prior research does not directly capture differences. Previous study tasks have typically considered one component (Fay & Klahr, 1996; Klahr & Chen, 2003; Pillow, 2002). Additionally, studies that consider reasoning about observable or unobservable properties have not incorporated a certainty judgment scale to capture children’s reasoning. The proposed study will assess observable or unobservable properties within the same task. This will make it possible to examine differences in children’s certainty judgments across age groups by comparing observable and unobservable properties directly.

Research Question 3: How do children’s certainty judgments differ across varying levels of evidence (conclusive, inconclusive, and guess) when reasoning about investigation outcomes?

At approximately 5- to 6-years-old children begin to recognize reasoning and guessing as distinct processes that are associated with different degrees of certainty (Pillow, 2002; Pillow et al., 2010; Pillow & Anderson, 2006). In the procedures by Pillow and colleagues the child
reached a conclusion and then rated their certainty. The current investigation will be an extension which investigates if children understand that conclusions based on reasoning are different from those based on guessing. The present study will specifically examine whether children recognize reasoning about unknown matters in a scientific context as different from guessing, or whether they understand it to be all speculative. The current task will include varying levels of evidence: conclusive, inconclusive, and guess.

Research Question 4: Does performance evaluating levels of evidence (conclusive, inconclusive, and guess) differ across age groups?

Sodian et al. (1991) examined first and second grade student’s ability to distinguish conclusive from inconclusive experiments. Results indicated that children of both age groups could distinguish a conclusive from inconclusive test. Nonetheless, results demonstrated that second graders (86% correct) performed significantly better than the first-grade students (55% correct). Furthermore, Fay and Klahr (1996) found 5-year-old children (76% correct) performed better than 4-year-olds (59% correct) when identifying a definitive outcome. Klahr and Chen (2003) extended the work of Fay and Klahr and found a significant difference between 4- and 5-year-olds in overall performance. Five-year-old children's performance increased from 39% correct to 51%, while four-year-old children's performance was not significant (35% to 37%). However, previous studies such as Sodian et al. (1991) included questions about whether one ‘would know’ or ‘not know’ an outcome and guessing was not included. Additionally, earlier investigation used a dichotomous response option and did not consider judgments of certainty as evaluative tools.
Results by Pillow and colleagues demonstrated that children aged 5- to 9-years-old identified a deduction as highly certain; however, younger children also overrate the certainty of guesses (Pillow et al. 2000, 2010; Pillow & Pearson, 2009, 2012). The present study will further examine when do children recognize that deductions are more certain than guesses, and how does certainty judgments change with age? We expect the certainty judgments of older children to be greater for conclusive items compared to younger children. Additionally, it is anticipated that children will improve at identifying guesses as uncertain with age.
CHAPTER 3

METHOD

Participants

This experiment was a cross-sectional study with 70 participants. Participants were 18 kindergarten students (10 females, 8 males; \( M \) age = 5 years, 5 month; \( SD = 4.31 \) months; age range = 5 years, 0 months to 6 years, 1 month), 18 first-grade students (10 females, 8 males; \( M \) age = 6 years, 8 months; \( SD = 8.99 \) months; age range = 6 years to 8 years, 3 months), 16 third-grade (7 females, 9 males; \( M \) age = 9 years, 5 month; \( SD = 5.52 \) months; age range = 8 years, 7 months to 10 years, 6 months), 18 adults (14 females, 4 males; \( M \) age = 22 years, 9 months; \( SD = 74.53 \) months; age range = 18-34 years). The children and adults were recruited through online research participation forums. Adults were also recruited from an entry-level psychology course at a large Midwestern University. The participant sample was diverse including Black/African American (42%), White/European American (40%), Hispanic/Latinx (9%), Asian/Southeast Asian (4.5%). The remaining participants elected not to disclose race or ethnic information (4.5%). All demographic information was provided by a parent/guardian for children or self-reported by adult participants. The study was approved by the university’s Institutional Review Board. Parents provided electronic written consent for participation and children gave spoken assent. Adults gave electronic written consent for participation. Study procedures were administered via Zoom Video Communications, an online video conferencing platform.
Materials

Participants used a computer or laptop with a camera and either speakers or headphones to communicate during the session. The experimenter shared their screen to direct the experimental procedure.

During the warm-up familiarization task, participants were shown an image of a familiar object and an image of an object expected to be unfamiliar on the screen separately. For the Box Task (*observable*), participants were presented with a series of pictures on the screen depicting colored boxes (e.g., a pair of identical blue boxes, or two blue boxes and two red boxes). Participants were told that there were keys or rocks in each box, and the corresponding sound effects were played. For the Animal Task (*unobservable*), participants were shown a novel animal, animal nest, and various animal food items (e.g., berries, worms, leaves). During each trial, a different food item was depicted as being left in the animal’s nest. All images were animated on the screen. The rating scale included an animated sequence of five vertical bars increasing in height from left to right (0.76, 2.54, 5.08, 7.62, 10.16 cm.) and width measured 3.33 cm. Each vertical bar of the 5-point scale had a corresponding number and description: 1-not sure at all, 2-a little sure, 3-medium, 4-kind of sure, 5-very sure. The rating scale will more accurately detect the differences in children’s evaluations of certainty for each trial that would not be captured by dichotomous response options.

Procedure

The experimenter interacted with participants in real-time via Zoom. The children and adults participated individually. The familiarization procedure was presented first, followed
by the Box Task or the Animal Task. Both tasks assessed children’s evaluation of the certainty of
inferences when reasoning across varying levels of evidence: conclusive, inconclusive, and
guess. Each child first received two warm-up questions followed by six trials of the main task
including three Box Task trials (observable) and three Animal Task trials (unobservable). The
order of the Box and Animal tasks was counterbalanced across participants (see Appendix A for
the complete study procedure).

The familiarization procedure was intended to acquaint participants with the certainty
rating scale and assess their ability to use it appropriately. Participants were shown a picture of a
familiar object (e.g., apple) and a picture of an unfamiliar object (e.g., a racing bicycle pedal).
They were asked to rate their certainty about the name of the object using the 5-point certainty
scale, 1 (not certain) to 5 (most certain). For example, children were presented a picture of an
apple and certainty scale and asked, “Do you know what this is called? How sure are you that
you know what this is called? Are you very sure or not sure at all?”). Children were encouraged
to point to the scale in addition to saying the corresponding words (e.g., “very sure”) or
indicating the associated number on the scale (e.g., “5”). The order of the two familiarization
trials was counterbalanced across participants.

The Box Task (observable) assessed participant’s ability to reason about immediately
observable properties using visual and auditory cues. The experimenter introduced the task with
a simple demonstration. Different color boxes were shown, and the participants were told that
some boxes have rocks inside them and others have keys inside them. Corresponding sound
effects were played when the bell or rattle sound was referenced. The experimenter then hid the
boxes behind a curtain. Next, part of a single box was revealed behind the animated curtain. The
experimenter pointed out that now they could see just the color of the box. There were three trials: one conclusive trial, one inconclusive trial, and one guess trial. For the trial, participants were shown pictures of different color boxes separately, two blue boxes with keys inside and two red boxes with rocks inside. Participants were asked corresponding knowledge check questions, “What’s inside the blue box?” and “What’s inside the red box?” Then the two blue boxes and two red boxes were shown together, and the experimenter explained, “Now I am going to hide one of these boxes”. The experimenter then showed the evidence and asked the certainty judgment question: (a) Evidence: “Look, it’s a [blue/red] box. Do you know if this [blue/red] box has a bell or rattle inside it?”, and (b) Certainty judgment: “How sure are you? Are you very sure or not sure at all?” For the observable inconclusive trial, two boxes of the same color were presented, but they each contained a different item inside. First, the experimenter explained that there were two orange boxes, one orange box had a rattle inside and the other orange box had a bell inside. Similarly, two purple boxes were presented, one with a bell inside and the other with a rattle inside. Corresponding sound effects were presented with each box. Participants were first asked two knowledge check questions for each orange and purple box (e.g., “What’s inside this orange box? Now what’s inside this orange box?”). Then the two orange boxes and two purple boxes were shown together, and the experimenter explained, “Now I am going to hide one of these boxes”. The experimenter then presented the evidence and asked the certainty question: (a) Evidence: “Look, it’s a [purple/orange] box. Do you know if this [purple/orange] box has a bell or rattle inside it?” and (b) Certainty judgment: “How sure are you? Are you very sure or not sure at all?” For the observable guess trial, participants were presented with eight boxes, two boxes for each of the four colors (gray, green, yellow, pink) and told that some of the boxes had
pennies inside and some had marbles inside. Corresponding sound effects were played with no association with any of the presented boxes. Participants were then asked the evidence and certainty questions: (a) Evidence: “Look, it’s a [pink/green/gray/yellow] box. Do you know if this [pink/green/gray/yellow] box has pennies or marbles inside it?”, and (b) Certainty judgment: “How sure are you? Are you very sure or not sure at all?” The experimenter displayed corresponding pictures and relevant sound effects for each trial. The certainty rating scale was presented simultaneously when the evidence was presented and the certainty judgment question was asked. The order of conclusive, inconclusive, and guess trials was counterbalanced across participants within each age group.

The Animal Task (unobservable) assessed children’s ability to reason about unobservable properties using animated visual cues. Participants were asked to reason about a new animal’s abilities that was not immediately observable by making an inference about the food that was left in the animal’s nest (see Appendix B for corresponding features across tasks). The experimenter started the Animal Task with a simple demonstration. First a food item was presented in a familiar environment and then the food item was shown in an animal nest (e.g., first picture were grapes in a tree and the next picture were grapes in the animal nest). The experimenter then presented an introduction sharing that there was a new island that was found, and people would like to learn more about the new animals on the island. The participants were told that animals that eat different foods are good at different things therefore, the task will ask them to look in the animals’ nests after the animals eat to learn what the animal is good at doing. For the unobservable conclusive trial, participants were shown a picture of carrots in a nest and then a picture of leaves in a nest separately. Participants were told that animals that eat certain foods are
good at doing different things. For example, “Some animals eat carrots and they drop carrots in their nests after they eat. All animals on the island that eat carrots are good at hearing sounds from far away,” and then “People have found that some animals eat leaves, and they drop leaves behind in their nests after they eat. All animals on the island that eat leaves are good at seeing things in the nighttime when it is dark.” Following the new information, a corresponding knowledge check question was asked (e.g., “What are animals that eat carrots good at doing?” and “What are animals that eat leaves good at doing?”). Participants were then shown an animation of the new animal leaping from the nest to reveal one of the food items (e.g., “Here is a new animal called a Moki. No one has seen the Moki up close, but it always drops food behind in its nest.”) The experimenter then showed the evidence and asked the certainty judgment question, Evidence: “Look, we found [carrots/leaves] in the Moki’s nest. Do you know if the Moki is good at hearing sounds from far away or good at seeing things in the nighttime?” Certainty judgment: “How sure are you? Are you very sure or not sure at all?” The unobservable inconclusive trial involved animals that ate the same food but were good at different abilities. Participants were shown a picture of berries in a nest and then a picture of corn in a nest. They were then told, “People have found that some animals eat berries and drop berries behind in their nests. Some animals that eat berries are good at seeing very small things but other animals that eat berries are good at smelling things from far away” and “People have found that some animals eat corn and drop corn behind in their nests. Some animals that eat corn are good at seeing very small things but other animals that eat corn are good at smelling things from far away.” Participants were then shown an animation of the new animal leaping from the nest to reveal one of the food items and told, “Here is a new animal called a Pema. No one has seen the Pema up
close, but it always drops food behind in its nest”. Participants were asked corresponding knowledge check questions when each food item and ability were presented (e.g., “What are animals that eat berries good at doing?”/“What’s another thing that animals that eat berries are good at doing?” and “What are animals that eat corn good at doing?”/“What’s another thing that animals that eat corn are good at doing?”). Participants were then shown an animation of the new animal jumping from the nest to reveal either the berries or corn. The evidence was then presented followed by the certainty judgment question. For example, “Here is a new animal called a Pema. No one has seen the Pema up close, but it always drops food behind in its nest. Evidence: “Look, we see [berries/corn] in the Pema’s nest. Do you know if the Pema is good at seeing very small things or good at smelling things from far away?” Certainty judgment: “How sure are you? Are you very sure or not sure at all?” For the unobservable guess trial, the experimenter explained that there was a new part of the island that was discovered, and people wanted to learn about the new animals’ abilities. Participants were shown animated pictures of four different types of foods and told, “We found a new part of the island with many new animals! Some of the animals eat worms, nuts, oranges, or peas. The animals on the island are good at doing different things. Some animals are good at seeing things from far away and some animals are good at hearing very quiet sounds.” The participants were asked knowledge check questions (e.g., “What are some of the new animals good at doing?” and “Tell me something else that some of the new animals are good at doing?”) followed by the presentation of the evidence and the certainty judgment. Participants were then shown an animation of the new animal leaping from the nest to reveal one of the four food items. Participants were told, “Here is a new animal called a Chiboo. No one has seen the Chiboo up close, but it always drops food behind in
its nest”. The evidence was then presented, Evidence: “Look, we see [worms/nuts/oranges/peas] in the Chiboo’s nest. Do you know if the Chiboo is good at seeing things from far away or good at hearing very quiet sounds?” followed by the Certainty judgment: “How sure are you? Are you very sure or not sure at all?” The order of conclusive, inconclusive, and guess trials for the Animal Task was counterbalanced across participants within each age group.

The study procedure was consistent for adults, with the addition that adult participants were informed that their responses would be used as a comparison with children and notified that the procedure had been designed for use with kindergarten and elementary school children.

Scoring

The certainty rating scale contained the numbers 1 (not certain) to 5 (most certain). For the familiarization task, mean ratings for the familiar and unfamiliar images were calculated for each age group (kindergarten, first grade, third grade, adult). Likewise, for the six experimental trials, a mean certainty rating was calculated for each combination of Condition (observable vs. unobservable) and Evidence Level (conclusive vs. inconclusive vs. guess) for each age group.
CHAPTER 4

RESULTS

To examine participants’ ability to use the rating scale appropriately, certainty ratings for the familiarization procedure were analyzed. For the familiar image, participants’ mean certainty ratings were kindergarten: $M = 4.67$, $SD = 0.59$; first grade: $M = 5.00$; third grade: $M = 5.00$; adult: $M = 5.00$. First grade, third grade, and adults all rated the familiar item as 5, thus the standard deviation for the three age groups was 0. For the unfamiliar image, participants’ mean certainty ratings were kindergarten: $M = 1.78$, $SD = 0.94$; first grade: $M = 1.50$, $SD = 1.09$; third grade: $M = 1.87$, $SD = 1.12$; adult: $M = 1.11$, $SD = 0.32$. A $4 \times 2$ (Age $\times$ Familiarity) analysis of variance (ANOVA) yielded a significant main effect of Familiarity, $F(1, 65) = 838.00$, $p < .001$, partial $\eta^2 = 0.93$, $MSE = 0.46$. All age groups rated themselves as more certain about the identity of the familiar object compared to the unfamiliar object. T-tests were computed to compare the performance of each age group to chance in the familiarization procedure. On each familiarization trial, the probability of choosing the correct answer by chance was 2.50. For the familiar object, performance was above chance for kindergarten: $t(17) = 15.47$, $p < .001$. First grade, third grade, and the adults all rated the familiar object as 5, thus a t-value could not be computed because the standard deviation was 0. Performance was above chance across age groups for the unfamiliar object: kindergarten: $t(17) = 15.47$, $p < .001$; first grade: $t(17) = -3.86$, $p = .001$; third grade: $t(14) = -2.18$, $p = .04$; adults: $t(17) = -18.22$, $p < .001$. All age groups demonstrated strong understanding of how to use the rating scale appropriately.
For the experimental procedure, mean certainty ratings and response patterns were analyzed with a 4×2×3 ANOVA (Age × Condition × Evidence Level), with Age (kindergarten, first grade, third grade, adults) as a between-subjects factor and Condition (observable vs. unobservable) and Evidence Level (conclusive vs. inconclusive vs. guess) as within-subjects factors. The mean certainty ratings and standard deviation by age and level of evidence for the Observable and Unobservable conditions is presented in Table 1. The ANOVA yielded a significant main effects of Evidence Level, $F(1, 130) = 42.51, p < .001$, partial $\eta^2 = 0.39$, $MSE = 1.52$, which was qualified by a significant Age × Evidence Level interaction, $F(6, 130) = 3.93, p = .001$, partial $\eta^2 = 0.15$, $MSE = 1.52$, a significant Condition × Evidence Level interaction, $F(2, 130) = 8.48, p < .001$, partial $\eta^2 = 0.11$, $MSE = 0.88$, and a significant Age × Condition × Evidence Level interaction $F(6, 130) = 2.39, p = .03$, partial $\eta^2 = 0.09$, $MSE = 0.88$.

To examine these effects further, separate 4 × 3 ANOVAs (Age × Evidence Level) were conducted for each Condition (observable vs. unobservable). For the Observable condition, there was a significant effect of Evidence Level, $F(2, 130) = 57.78, p < .001$, partial $\eta^2 = 0.47$, $MSE = 0.99$, and a significant Age × Evidence Level interaction, $F(6, 130) = 2.93, p < .01$, partial $\eta^2 = 0.11$, $MSE = 0.88$. Mean certainty ratings by age and level of evidence for the Observable condition are shown in Table 1. Separate one-way ANOVA’s yielded significant effect of Evidence level for each age group (kindergarten: $F(2, 32) = 9.18, p = .001$, partial $\eta^2 = 0.36$, $MSE = 0.86$, Tukey’s honestly significant difference (HSD) = 0.78; first grade: $F(2, 34) = 10.33, p < .001$, partial $\eta^2 = 0.38$, $MSE = 1.11$, Tukey’s HSD = 0.86; third grade: $F(2, 30) = 13.02, p <$
Table 1

Mean Certainty Ratings (and Standard Deviation) by age and evidence level for the Observable and Unobservable conditions.

<table>
<thead>
<tr>
<th>Age</th>
<th>Conclusive</th>
<th>Inconclusive</th>
<th>Guess</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Kindergarten</td>
<td>4.71&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.59</td>
<td>3.35&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>First Grade</td>
<td>4.61&lt;sup&gt;cd&lt;/sup&gt;</td>
<td>0.61</td>
<td>3.39&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Third Grade</td>
<td>4.69&lt;sup&gt;ef&lt;/sup&gt;</td>
<td>0.87</td>
<td>3.38&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>Adult</td>
<td>4.89&lt;sup&gt;gh&lt;/sup&gt;</td>
<td>0.47</td>
<td>2.67&lt;sup&gt;g&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>Conclusive</th>
<th>Inconclusive</th>
<th>Guess</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Kindergarten</td>
<td>3.56&lt;sup&gt;ij&lt;/sup&gt;</td>
<td>1.29</td>
<td>4.17</td>
</tr>
<tr>
<td>First Grade</td>
<td>4.17</td>
<td>0.86</td>
<td>3.44</td>
</tr>
<tr>
<td>Third Grade</td>
<td>4.44&lt;sup&gt;ik&lt;/sup&gt;</td>
<td>0.81</td>
<td>3.56</td>
</tr>
<tr>
<td>Adult</td>
<td>4.83&lt;sup&gt;ilm&lt;/sup&gt;</td>
<td>0.38</td>
<td>3.28&lt;sup&gt;l&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Note. Scale ranges from 1 to 5. Higher numbers correspond to greater certainty. For each age group, means that have superscripts in common are significantly different from each other.
.001, partial $\eta^2 = 0.46$, $MSE = 0.82$, Tukey’s HSD = 0.79; adult: $F(2, 34) = 32.43, p < .001$, partial $\eta^2 = 0.66$, $MSE = 1.13$, Tukey’s HSD = 0.95). The certainty response patterns across levels of evidence for kindergarteners were different from first graders, third grades, and adults. Kindergarten children gave significantly higher certainty ratings for the conclusive compared to the inconclusive trials ($p = .001$, 95% C.I. = [0.53, 2.18]), but no statistical difference was found between conclusive and guess trials or inconclusive and guess trials for kindergarteners. In contrast, first graders, third graders, and adults, reported significantly higher certainty ratings for the conclusive trials compared to the inconclusive and conclusive trials compared to the guess trials (first grade: conclusive vs. inconclusive, $p = .005$, 95% C.I. = [0.35, 2.09], conclusive vs. guess, $p = .001$, 95% C.I. = [0.58, 2.42]; third grade: conclusive vs. inconclusive, $p = .007$, 95% C.I. = [0.34, 2.29], conclusive vs. guess, $p = .002$, 95% C.I. = [0.58, 2.42]; adult: conclusive vs. inconclusive, $p < .001$, 95% C.I. = [1.18, 3.26], conclusive vs. guess, $p < .001$, 95% C.I. = [1.68, 3.65]. However, no statistical difference was found between the inconclusive and guess trials for first graders, third graders, and adults in the Observable condition. Figure 1 shows the mean certainty ratings by age and level of evidence for the Observable condition.

For the Unobservable condition, there was a significant effect of Evidence Level $F(2, 132) = 10.39, p < .001$, partial $\eta^2 = 0.37$, $MSE = 1.40$, and a significant Age $\times$ Evidence Level interaction $F(6, 132) = 3.99, p = .001$, partial $\eta^2 = 0.15$, $MSE = 1.40$ (see Table 1). Separate one-way ANOVA’s yielded significant Evidence level effects for third grade, $F(2, 32) = 4.20, p = .02$, partial $\eta^2 = 0.23$, $MSE = 0.97$; Tukey’s HSD = 0.86) and adults $F(2, 34) = 12.87, p < .001$,
Figure 1: Mean certainty ratings by age and evidence level for the Observable condition.

partial $\eta^2 = 0.43$, $MSE = 1.61$, Tukey’s HSD = 1.04) but not for kindergarten or first grade children. Third graders rated conclusive trials as significantly higher than inconclusive trials ($p = .04$, 95% C.I. = [0.03, 1.72]), but there was no statistical difference between conclusive and guess or inconclusive and guess trials. Adults’ ratings were significantly higher for conclusive compared to inconclusive trials ($p = .007$, 95% C.I. = [0.39, 2.72]) and between conclusive and guess trials ($p = .001$, 95% C.I. = [0.85, 3.26]) but no statistical difference was found between inconclusive and guess trials. Figure 2 shows the mean certainty ratings by age and level of evidence for the Unobservable condition.
Table 1 shows the mean certainty ratings and standard deviation by age and evidence level for both the Observable and Unobservable conditions. In addition to demonstrating differences within age groups as previously described, the table presents information across age groups. For example, kindergarten children reported significantly higher observable guess ratings compared to the adult participants during the Observable condition. Kindergarteners reported greater variation in responses across both conditions (e.g., kindergarten: observable inconclusive = 3.35 and kindergarten: unobservable inconclusive = 4.17. First, third, and adult participants reported similar ratings for both conditions across levels of evidence (e.g., first grade: observable guess = 3.11 and first grade: unobservable guess = 3.89; third grade:
observable guess = 3.19 and third grade: unobservable guess = 3.56; adult: observable guess = 2.22 and adult: unobservable guess = 2.78). Results demonstrated less variance for older children and adults concerning rating responses across conditions.

Results indicate a difference in reasoning between observable and unobservable evidence across levels of evidence. This is particularly evident when comparing kindergarten certainty ratings to the older age groups. As predicted, all age groups successfully distinguished conclusive items as more certain than the inconclusive and guess items in the Observable condition. However, kindergarten children rated the guess trial which required reasoning about a random outcome as more certain than the inconclusive trials which present an intermediate level of evidence. First and third grade children demonstrated understanding of the varying level of evidence as they rated the guess trial as less certain than the inconclusive trial. Adult responses indicate a strong differentiation of the levels of evidence and reported a greater difference by rating inconclusive and guess trials as much lower certainty compared to the kindergarten and elementary school age participants.

Kindergarten children did not differentiate evidence level in the Unobservable condition. Certainty ratings were higher for inconclusive compared to the conclusive trial. Similarly, first grade children demonstrated difficulties distinguishing between levels of evidence which differs from their Observable condition performance. Patterns of response demonstrate that first grade children did report the conclusive trial as slightly more certain but the differences between the inconclusive and guess were marginal. Third grade children displayed some improvement as they reported significantly greater certainty for conclusive compared to inconclusive trials. Adults reported greater certainty for the conclusive trials for both inconclusive and guess trials.
However, third graders and adults’ certainty ratings for the inconclusive trial were not significantly different from guess trials. Nonetheless, the range of certainty ratings was much greater across the three levels of evidence for adults suggesting greater recognition between the levels of evidence. Participants’ certainty ratings suggest a progressive differentiation between conclusive and inconclusive trials with increased age, especially for the Unobservable outcomes.
CHAPTER 5

DISCUSSION

The present study investigated children’s evaluation of the certainty of conclusions during scientific reasoning tasks. The objectives of the current study were (a) to examine children’s reasoning about observable vs. unobservable properties, (b) compare children’s judgments of the certainty of inferences when reasoning about varying levels of evidence, and (c) assess differences in evaluations of certainty across age groups.

Science inquiry relies on applying logical reasoning processes to the evidence available to make inferences and generate appropriate outcomes. Thus, scientists and researchers often examine information and draw conclusions about properties or constructs that are not immediately verifiable. They must evaluate evidence to draw conclusions or determine that it is necessary to reexamine and revise previous conclusions. Prior investigations have involved children reasoning about temporary situations, observable items, and immediately verifiable conclusions (Klahr & Chen, 2003; Piekny & Maehler, 2013; Pillow et al., 2002, Pillow & Anderson, 2006; Pillow & Pearson, 2012; Vilma, 2019). Previous studies did not investigate how reasoning about observable vs. unobservable properties or information that was not immediately verifiable could influence children’s reasoning. The current investigation presented children with evidence and asked them to make conclusions about observable properties vs. unobservable properties while evaluating varying levels of evidence. Additionally, previous scientific reasoning investigations have utilized dichotomous task paradigms resulting in limited response
options. The current study included a certainty judgment rating scale to capture reasoning responses.

To investigate the study objectives the procedure included two Conditions (observable vs. unobservable) each involved reasoning about three Evidence levels (conclusive vs. inconclusive vs. guess). For the Observable condition all age-groups rated the conclusive item as more certain than the inconclusive and guess items. Participants’ certainty ratings indicated that children and adults could recognize and provide different certainty responses to varying levels of evidence. However, results suggest this differentiation was limited to the Observable condition for the younger children. In the Unobservable condition both kindergarten (approximately 5 years old) and first grade (approximately 6 years old) children demonstrated difficulties discriminating between levels of evidence. By third grade (approximately 9 years old) children begin to accurately differentiate the levels of evidence during the Unobservable condition. Adults consistently reported conclusive evidence as more certain than inconclusive and guess information for both the Observable and Unobservable conditions. These results add to the literature on scientific reasoning development, specifically the extent of abilities when evaluating varying levels of evidence.

The following sections will discuss the results of the present study in relation to the study objectives and previous investigations. The results of the Observable condition will be discussed by age group followed by a description of the consistent findings between the current results and findings of previous investigations that used observable properties. This section will conclude with a discussion of the patterns of responses that were displayed across age groups during the Observable condition. The ensuing section will focus on the Unobservable condition with an
emphasis on current study and the patterns of responses across age groups. Performance in the Unobservable condition compared to the Observable condition across age groups will then be discussed. The following section will consider possible interpretations of results and alternative explanations as to why children performed differently across the two conditions. The final section will consider the 5-point certainty rating scale that was implemented in the study procedure.

The Observable condition examined the ability to make accurate inferences about immediately observable properties during evidence evaluation. The study task (Box Task) assessed children’s reasoning concerning readily observable items (e.g., different color boxes) and asked them to make an inference about an item inside the box based on an observable property (e.g., color of box). Results demonstrated that children distinguished among different levels of evidence in their certainty judgments. As expected, all age groups on average rated the conclusive items as more certain than inconclusive and guess items. Contrary to expectations, certainty judgments of first and third grade children were not greater than kindergarten children for the conclusive item during the Observable condition. There were no age differences when reasoning about the conclusive evidence. This indicates that kindergarteners responded the same as first graders, third graders, and adults for the observable conclusive trial. However, kindergarten children demonstrated different patterns of responses compared to the older children and adults when reasoning about the inconclusive and guess items. Kindergarten children reported the observable guess trial as more certain than the observable inconclusive trial. The response patterns for the first graders, third graders, and adults were parallel. First graders, third graders, and adults accurately rated the observable guess trial as less certain
compared to the observable inconclusive trial. However, for the observable guess trial kindergarten children gave higher certainty ratings than did adults. The primary aim of the investigation was to examine children’s logical reasoning processes. However, an adult comparison group was included to compare and evaluate performance across age groups. As expected, adults demonstrated a mature understanding when reasoning across the varying levels of evidence. Findings are consistent with previous developmental investigations which included an adult comparison group (Pillow & Pearson, 2009; 2012; Pillow et al. 2010. The current study demonstrated evidence for early competence of basic scientific reasoning abilities. By 5-years-old children can successfully identify conclusive from inconclusive or guess information when reasoning about observable properties. This is counter to Kuhn and colleagues’ assertion (Kuhn, 2011; Kuhn & Pearsall, 2000; Kuhn et al., 1988) that children are unable to correctly evaluate evidence information prior to approximately age 10-11 years old. However, it is important to note that the current investigation exclusively considered evidence evaluation abilities, it did not include theory-based reasoning items that were incorporated into Kuhn’s scientific reasoning tasks. The current study aimed to examine if children can use the presented evidence to draw appropriate conclusions. Results provided further support for the early emergence of scientific reasoning skills. This is particularly evident when task procedures involve concrete items, immediately available outcomes, and the evidence information is conclusive.

The present results are consistent with previous literature indicating children’s ability to successfully reason about observable properties. Prior investigations involved evaluating experiments or designing an appropriate experiment for testing using concrete items (Klahr & Chen, 2003; Piekny et al., 2013, 2014; Sodian et al., 1991; van der Graaf et al., 2016). Chen and
Klahr (1999) examined children’s (aged 7-10) ability to successfully engage in physical tasks during scientific reasoning (e.g., canal task, spring task, ramps task). The procedure involved children reasoning about variables and corresponding items of different weights, sizes, and positions. Additionally, studies by Pillow and colleagues (e.g., 2006, 2009, 2015) involved participants reflecting on their own reasoning using certainty judgments ratings. Results demonstrated that children 5- to 9-years-old could make accurate inferences when reasoning about tangible items. Study procedures exclusively involved reasoning about temporary situations, observable items (e.g., stuffed animals, marbles, toy horses) and reasoning about observable properties such as color and size (e.g., red toy dog/yellow toy dog or small toy bear/big toy bear). Prior investigations consistently used observable properties during developmental assessments.

The present results are consistent with previous literature indicating children’s ability to successfully reason about varying levels of evidence when reasoning about observable properties. Scientific reasoning procedures have used observable properties to examine children’s ability to evaluate levels of evidence (Koerber et al., 2005; Kuhn, 1989; Piekny et al., 2014; Piekny & Maehler, 2013).

In a similar task as Koerber et al., (2005), Piekny and colleagues presented children aged 4-13 pictures that depicted children with healthy and unhealthy teeth and asked to reason about the presented evidence. Results demonstrated that children could reason about covariation evidence outcomes particularly when the evidence was an example of perfect covariation (e.g., 20-picture set with 10 children with healthy teeth and green chewing gum and 10 children with unhealthy teeth and red chewing gum). Fay and Klahr’s (1996) target box paradigm presented 4-
and 5-year-old children with a target object (e.g., curved pieces of construction paper) and then showed two boxes with varying shaped construction pieces. Results demonstrated that preschool children can successfully differentiate determinate (e.g., curved with square construction pieces vs. squared with straight construction pieces) from indeterminate (e.g., curved with square construction pieces vs. curved with straight construction pieces) particularly if the materials are concrete, all the information is visually available (e.g., each available box is open), and the outcome allows for immediate verification (e.g., reviewing contents of each available box).

The literature demonstrates the emergence of early competence in evidence evaluation abilities in some situations. However, different patterns of responses found for the observable versus the unobservable condition in the current investigation provide additional support that evidence evaluation skill is incomplete prior to approximately 6-years-old and may improve with age. As anticipated, all age groups rated conclusive evidence as the most certainty. However, in the Observable condition kindergarten children (approximately aged 5) rated the guess item as more certain than the inconclusive item. Kindergarten children performed better at evaluating conclusive information than inconclusive information. However, by approximately 6-years-old children’s ratings were similar to the adult pattern. First graders (approximately aged 6), third graders (approximately aged 9) and adults rated the guess item as the least certain evidence level. Kindergarten children may experience a greater challenge when reasoning about inconclusive evidence.

The present results are similar to previous investigations that showed inconsistent patterns of response for younger participants. Tullos and Woolley (2009) examined children’s ability to make an inference about determinant vs. indeterminate evidence outcomes. Children
aged 4- to 8-years-old were shown a set of animals and asked to make an inference about the reality status of new animals based on the evidence presented. Children were presented three different types of evidence: supporting, irrelevant, and no evidence when asked to expect the contents of the evidence box. Results demonstrated that 6- to 8-year-olds could differentiate the varying levels of evidence and could make correct inferences about the reality status of the animal based on the evidence they observed. However, 4- to 5-year-olds had difficulty differentiating the different levels of evidence which resulted in poor performance. Fay and Klahr’s (1996) target box paradigm demonstrated that 4- and 5-year-old children had difficulty reasoning about indeterminate outcomes. For example, when children were asked, “Is this a time that you can tell which box I used to make this or is this a time that you cannot tell which box I used?”. Outcomes in which the construction pieces that were identical to the target object were presented in multiple boxes (e.g., curved with square construction pieces vs. curved with straight construction pieces) instead of just one box presented increased difficulty for the younger children.

Additionally, Pillow and colleagues found that young children have difficulties evaluating inconclusive or guess information and have the tendency to incorrectly rate guesses as equally certain than deductive inferences that are conclusive (Pillow et al., 2000; Pillow & Pearson, 2009). Pillow et al., examined evidence evaluation abilities by asking participants to differentiate a deductive inference compared to indeterminate evidence resulting in an inductive inference or (Pillow & Anderson, 2006; Pillow & Pearson, 2012; Pillow et al., 2010). Results demonstrated that children as young as 5 years old can consistently report greater certainty for deductive inference than guessing (Pillow 2002; Pillow & Anderson, 2006; Pillow and Pearson,
2009). However, younger children (aged 4-5) did not differentiate levels of evidence and the strength of the induction influenced certainty judgments. By 8- to 9-years-old, children demonstrated improvements in reasoning by successfully identifying that deductions were more certain than strong inductions and strong inductions were more certain than weak inductions and guesses (Pillow & Pearson, 2012; Pillow et al., 2010). Pillow and Pearson (2009, 2012) showed that 6-8-year-olds can distinguish their own and another person’s deductive inferences as more certain than guesses. However, 4-5-years-old exhibited inconsistent response patterns as they did not report differences in certainty judgments between inferences and guess items. It was concluded that by age 4 children display basic understanding of deduction however, young children (aged 4-5) report more inconsistencies in response patterns when asked to reason about varying levels of certainty.

As suggested by prior and current findings, the ability to differentiate levels of evidence may be incomplete for younger children. That is, conclusive information with observable items could be more accessible to younger children particularly, if the information is concrete, involves familiar properties (e.g., color, size, shape), and the outcomes are definitive. This situation provides a greater simplicity that younger children can successfully engage in. However, the increase in ambiguity or the lack of a definitive outcome may cause reasoning about inconclusive evidence or guess situations to present more of a challenge even if reasoning about observable properties.

Furthermore, kindergarten children may not have identified the task as having three distinct levels of evidence. That is, kindergarten children may not have recognized the difference in certainty between the inconclusive information (intermediate evidence outcome) compared to
the guess information. Thus, kindergarten certainty ratings for inconclusive and guess resulted in an inconsistent pattern of response compared to the first graders, third graders, and adults. The difference between inconclusive and guess information is relatively subtle and neither leads to a certain outcome. It is possible that the inconclusive information increases difficulty particular for kindergarten children but continues to develop during the early elementary school years. The current results indicate that by first grade (approximately 6 years old) children can successfully make this differentiation and reason about distinct levels of evidence when presented with observable properties. The current results show that children can recognize that conclusive evidence is more certain than inconclusive and guess information when reasoning about observable information.

The Unobservable condition examined children’s reasoning about unobservable properties during scientific reasoning tasks. Additionally, it examined children’s reasoning about varying levels of evidence when the presented information was not readily available or verifiable. Scientists use evidence and reasoning to draw conclusions about properties or events that they do not yet know or are not immediately perceptible. Thus, the Unobservable condition was more comparable to real world scientific examples. The Unobservable condition (Animal Task) asked children to make an inference about an unobservable internal property (e.g., hearing sounds from far away) of a fictional animal based on the type of animal food that is found in the animal’s nest.

The current study showed that younger children had difficulty accurately differentiating levels of evidence during the Unobservable condition. Patterns of responses demonstrate that there was a lack of understanding about the connection between the levels of evidence and
corresponding degrees of certainty. Students in kindergarten reported greater certainty for the 
*unobservable inconclusive* trial compared to the *unobservable conclusive* and *unobservable
guess* trials. First graders rated the *unobservable conclusive* trial as the most certain but reported
the *unobservable guess* trial as more conclusive than the *unobservable inconclusive* trial. Third
grade children reported the *unobservable conclusive* trial as the most certain and reported nearly
identical certainty ratings for both the *unobservable inconclusive* and *unobservable guess* trials.
The adult participants exhibited a clear understanding of the varying level of evidence. Adults
reported greater certainty ratings for the *unobservable conclusive* trial followed by the
*unobservable inconclusive* trial and the *unobservable guess* trial was identified as the least
certain. Kindergarten children did not identify varying levels of evidence in their certainty
judgments. First grade children rated the conclusive item as more certain than inconclusive and
guess but reported the guess item as more certain than the inconclusive item. However, by third
grade children rated the conclusive information as more certain and rated the inconclusive and
guess items being less certain.

The current study allowed for a direct comparison between Unobservable condition
(Animal Task) with the Observable condition (Box Task). This comparison provided an
extension of previous work that relied on observable properties. There exists a debate within the
literature concerning the extent to which young children can accurately reason about observable
compared to unobservable or non-obvious properties. Early investigations revealed low
performance when children were asked to reason about properties they could not directly observe
(Bruner et al., 1966; Carey, 1985; Piaget, 1929). Researchers argued that young children were
unable to correctly evaluate information about non-obvious properties (Carey, 1985, 1988;
Piaget, 1929). Carey (1985, 1988) concluded that children aged 4 to 5 have a limited understanding of internal properties and it was not until 10 years old that children can consistently understand and correctly reason about internal biological properties. In contrast, Hatano and Inagaki (2008) suggested that 5-year-old children have a rudimentary understanding of internal biological properties and can use this information to identify categorical differences (e.g., ability to grow or not) between of living (e.g., plants and humans) and nonliving examples. Additionally, Piekny and Maehler (2013) used a modified version of Sodian et al.’s (1991) mouse task to evaluate children’s ability to differentiate a conclusive from inconclusive experiment. Participants were presented with a story introduction and tasked with identifying the size of a mouse (e.g., big mouse or small mouse) considering the features of the ‘mouse houses’. Children were informed that the big mouse could only enter the ‘mouse house’ with the big door however, the small mouse could enter the ‘mouse houses’ with the big door and the ‘mouse house’ with the small door. For example, children were asked in what ‘mouse house’ should the ‘food’ be placed for the mouse to access the food to determine the size of the mouse. Results by Piekny and Maehler (2013) and Sodian et al.’s (1991) determined that by 5- and 6- year-olds could identify that the food should be placed in the small ‘mouse house’ (conclusive test) vs. the big ‘mouse house’ (inconclusive test). Results demonstrate that children can correctly use evidence to make inferences about events that are not directly observable. However, the procedure of the mouse task involved reasoning about observable properties (e.g., size of mouse and ‘mouse house’ door). The current study design provided a method to explore the developmental trajectory when reasoning about observable vs. unobservable properties within the same task.
Additionally, the study design allowed an assessment of evidence evaluation when the information is not immediately available or verifiable. Prior studies have predominantly used observable properties that were verifiable to examine reasoning in early childhood (Koerber et al., 2005; Kuhn, 1989; Pillow et al., 2010; Tullos & Woolley, 2009). Previous investigations did examine children’s reasoning about unobservable properties while reasoning about various levels of evidence. The current procedure also examined the ability to reason about unknown information and not temporary facts used in previous investigations. Previous studies demonstrated that by age 5 children can effectively evaluate their own reasoning and differentiate levels of certainty (Ghetti et al., 2013; Pillow & Anderson, 2006; Pillow & Pearson, 2012). Findings by Pillow and colleagues (Pillow et al., 2000, 2010; Pillow & Pearson, 2009) illustrate a detailed record of the capacity of children to reason about varying levels of evidence. For example, children as young as 4-5-years old can report greater certainty during deductive inference than guessing (Pillow et al., 2000; Pillow & Pearson 2009). Additionally, kindergarten children began to distinguish deduction information from induction and guessing (Pillow et al., 2010).

The results of the Unobservable condition were informative as it allowed a direct comparison to the Observable condition. The Unobservable condition involved reasoning about unknown information about characteristics and corresponding skills that were not directly observable. Findings suggest that kindergarten children are unable to differentiate levels of evidence when reasoning about unobservable properties. Kindergarten children’s patterns of response were different than the first graders, third graders, and adults for the Unobservable condition. Results show that kindergarten children reported the inconclusive item as more certain.
than the conclusive item. Kindergarten children were the only age group to report a conclusive item as less certain than inconclusive and guess item for either condition. Results suggest that by first grade children can begin to reason about unobservable properties. The first graders in the current study correctly identified the conclusive evidence as more certain than inconclusive and guess evidence. However, first graders reported the guess item to be more conclusive than the inconclusive item. This result is suggestive of the potential increase in difficulty when reasoning about unobservable properties since the same first grade students reported accurate certainty response patterns during the Observable conditions. Patterns of response for kindergarten and first grade children indicate that it is not until approximately third grade that children can evaluate varying levels of evidence when information is not observable or readily verifiable. Previous investigations showed that preschool children can differentiate varying levels of evidence (Fay & Klahr, 1996; Klahr & Chen, 2003) involving observable properties, however, the present study suggests that this early evidence evaluation skills may not directly transfer when young children are reasoning about unobservable properties. The results emphasize that reasoning about observable properties vs. unobservable properties can impact reasoning particularly for children prior to approximately 9-years-old.

The increase in difficulty during the Unobservable condition may lead to the age-related changes exhibited. Current results show that the kindergarten and first graders had found it challenging to reflect on their own reasoning and report correct certainty ratings; however, third graders exhibited improvements. The Unobservable condition revealed age-related improvements demonstrated in prior investigations (Pillow, 2002; Pillow et al., 2000, 2010). By approximately 8-9-years old children can more accurately differentiate levels of evidence and
provide different types of explanations related to each (Pillow et al., 2010). As anticipated children demonstrated improvements at identifying guesses as uncertain with age which was especially evident during the Unobservable condition. By third grade, children reported the conclusive items as the most certain and reported nearly the same certainty ratings for the inconclusive and guess trials. The adult participants reported greater certainty ratings for the conclusive items followed by the inconclusive and the guess items were identified as the least certain. Results demonstrated that by third grade (approximately aged 9) children can differentiate varying levels of evidence when reasoning about unobservable properties. Adults displayed a consistent understanding of evidence evaluations during the task. Results show a statistical difference in mean values for the unobservable inconclusive trials between kindergarten and third grade students and kindergarten students and adults. Prior investigations have demonstrated that children have difficulty reasoning about inconclusive or intermediate evidence outcomes. This seems to be particularly evident for younger children aged 4- to 5-years old (Fay & Klahr, 1996; Klahr & Chen, 2003; Pillow & Anderson, 2006). Thus, reasoning about inconclusive evidence during the Unobservable condition may have presented an increased challenge for younger children particularly when reasoning about the inconclusive and guess items. This is particularly evident for the kindergarten children as they reported the least accurate responses during the Unobservable condition.

The difficulty for kindergarten and first grade children to report accurate certainty ratings during the Unobservable condition is potentially due to an increased reasoning challenge during the task. As indicated by prior research young children can identify conclusive items but more ambiguous levels of evidence or guessing can be more challenging (Pillow, 2002; Pillow &
Anderson, 2006; Pillow & Pearson, 2012). The Unobservable condition involved reasoning about an unknown internal property of a novel animal. The task presented an additional reasoning challenge that was not present in the Observable condition. Thus, differentiating three levels of evidence specifically inconclusive and guess information while the information is not directly observable or verifiable increased the complexity for kindergarten and first grade children resulting in inaccurate certainty responses. However, various factors could influence children’s certainty judgments.

Working memory capacity must be considered in relation to performance differences across the Observable and Unobservable conditions. Working memory involves the cognitive processes that temporarily maintain and manipulate information without the presence of extremal cues to complete a mental task (Best & Miller, 2010; Cowan, 1998; Huizinga et al., 2006). There is significant evidence demonstrating the relationship between working memory capacity and reasoning performance (Best & Miller, 2010; Handley et al., 2004; Simms et al., 2018; Süß et al., 2002). For example, Handley et al., (2004) found that a measure of working memory predicted logical reasoning task performance, but the same measure was independent of belief-based problems in a sample of 10-year-old children. Results suggest a critical function of working memory particularly during logical reasoning mental tasks. Researchers have also reported that the task complexity influences working memory (Best & Miller, 2010; Simms et al., 2018). That is, an increase in task difficulty requires additional working memory demands. Changes in working memory development have been demonstrated to impact performance during reasoning tasks. Simms et al. (2018) demonstrated that 5- to 11-year-old children’s individual differences in working memory predicted performance during a reasoning mapping task. Children completed
three executive functioning and working memory tasks (e.g., List Sorting Working Memory task) and an analogical mapping task depicted with sets of pictures. Results revealed the working memory battery to be the most accurate predictor of the reasoning task, highlighting the relationship between working memory development and reasoning task performance. Studies have demonstrated a gradual increase of working memory improvement from childhood into adolescence in which skills (e.g., improved ability to hold and manipulate more information and perform tasks that require greater mental representation demands) become more aligned with adult performance (Best & Miller, 2010; Simms et al., 2018; Süß et al., 2002). Huizinga et al. (2006) examined executive functioning mechanisms across four age groups (aged 7, 11, 15, and 21). Results demonstrated continued improvement of working memory performance across age groups providing evidence for the developmental trend involved in working memory capacity.

Working memory demands may have differed across the two conditions in the present study. First, the perceptual experience and the familiarity of the items were distinct for each condition. The Observable condition was designed to present sensory information that could be immediately experienced (e.g., box color, sound of ringing bell). Additionally, though the corresponding features during each condition were parallel (see Appendix B), the Observable condition allowed for reasoning about familiar properties such as box color and sound (e.g., ringing). That is, participants were asked to draw conclusions about a potential item inside a box (e.g., a bell) based on the evidence presented. In contrast, the Unobservable condition required participants to reason about verbal information concerning an animals’ diet and concomitant abilities. Specifically, successful reasoning would involve maintaining the specifics about an animal food item (e.g., leaves) and its corresponding ability (e.g., “good at seeing things in the
nighttime when it is dark”) while reporting about the evidence being presented. Although the particular pairings of color and sound cue were novel, the colored boxes and potential items inside the boxes (e.g., bells, marbles, keys) are familiar everyday objects to children, whereas animal diets, internal properties, and unobservable abilities are not concepts children typically encounter. Thus, the Observable condition may have allowed for reasoning about more memorable information as it included a direct perceptual experience and concrete properties involving familiar items. In contrast, during the Unobservable condition the working memory demands were potentially increased due to the lack of familiarity with the condition components which involved reasoning that was less concrete. Second, working memory demands potentially differed between the two conditions due to the differences in complexity. Researchers have shown that reasoning about inconclusive information (Fay & Klahr, 1996; Klahr & Chen, 2003; Piekny & Maehler, 2013) or uncertainty (Flavell et al., 1985; Lyons & Ghetti, 2011; Pillow & Anderson, 2006) presents increased difficulty throughout childhood that improves with age. Reasoning about inconclusive and guess items is possibly more complex in the Unobservable condition which resulted in significant working memory demands. Inconclusive and guess information presents an increased challenge for children, thus considering these items when reasoning about unobservable properties seemingly increases the task complexity for the younger participants. Results indicate that the Unobservable condition may have been taxing to working memory particularly for the kindergarten children due to this increase in reasoning complexity. Thus, resulting in their difficulty differentiating the levels of evidence for the younger age groups.
The differences in memory demands may have contributed to distinct patterns across age groups in the two conditions. Similar to previous investigations the current study demonstrated improvement in differentiating levels of evidence with age. The Unobservable task when reasoning about more ambiguous items resulted in more effortful working memory demands. This presented difficulty for younger children however, by third grade, children were presumably better able to manage the increase in working memory demands compared to the Observable condition. This allowed for maintaining the necessary information for each trial and reporting accurate certainty rating responses. The simplicity of the task probably did not present working memory challenges for adult participants which resulted in the similar patterns of response across both conditions. Additionally, the type of reasoning involved in the procedure may have magnified demands on working memory for the younger participants. Investigations have suggested that working memory is particularly relevant during logical reasoning tasks (Handley et al., 2004). This would imply that evidence evaluation tasks may be more reliant on working memory abilities. Thus, younger children with less advanced working memory capacity may experience challenges. It would be valuable to further investigate the influence of working memory capacity when reasoning about observable vs. unobservable properties.

A distinction to consider is if participants reasoned about individual or category membership differently across the two conditions. For the Observable condition information about boxes is generally presented (e.g., blue boxes have keys inside) and participants are then asked to reason about a specific box (e.g., “It’s a blue box. Do you know if the blue box has keys or rocks inside it?”). For the Unobservable condition, a new animal is generally presented, however, a specific species is named (e.g., “Here is a new animal called a Moki”). Although
participants were asked to reason about a specific individual item in both conditions during evidence evaluation (e.g., a single blue box or food left in the nest of a single Moki), for the Unobservable condition younger children may have reasoned that the individual animal being discussed represented a species. Evidence suggests that children as young as 4-years-old, can successfully reason about category membership and make appropriate inferences when presented with novel or unfamiliar categories (Davidson & Gelman, 1990) and this extends to reasoning separate from perceptual information including unobservable properties (Gelman & Markman, 1986). Similarly, Cimpian (2016) suggested that the development of category representations of categories emerges early and argues that there exists a developmental preference for categorical representation. Specifically, developing cognitive processes are better structured to reason about category membership information. Nonetheless, the Unobservable condition may have created different reasoning demands which could have resulted in reasoning about a species compared to an individual item. Younger children’s certainty ratings may have been impacted by this distinction. For example, less certainty could be reported when asked to reason about the new animal if it was believed to be representative of larger species compared to reasoning about a singular item (e.g., blue box). Subsequent investigations could include more consistent language during the Unobservable task to emphasize reasoning about an individual animal which would form additional similarities across the two conditions. Further investigations considering reasoning about both individuals and categories and its influence on scientific reasoning would be beneficial.

The extent of experience with reasoning about unobservable properties may impact performance outcomes. Younger children may not be engaging in the pattern of reasoning
required for the Unobservable condition. Kindergarten and first grade learning materials likely incorporate reasoning about concrete and observable content in efforts to maximize comprehension. Thus, younger children may not fully understand that in a scientific situation it is possible to learn things through evidence and reasoning when information is not immediately available. By third grade children may have explicit training or experience reasoning about information that is not observable or verifiable and making appropriate inferences. This is possibly a result of more advanced learning materials and further development of logical reasoning processes. Thus, by third grade responses were more similar to adult response patterns in the present study.

Alternatively, the differences captured may be exhibited due to the response rating scale used in the current investigation. The current investigation used a modified 5-pont certainty rating scale (Pillow et al., 2015). This allowed for greater availability of response options. Scientific reasoning studies frequently use dichotomized forced choice response options such as, “Know for sure”/“Not know for sure” or “Is this a time you can tell/Is this a time you cannot tell” (Fay & Klahr, 1996; Klahr & Chen, 2003). The rating scale could have more accurately detected differences in evidence evaluations across age groups that previous scientific reasoning procedures did not permit.

To summarize, when comparing findings of previous studies to the present results it suggests a difference when reasoning about observable properties compared to unobservable. That is, reasoning about observable items may be more accessible in kindergarten (approximately 5 years old). However, reasoning about unobservable properties may not develop
until third grade (approximately 9 years old). Current findings suggest the scope or nature of observability vs. nonobservability impacts children’s reasoning.

In conclusion, the implications of study results will be discussed followed by recommendations for future investigations. Scientific reasoning investigations inform the understanding of cognitive and logical reasoning processes during development and has the prospect of informing science education. The current study allows for an increased understanding of children’s inferential activities, reasoning about observable and unobservable properties, understanding of differentiating levels of evidence, and their ability to reflect on their own reasoning to report certainty judgments. The present and related findings can be used to develop science education curriculum. A more accurate and in-depth knowledge of the scope of children’s scientific reasoning abilities can assist in advising the science content being administered at specific grades, the instruction being delivered, and the corresponding learning materials being applied. Specifically, it would be beneficial to help children understand during science instruction that relevant information goes beyond following protocols or procedures with known outcomes. Expressing the understanding that it is not always the case that science investigations involve deliberating about known information but that it often requires reasoning about unknown information where evidence is evaluated to draw appropriate conclusions. Current study results demonstrate that this line of scientific reasoning exploration can be included at approximately third grade. Children have the ability to start considering reasoning about unknown information and the procedures to take to address the gaps in knowledge.

Additionally, scientific reasoning abilities are a critical and required skill in current society (Fischer et al., 2014, Golumbic et al., 2023; Osborne, 2013). It allows the ability to
successfully navigate the significant amount of information that is transmitted and accessible. For example, understanding if the information of the evidence available is conclusive or inconclusive or the inferences that can be made from the presented evidence. These skills are necessary in both educational and non-academic contexts.

Modifications to the current study procedure could offer a more precise representation of scientific reasoning skills. The present investigation exclusively relied on a certainty rating scale to capture participant responses. It would be beneficial for subsequent investigations to include a justification prompt in addition to the certainty rating scale responses. This could provide supplementary information on the reasoning involved in children’s responses during the task. For example, a justification prompt included in the current procedure could provide information to clarify differences when reasoning during the Observable vs. Unobservable condition. This would be beneficial in efforts to identify differences between kindergarten and first grader children that were present in both conditions and differences between younger children (5- and 6-year-old) and third grade children (9-year-olds) detected in the Unobservable condition. Asking justification prompts to explain answers could increase understanding of scientific reasoning abilities across age groups.

The modality in which the task was administered must also be considered. The current study procedure was presented online. This allowed for recruitment due to increased restrictions at local daycare facilities and elementary schools. Prior investigations did not utilize online experimental procedures. Though the present study is using the same theoretical frameworks as previous investigations, the study design results in a comparison of different modalities of investigations. It could be informative to compare the online procedure to an equivalent in-
person procedure. The revised study design would generate questions on how reasoning about the different property types translates between the different procedural modalities. For example, would children’s certainty judgments of observable items be more certain when manipulating an object in-person or would children be more or less certain when reasoning about unobservable properties compared to the online procedure. Results could inform the methodology of subsequent investigations. As online procedures improve in addition to children’s familiarity with engaging and communicating on screen, there may be little differentiation. However, ensuing studies are necessary to determine how different modalities may affect reasoning performance or certainty judgments.

The patterns of performance of the present investigation suggest that: (a) evaluation of the certainty of inferences about observable properties differ from evaluation of the certainty of inferences about unobservable properties for young children, (b) performance evaluating varying levels of evidence is dependent on if the information involves observable or unobservable properties and (c) reasoning about inconclusive evidence with unobservable properties or unverifiable information may require advanced skills. The present study design allowed the opportunity to examine the differences in performance when reasoning about observable vs. unobservable directly. Since previous studies relied on concrete items that were readily available and verifiable, the present investigations extended the literature to assess children’s ability to reason about unobservable properties and unknown information that was not immediately verifiable.

To conclude, the current findings demonstrate that children can identify different levels of evidence when the information is observable and have difficulty when the information
involves unobservable properties. By kindergarten children can successfully differentiate levels of evidence during the Observable condition. Current results provide further support for early emergence of scientific reasoning abilities. The comparison of the Unobservable condition demonstrated improved performance with age, showing it was not until approximately third grade that students could successfully engage in such tasks. The current investigation fosters greater understanding of the extent of abilities for scientific reasoning across age groups. Current findings provide further knowledge of how young children’s evaluation of varying levels of evidence can be influenced by the extent of the observability or nonobservability of study materials. The results of the study are expected to be of interest to developmental and cognitive psychologists, as well as science educators. The present findings have the potential to inform instruction in science, logical reasoning processes across age groups, and further research concerning scientific reasoning development.
REFERENCES


APPENDIX A

EVIDENCE EVALUATION TASK
Evidence Evaluation Task

Warm-up (certainty scale):
Familiarize participants with the certainty rating scale.

We’re going to play a game. In this game, I am going to show you some pictures and I’m going to ask you some questions. I would like you to tell me if you are very sure, a little sure, or not sure about things. To tell me how sure you are, you can point to a number on the screen and then say the number out loud. *(Rating scale: 1-not sure at all, 2-a little sure, 3-medium, 4-kind of sure, 5-very sure).*

1. Here is a picture. Do you know what this is called? How sure are you that you know what this is called? Are you very sure, a little sure, or not sure?”
   - Picture of a familiar object (a photograph of an apple)

2. Here is a picture. Do you know what this is called? How sure are you that you know what this is called? Are you very sure, a little sure, or not sure?
   - Picture of an unfamiliar object (a photograph of a racing bicycle pedal)

Condition 1: Boxes

Warm-up (Boxes)
Familiarize participants with curtain opening during box task.

Here we have some boxes. They are different colors, dark green, light blue, pink, and brown. We can see the whole box here. Now I am going to hide one of these boxes behind this curtain.

Look, when we open the curtain, we can now see part of the box. Now you can see the color of the box.

Here we have some boxes. They are different colors, dark green, light blue, pink, and brown. Look, you can open a box and put things inside.

Experiment (Boxes)
We are going to play a game with some boxes. They are different colors, and some of them have things inside them that make a sound. Let’s see if we can find out what’s inside them.

Observable Conclusive:
Here are two blue boxes. The blue boxes have keys inside. When you shake a blue box, it makes a sound like keys. Like this…*(key sound effect)*.

- Knowledge check 1: What’s inside the blue box?
Here are two red boxes. The red boxes have rocks inside. When you shake a red box, it makes a sound like rocks. Like this…(*rocks sound effect*).

- Knowledge check 2: What’s inside the red box?

Now I am going to hide one of these boxes.

Evidence: Look, we can see the box now. It’s a [blue/ red] box. Do you know if the [blue/ red] has keys or rocks inside it?
- Certainty judgment: How sure are you? Are you very sure or not sure at all?

Observable Inconclusive:
Here are two orange boxes. One orange box has a rattle inside. When you shake this orange box, it makes a sound like a rattle. Like this…(*rattle sound effect*).
The other orange box has a bell inside. When you shake this orange box, it makes a sound like a bell. Like this…(*bell sound effect*).

- Knowledge check 1: (a) What’s inside this orange box? (b) What’s inside this orange box?

Here are two purple boxes. One purple box has a rattle inside. When you shake this purple box, it makes a sound like a rattle. Like this…(*rattle sound effect*).
The other purple box has a bell inside. When you shake this purple box, it makes a sound like a bell. Like this…(*bell sound effect*).

- Knowledge check 2: (a) What’s inside this purple box? (b) What’s inside this purple box?

Now I am going to hide one of these boxes.

Evidence: Look, it’s a [purple/orange] box. Do you know if the [purple/orange] has a bell or rattle inside it?
- Certainty judgment: How sure are you? Are you very sure or not sure at all?

Observable Guess:
Here we have some boxes. There are light purple, green, gray, and yellow boxes. Some of the boxes have pennies inside. When you shake these boxes, they make sounds like pennies. Like this…(*pennies sound effect*). Some boxes have marbles inside. When you shake these boxes, they make sounds like marbles. Like this…(*marbles sound effect*).

- Knowledge check 1: What is one sound that the boxes make?
• Knowledge check 2: Tell me another sound that the boxes make?

Now I am going to hide one of these boxes.

Evidence: Look, it’s a [light purple/green/gray/yellow] box. Do you know if the [[light purple/green/gray/yellow] has pennies or marbles inside it?

• Certainty judgment: How sure are you? Are you very sure or not sure at all?

**Condition 2: New Animal**

Warm-up (New animal)
(Each item will be shown in a different context, then shown in an animal nest (e.g., grapes in a tree then see grapes in a nest.)

Here we have some grapes. Look, we can now see the grapes in a nest.
Here we have some bugs. Look, we can now see the bugs in a nest.

Experiment – Introduction (New animal)
Some people just found a new island. We want to learn more about the new animals that live on the island. Animals that eat different foods are good at different things. We are going to look at the food in the animals’ nests after they eat to help us learn about what the animals are good at doing.

Unobservable Conclusive:
People have found that some animals eat carrots and they drop carrots in their nests after they eat. All animals on the island that eat carrots are good at hearing sounds from far away.

• Knowledge check 1: What are animals that eat carrots good at doing?

People have found that some animals eat leaves and they drop leaves behind in their nests after they eat. All animals on the island that eat leaves are good at seeing things in the nighttime when it is dark.

• Knowledge check 2: What are animals that eat leaves good at doing?

Here is a new animal called a Moki. No one has seen the Moki up close, but it always drops food behind in its nest.

• Evidence: Look, we found [carrots/leaves] in the Moki’s nest. Do you know if the Moki is good at hearing sounds from far away or good at seeing things in the nighttime?

• Certainty judgment: How sure are you? Are you very sure or not sure at all?
Unobservable Inconclusive:
People have found that some animals eat berries and drop berries behind in their nests. Some animals that eat berries are good at seeing very tiny things but other animals that eat berries are good at smelling things from far away.

- Knowledge check 1: (a) What is something animals that eat berries good at doing? (b) What's another thing that’s animals that eat berries good at doing?

People have found that some animals eat corn and drop corn behind in their nests. Some animals that eat corn are good at seeing very tiny things but other animals that eat corn are good at smelling things from far away.

- Knowledge check 2: (a) What is something animals that eat corn good at doing? (b) What’s another thing that animals that eat corn good at doing?

Here is a new animal called a Pema. No one has seen the Pema up close, but it always drops food behind in its nest.

- Evidence: Look, we see [berries/corn] in the Pema’s nest. Do you know if the Pema is good at seeing very tiny things or good at smelling things from far away?
- Certainty judgment: How sure are you? Are you very sure or not sure at all?

Unobservable Guess:
We found a new part of the island with many new animals! Some of the animals eat worms, nuts, oranges, or peas. The animals on the island are good at doing different things. Some animals are good at seeing things from far away and some animals are good at hearing very quiet sounds.

- Knowledge check 1: What are some of the new animals good at doing?
- Knowledge check 2: Tell me something else that some of the new animals are good at doing?

Here is a new animal called a Chiboo that eats [worms/nuts/oranges/peas]. No one has seen the Chiboo up close, but it always drops food behind in its nest.

- Evidence: Look, we see [worms/nuts/oranges/peas] in the Chiboo’s nest. Do you know if the Chiboo is good at seeing things from far away or good at hearing very quiet sounds?
- Certainty judgment: How sure are you? Are you very sure or not very sure?
APPENDIX B

CORRESPONDING FEATURES ACROSS TASKS
## Corresponding Features Across Tasks

<table>
<thead>
<tr>
<th>Task (observable)</th>
<th>Item: box</th>
<th>Feature: sound effect (e.g., bell)</th>
<th>Shows evidence: curtain</th>
<th>Evidence: color (e.g., blue box)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Box Task</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Animal Task (unobservable)</td>
<td>Item: new animal</td>
<td>Feature: skill (e.g., good at smelling things from far away)</td>
<td>Shows evidence: nest</td>
<td>Evidence: food (e.g., carrot)</td>
</tr>
</tbody>
</table>