Children Revise Their Core Beliefs About Objects, Agents, and Social Groups With Statistical Evidence

By

Rongzhi Liu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Fei Xu, Chair
Professor Jan Engelmann
Professor Mahesh Srinivasan
Professor Arianne Eason

Fall 2023

Children Revise Their Core Beliefs About Objects, Agents, and Social Groups With Statistical Evidence

Abstract

Children Revise Their Core Beliefs About Objects, Agents, and Social Groups With Statistical Evidence

by

Rongzhi Liu

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor Fei Xu, Chair

Humans are powerful Bayesian learners – we rationally update our beliefs given new statistical evidence. In human history, there is ample evidence that we can completely overturn our previous beliefs and theories given new environmental input; we can also forego past tools, technologies, and social arrangements to adapt to new environments. However, are there limits to our ability to rationally revise our beliefs? Are there beliefs that are so entrenched that cannot be revised? Previous research has shown that human adults, but not non-human animals (i.e., chickens) can override a hard-wired perceptual "light-from-above" prior with new evidence. This raises the possibility that the learning mechanisms in humans may be more powerful than those in non-human animals. The current dissertation investigates whether beliefs in the most fundamental domains of human knowledge – the core knowledge systems of objects, agents, and social beings – can be revised given a small amount of counterevidence.

Chapter 1 outlines the theoretical framework of this research and describes the three core knowledge systems that are investigated in Chapters 2–4. Chapter 2 assesses whether 4- to 6-year-old children and adults can revise three core knowledge principles in the object system – solidity (solid objects cannot pass through each other), continuity (objects traverse spatiotemporally connected paths), and contact (objects cannot interact at a distance). The findings show that children and adults can revise their beliefs about these object principles given a small number of violations of each principle. Chapter 3 uses the same methods as Chapter 2 to assess whether 4- to 6-year-old children and adults can revise three core knowledge principles in the agent system – goal (agents' actions are goal-directed), efficiency (agents take the most efficient means to achieve their goals), and sampling (when an agent chooses objects that are in the minority of a population, they prefer that type of object). The findings show that children and adults can also revise their beliefs about these agent principles given a small number of violations of each principle. Furthermore, the agent principles are more easily revised than the object principles, suggesting that the agent system is more flexible than the object system. Chapter 4 investigates whether a novel paradigm of presenting statistically representative counterevidence can change 5- to 6-year-olds' intergroup biases. The findings show that this paradigm successfully changes children's biases about minimally defined social groups, and provide preliminary evidence that this paradigm is also effective in changing children's racial

biases. Finally, Chapter 5 synthesizes the findings from Chapters 2–4, and discusses the implications of this research for theories of cognitive development and social development.

In conclusion, the current dissertation demonstrates that humans have powerful learning mechanisms and suggests that we might have the ability to revise *any* beliefs with new evidence. Future research will investigate how robust and long-lasting the belief revisions are, and whether this ability to revise even our most fundamental beliefs is unique to human learners.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am most grateful to my advisor, Fei Xu, who has generously provided so much guidance and support throughout my graduate career. I am inspired by every conversation and interaction we had over the past five and a half years. She has made me a better researcher in so many ways, and I could not imagine having a better advisor than her. I am also thankful to my committee members, Jan Engelmann, Mahesh Srinivasan, and Arianne Eason, for the inspiring and constructive questions and suggestions that improved my research tremendously. I would like to thank Gil Diesendruck, who has been an amazing collaborator and a terrific mentor to me. I would also like to thank my undergraduate advisor, Stella Christie, for opening the door to cognitive development for me.

I have had the privilege of working with a group of highly motivated and passionate undergraduate research assistants from UC Berkeley and other Universities: Amy Abad, Cindy Chen, Michelle Chen, Yifei Chen, Tiffany Chiang, Alexis Chin, XiLin Choi, Elizabeth Cholakyan, Christiane Helmer, Jennifer Herrera, Connie Lam, Jolie Lam, Nina Li, Keyu (Amber) Liu, Bailey Martinez, Christina Nguyen, Sabrina Ni, Jimin Oh, Bethany Ou, Vindhya Srikanth, Yuanzhi Qin, Lydia Qu, Cathy Wang, Shengqi (Iris) Zhong, Hang Zhou. None of the research in my dissertation would be possible without their hard work and dedication.

I would like to thank our lab managers – Min (Tina) Tang, Kaitlyn Tang, Gwyneth Heuser, Emma Roth, Harmonie Strohl – who have been incredibly helpful and have made my life so much easier. I want to thank my peers and friends who have provided so much support and inspiration for my research – Yuan Meng, Shaun O'Grady, Ruthe Foushee, Elena Luchkina, Roya Baharloo, Stephanie Alderete, Alyson Wong, Elena Leib, and Rebecca Zhu.

I am grateful for the financial support provided by the American Psychological Foundation Elizabeth Munsterberg Koppitz Graduate Student Fellowship, the National Science Foundation, and the UC Berkeley Adrian Hao Yin Ü Gateway Graduate Fellowship. I thank the Lawrence Hall of Science, the Bay Area Discovery Museum, Kensington and Madera Elementary schools for their support in my data collection. I thank all of the children, their parents, and adults who participated in my studies.

I thank my parents, who have provided me unconditional support over the years. I thank my amazing friends for enriching my life outside of research – Jinzi Shen, Zirui Zhou, Fuyi Zhao, Senmiao Sun, Yanyan Ren, Runlei Zhang, Mingyu Yuan, Mengsong Ouyang, Keming Zhang, Yang Lyu, Xingjian Wu, Menglin Sun – thank you for sharing the joy of hiking, skiing, surfing, traveling, board games, and food with me! And my partner, Zhuofan Li, who I know will always be there for me.

Chapter 1: Introduction

1.1. Background

Human children and adults are remarkable learners. Young children acquire their native languages within the first few years of life, in the absence of any formal instructions. Babies born into different societies learn radically different cultural practices and social norms. Adults have invented complex technologies such as smartphones, computers, and artificial intelligence; children and adults in modern societies learn to interact with these technologies rapidly. More impressively, we can overturn our previous beliefs, theories, and social norms, and develop completely new beliefs, theories, and norms. Einstein's theory of general relativity replaced Newtonian mechanics, changing what people believed about how the universe worked for hundreds of years. Children and adults who immigrate to foreign countries can adapt to completely different cultures and learn new languages and new social norms.

Indeed, many cognitive scientists and developmental psychologists have argued that one of the hallmarks of human learning is that we form beliefs and revise them given new evidence (Chater & Oaksford, 2008; Gopnik & Wellman, 2012; Piaget, 1954; Tenenbaum et al., 2011; Xu, 2019; Ullman & Tenenbaum, 2020). However, are there limits to humans' ability to rationally revise our beliefs? Are there beliefs that are so entrenched that cannot be revised?

An example of a deeply ingrained, innate prior in the perceptual domain is the "light-from-above" prior. Humans and nonhuman animals hold a strong assumption that the light source is always from above. This assumption is rational since for all species living on Earth, the primary light source – the Sun – is always overhead. Researchers have investigated whether this "light-from-above" prior is subject to revision in humans and nonhuman animals. Hershberger (1970) found that chickens continue to assume the "light-from-above" prior after being reared in a cage with light from below for several weeks. In contrast, Adams, Graf, and Ernst (2004) showed that human adults, when given 1.5 hours of haptic information that goes against the "light-from-above" prior, adapted quickly when asked about the light source, and they generalized their revised beliefs about the light source to a different task. Thus, nonhuman animals cannot revise this strong perceptual prior given weeks of counterevidence, whereas human adults can revise it with a relatively small amount of counterevidence. Compared to nonhuman animals, humans might have more powerful and flexible learning mechanisms that allow us to adapt to more diverse environments and learn completely different beliefs and principles.

If human learners can revise the hard-wired perceptual priors such as the "light-from-above" prior, can we also revise other deeply entrenched beliefs – for example, our beliefs about how objects move, how agents behave, and our deeply ingrained biases toward social groups? This is the focus of the current dissertation.

1.2. Rational Constructivism

One of the leading theories on learning and belief revision is Rational Constructivism (Fedyk & Xu, 2018; Gopnik & Wellman, 2012; Tenenbaum et al. 2011; Xu, 2019; Xu & Kushnir, 2012, 2013). According to this framework, human infants begin with a set of proto-conceptual primitives. These proto-conceptual primitives go beyond the sensorimotor primitives as argued by Piaget (1954), but they are also not fully conceptual (as argued in Carey, 2009; Spelke et al., 1992) – they are not part of a *language of thought* (Fodor, 1975). Object, number, agency, space, and causality are 5 candidate proto-conceptual primitives that have been identified. The final outcome of learning and development is a set of domain-specific intuitive theories (e.g., intuitive physics, intuitive psychology, intuitive biology). Intuitive theories are similar to scientific theories in the sense that they are causal, explanatory frameworks that support predictions and actions, and that they can undergo belief revision and conceptual change.

Importantly, infants, children, and adults possess a large toolbox of learning mechanisms that drive their learning, development, and conceptual change from the initial state to the final state. These learning mechanisms include language and symbol learning, Bayesian inductive learning which supports rational belief revision, and constructive thinking (e.g., analogy, explanation, and mental simulation) which supports conceptual change. Moreover, children are active learners; besides learning by processing new evidence in their environment, children also actively interact with their environment to generate new data that are conducive to their learning.

The "rational" in Rational Constructivism bears on the key mechanism that drives learning in infants, children, and adults – Bayesian inductive learning. The Bayesian framework uses the Bayes' Rule as the formal tool to capture how learners should update their beliefs given new evidence:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in H} P(d|h')P(h')}$$

The Bayes' Rule computes the posterior probability $P(h|d)$, which is the learner's degree of belief in hypothesis h given data d. The posterior probability is proportional to the product of the prior probability and the likelihood. The prior probability $P(h)$ is the learner's degree of belief in hypothesis h before observing the new data. The likelihood $P(d|h)$ is the probability that the learner would observe data d if hypothesis h were true. Thus, the learner's posterior belief in the hypothesis should rationally integrate the strength of their prior belief and the strength of the evidence.

1.2.1. Rational belief revision in children and adults

In the past few decades, researchers have investigated how infants, children, and adults learn, and in particular, whether they can use new evidence to update their beliefs. A large body of research now suggests that humans are Bayesian learners – we learn from new evidence, and rationally update our beliefs in various domains.

Observing new evidence triggers explanations, explorations, and belief revision. Outcomes inconsistent with children's prior beliefs trigger their explanatory reasoning in the physical domain (Legare et al., 2010) and the psychological domain (Legare et al., 2016). Those explanations lead to exploratory, hypothesis-testing behaviors that are conducive to learning (Legare, 2012). For example, one study showed that when children observed evidence that violated their theories about shadow size, they were more likely to perform informative experiments to test their beliefs (van Schijndel, et al., 2015). In a similar study, children who believed that objects balance at their geometric center (instead of their center of mass) were shown evidence that violated their initial theories. These children spent a longer time playing with the balance and blocks. While they played, if they found an auxiliary variable (i.e., a magnet was attached to the object), they explained away the disconfirming evidence; but if no auxiliary variable was found, they correctly revised their theories about balance (Bonawitz et al., 2012). In another study, children formed initial beliefs about the location of a reward, and then they were given new, social evidence inconsistent with their initial beliefs – reasons provided by a disagreeing social partner. This study found that children can revise their beliefs in light of new, social evidence, and their belief revision depended on the strength of their initial beliefs and the quality of the reasons provided by the social partner (Schleihauf et al., 2022). Finally, adults' misconceptions about specific domains (e.g., the variables that affect the pendulum period) can be revised when they are given evidence that contradicts their initial beliefs (Masnick et al., 2017).

Moreover, observing new evidence supports infants and children to acquire new variables and new concepts. They can acquire new variables that are important for reasoning about physical events (Baillargeon, 2008). For example, in covering events, infants do not attend to the variable height until 12 months of age. In one study (Wang & Baillargeon, 2008), 9-month-olds were shown pairs of new evidence where a short cover would only partially hide the object, and a tall cover would fully hide the object. After observing these events, 9-month-olds learned to attend to the variable height in covering events. Similarly, children acquire new concepts and beliefs about matter, weight, density, and friction as they accumulate evidence about objects (e.g., Siegler, 1996; Smith, Carey, & Wiser, 1985). In the domain of agents, children can acquire new Theory of Mind concepts after being shown belief-violating evidence. Three-year-olds who had not passed the False Belief (FB) tasks received 12 microgenetic sessions, where they learned about the actual outcomes (i.e., outcomes that contradicted their original predictions) of the stories in the FB tasks. After the sessions, children showed significant improvement in their performance in the FB tasks (Amsterlaw & Wellman, 2006). Children can also acquire an understanding of the subjectivity of preference given new evidence. Repacholi and Gopnik (1997) found that infants do not understand that others can have preferences different from their own until 18 months of age – when younger infants were asked to offer other agents some food, they offered food that they preferred themselves regardless of the preferences other agents have expressed. However, after observing two experimenters expressing different preferences repeatedly, even 14-month-olds can represent others' preferences that are different from theirs (Doan et al., 2015). In addition, when 16-month-olds were given strong evidence that an agent had a different preference – when the agent chose 6 boring toys from a jar containing 13% of boring toys and 87% of interesting toys – 16-month-olds inferred that the agent preferred the boring toys (a preference different from their own), and they were more likely to offer a boring toy to the agent (Ma & Xu, 2011).

Furthermore, children's and adults' learning and belief revisions appear to be rational – they are consistent with the Bayesian framework. Researchers have used Bayesian probabilistic models to capture how children revise their beliefs about spatial contiguity (Kushnir & Gopnik, 2007), how they make causal inferences given statistical evidence (Schulz et al., 2007), how they revise their higher-order beliefs (Kimura & Gopnik, 2019), as well as the developmental course of preference understanding (Lucas et al., 2014b). Similarly, adults' causal learning rationally integrates the strength of their prior knowledge about the probability of different forms of causal relationships and the strength of the evidence (Lucas & Griffiths, 2010; Griffiths et al., 2011; Lucas et al., 2014a). In complex scenes of object interactions, adults rationally update their beliefs about object properties given new observations and mental simulations (Hamrick et al., 2016; Allen et al., 2020). Adults can also reason about agents in complex scenes, and they update their beliefs about agents' mental states and the environment given new evidence (Baker et al., 2017, Jara-Ettinger et al., 2020; Shu et al., 2021).

Thus, past research has demonstrated that children and adults are powerful and rational learners in numerous domains. However, are there limits to what we can learn? When we are given new evidence, can we revise *any* beliefs, even the most fundamental beliefs we hold since infancy?

1.3. The core knowledge systems

In this dissertation, I investigate this question by exploring the malleability of children's and adults' beliefs in the most fundamental domains of knowledge – the core knowledge systems (Spelke, 1988, 2000, 2022; Spelke & Kinzler, 2007).

The core knowledge systems are a small number of systems of domain-specific knowledge, each accompanied by a set of principles. These core knowledge systems emerge early in development (Baillargeon, 2008; Spelke, 2022), are universal in humans across cultures (Barrett et al, 2013; Gordon, 2004), and are shared with some nonhuman animals (Hare et al., 2001; Regolin & Vallartigara, 1995). These systems support further learning throughout development, and they allow infants and children to construct intuitive theories such as intuitive physics and intuitive psychology (e.g., Carey, 1985; Gopnik & Meltzoff, 1997; Wellman & Gelman, 1992).

The leading theorist on the Core Knowledge view, Elizabeth Spelke, has argued that the core knowledge systems occupy the middle ground between perceptual systems and belief systems. They share some properties with perceptual systems (e.g., they are ancient and encapsulated) and share some properties with belief systems (e.g., they consist of interconnected, abstract concepts). Spelke argued that the core knowledge systems are likely to be innate, since abstract principles are hard to learn from experience and since some animals need these systems for survival from the beginning. Furthermore, she argued that these systems are encapsulated and unaffected by our beliefs and thoughts (Spelke, 2022). If these arguments are true, then beliefs in the core knowledge systems might be resistant to revision.

However, under the Rational Constructivist and the Bayesian frameworks, belief revision is always possible given the right kind of counterevidence, even when we have strong prior beliefs. Thus, the core knowledge systems are good candidates to probe the limits of humans' ability to rationally revise our beliefs.

In this dissertation, I focus on three of the core knowledge systems – the systems of objects, agents, and social beings.

1.3.1. Objects

The most well-studied core knowledge system guides how we represent and reason about objects. A set of spatiotemporal principles underlies object perception and guides how objects move and interact. These principles include solidity – objects cannot occupy the same space as other objects (Spelke et al., 1992), continuity – objects exist and move continuously in time and space (Aguiar & Baillargeon, 1999), contact – objects do not interact at a distance (Leslie & Keeble, 1987), and cohesion – objects move as connected and bounded wholes (Aguiar & Baillargeon, 1999). These principles emerge by 2.5 to 6 months of age in human infants.

1.3.2. Agents

Another well-studied core knowledge system guides how we represent and reason about agents and their intentional actions. Agents have intentions, and they can act on objects to cause changes in objects. A distinct set of principles underlies agents' actions – agents' intentional actions are directed to goals (Woodward, 1998), they choose efficient means to achieve their goals (Gergely & Csibra, 2003), and their preferences can be inferred based on violations of random sampling (Wellman et al., 2016). For ease of exposition, I will refer to these as the Goal principle, the Efficiency principle, and the Sampling principle from now on. These principles emerge by 6 to 12 months of age in human infants.

1.3.3. Social beings

The third, more recently identified system guides how we reason about people as social beings who interact with other individuals. People not only have intentions and act on objects (as described in the agent system), but they also share experiences with other individuals and connect to other individuals within a social network. We tend to categorize ourselves and others into social groups based on (even arbitrary) similarities, and we prefer members of our own groups over members of other groups. For example, infants prefer individuals who share similarities with themselves (Mahajan & Wynn, 2012), and they expect individuals who share similarities to show ingroup preferences toward each other (Bian et al., 2022). They preferentially look at or engage with individuals based on their race (Kelly et al., 2005; Bar-Haim et al., 2006), gender (Quinn et al., 2002), and the languages they speak (Kinzler et al., 2007). These tendencies emerge between 3 to 12 months of age in human infants. They constitute the core knowledge system of social beings and underlie our later-developing biases toward social groups (Spelke & Kinzler, 2007; but see Spelke, 2022, for a slightly different account of the system of social beings).

1.4. Précis

The current dissertation explores the limit of humans' ability to rationally revise our beliefs by investigating the malleability of children's and adults' beliefs in the most fundamental domains of knowledge – the core knowledge systems of objects, agents, and social beings.

Chapter 2 examines whether 4- to 6-year-olds and adults can revise their beliefs about 3 core knowledge principles about objects – Solidity, Continuity, and Contact. Past research has shown that observing violations of these principles promotes exploration and learning about the properties of the particular objects that violated the principles (Stahl & Feigenson, 2015; 2017; Perez & Feigenson, 2022). Chapter 2 demonstrates that when given a small number of violations of these principles, children and adults can revise their higher-level beliefs about the abstract principles governing object reasoning. About a third of the learners genuinely accepted the counterevidence and genuinely revised their beliefs about these principles. In addition, learners conservatively generalized their revised beliefs to new objects and new contexts in the same environment.

Chapter 3 examines whether 4- to 6-year-olds and adults can revise their beliefs about 3 core knowledge principles about agents – Efficiency, Goal, and Sampling. Past research has shown that violations of these principles allow children to update their beliefs about the agents who violated the principles (Colomer et al., 2020; Colomer & Woodward, 2023). Paralleling Chapter 2, Chapter 3 demonstrates that when given a small number of violations of the psychological principles, children and adults can revise their higher-level beliefs about the abstract principles governing agent reasoning. About half of the learners genuinely accepted the counterevidence and genuinely revised their beliefs about these principles. Furthermore, learners readily generalized their revised beliefs to new agents in the same environment.

A comparison of the results in Chapter 2 and Chapter 3 revealed some important domain differences. Children and adults have stronger prior beliefs for the object principles than the agent principles, and the object principles are harder to revise than the agent principles. These findings suggest that the agent system might be more flexible than the object system. Thus, Chapters 2 and 3 provide strong evidence that human learners can revise our deeply entrenched beliefs.

Chapter 4 focuses on another set of deeply ingrained beliefs in the core knowledge system of social beings – biases about social groups. Chapter 4 examines whether children can change their intergroup biases about minimal groups and racial groups given new evidence. Past studies showed that exposure to counterstereotypic exemplars had mixed effects on changing adults' and children's intergroup biases (Block et al., 2022; Gonzalez et al., 2017; Gonzalez et al., 2021; Lai et al., 2014). Using a novel paradigm, Chapter 4 demonstrates that a minimal intervention of showing children *statistically representative counterevidence* can change 5- to 6-year-olds' attitudes and beliefs about social groups, although given the small sample size, these results should be interpreted with caution.

Taken together, the current dissertation shows that the belief revision mechanisms in human learners are not subject to the same limitations as non-human animals. Given a small amount of counterevidence, children and adults can revise their deeply entrenched beliefs about objects and agents. Given minimal intervention, children can change their deeply ingrained biases about social groups. This research suggests that humans might have a unique ability to revise *any* beliefs given new evidence.

Chapter 2: Children and Adults Revise Core Knowledge Principles About Objects

2.1. Introduction

One of the core knowledge systems guides how we represent and reason about objects. The core principles in this system emerge by 2.5 to 6 months of age. These principles include solidity – objects cannot occupy the same space as other objects (Spelke et al., 1992), continuity – objects exist and move continuously in time and space (Aguiar & Baillargeon, 1999), cohesion – objects move as connected and bounded wholes (Aguiar & Baillargeon, 1999), and contact – objects do not interact at a distance (Leslie & Keeble, 1987). These principles support further learning in the physical domain (see Baillargeon, 2008 for a review). These principles also persist into adulthood. Adults' ability to track multiple, independently moving objects is disrupted when the objects violate the principle of continuity (Scholl & Pylyshyn, 1999) or cohesion (Scholl et al., 2001; vanMarle & Scholl, 2003). Are these most fundamental core principles about objects subject to revision once we acquire them? If children and adults are given enough evidence that violates these principles, will they rationally update their beliefs?

Past research has shown that violations of core knowledge principles about objects lead to enhanced attention, exploration, and learning. Studies using the violation of expectation paradigm (VOE) have demonstrated that infants are surprised and look longer at events that violate the solidity, continuity, cohesion, and contact principles than events that are consistent with these principles (Aguiar & Baillargeon, 1999; Leslie & Keeble, 1987; Spelke et al., 1992). Adults also reported being more surprised at apparent violations of the continuity and solidity principles compared to events that did not violate any core physical principles (Smith et al., 2020). Furthermore, surprise at violations of these principles promotes exploration and learning. Observing an object violate a core physical principle prompted infants to explore that object more (Stahl & Feigenson, 2015), with the goal of finding an explanation for the violation (Perez & Feigenson, 2022). Infants and children also showed enhanced learning for properties and novel words related to the objects that violated core physical principles (Stahl & Feigenson, 2015; 2017). Thus, observing violations of core physical principles provides an opportunity for learners to learn something special about the particular objects that violated the principles. Here we ask a different question about the role of counterevidence: Are the fundamental, abstract principles governing object reasoning, which are already present in infancy, revisable given multiple violations of the principles?

In the present chapter, we examine this question with 3 core principles that guide reasoning in the object system: the Contact principle – objects do not interact at a distance, the Continuity principle – objects exist and move continuously in time and space, and the Solidity principle – objects cannot occupy the same space as other objects. In 6 experiments, participants observed events that supported or violated these principles, or they did not receive any new evidence about these principles. Then, they were asked to make predictions about the outcomes of new events that varied in the extent to which they were different from the original events. For each new event, participants were asked to predict whether objects would behave in ways consistent with the principles or inconsistent with the principles.

We investigated this question with animated objects in virtual environments. A large body of past studies have investigated children and adults' physical reasoning in the virtual environment, and these studies showed that children and adults expect objects to interact in similar ways in the virtual environment as in the real world. For instance, infants expect animated objects projected on the screens to interact by making contact with each other (Leslie

& Keeble, 1987). Adults expect objects in virtual environments to behave in accordance with the continuity and solidity principles (Smith et al., 2020). Many other studies have used virtual environments to examine more complicated physical reasoning in children and adults. For instance, children and adults can access their real-world physical knowledge to control objects' speeds and moving trajectories in virtual environments (Daum & Krist, 2009; Huber et al., 2003). Adults can also infer the relative masses of objects (Hamrick et al., 2016), infer whether a stack of blocks would fall and in which direction (Battaglia et al., 2013), and learn to use objects in new ways to flexibly solve physical problems, e.g., how to use a catapult to make an object drop into a bucket (Allen et al., 2020) in virtual environments. Thus, in the present studies, we also use animated objects in virtual environments, and we expect our results to generalize to children's and adults' reasoning about objects in the real world.

We compare 3 hypotheses for whether children and adults would revise their beliefs about the core physical principles given counterevidence. Our first hypothesis is that the core knowledge principles about objects are so robust that upon encountering evidence violating these principles, the system responsible for object representation and reasoning will be disrupted. When we observe such violations, the object reasoning system would generate "error messages", and cannot compute the scenes involving the violations. As a result, children and adults would not be able to process the belief-violating evidence or use this evidence to revise their beliefs about the core physical principles. This hypothesis predicts that regardless of the evidence they observe, participants would predict outcomes consistent with the principles for new events. A few past studies provide support for this hypothesis. Adults can track multiple, independently moving objects even if they briefly disappear from the visual field, as long as accretion and deletion cues suggest the presence of an occluder. However, adults failed to track objects when the objects violated the continuity principle, e.g., when objects disappeared at an occluder's boundary (without deletion cues) and instantaneously reappeared at the other boundary (without accretion cues) (Scholl & Pylyshyn, 1999). Similarly, adults fail to track objects that violate the cohesion principle (Scholl et al., 2001; vanMarle & Scholl, 2003).

Our second hypothesis is that upon encountering evidence violating these principles, the object representation and reasoning system would continue to process the events. However, the core knowledge principles are relatively robust in the sense that children and adults would not readily accept the violations as counterevidence against the principles. Instead, they will try to come up with alternative interpretations to explain away the violations (e.g., the violation is a perceptual illusion), and not revise their beliefs about the principles. This hypothesis predicts that when children and adults are asked to make predictions about the same objects in the same contexts, they will predict outcomes inconsistent with the principles, since the alternative interpretations may still apply. However, when they are asked to make predictions about different objects in different contexts, the alternative interpretations no longer apply, and they will predict outcomes consistent with the principles. In addition, when asked to explain the belief-violating evidence, participants would be more likely to provide alternative interpretations of the evidence instead of stating that they accept the evidence. Past research suggests that people sometimes explain away counterevidence and refuse to revise their beliefs, especially when they have strong prior beliefs. For instance, stereotypes about social groups are often resistant to counterevidence. When people encounter counterstereotypic exemplars, they group those exemplars into a subtype and view them as exceptions to the group. Children (Hayes et al., 2003) and adults (Richards & Hewstone, 2001) fail to revise their stereotypes when they can subtype the counterstereotypic exemplars. Since children and adults have strong prior beliefs about the core physical principles, these beliefs might also be resistant to change.

Our third hypothesis is that children and adults will accept the evidence violating the principles, and use this evidence to revise their beliefs about the principles. This hypothesis predicts that participants who saw the belief-violating evidence would be more likely to predict outcomes that are inconsistent with the principles, compared to those who saw the belief-consistent evidence and those who did not receive new evidence. In addition, when asked to explain the belief-violating evidence, participants would state that they have accepted the violations of the principles in the belief-violating evidence. After children and adults accept the counterevidence and revise their beliefs about the principles, there are two possibilities for how far they generalize their revised beliefs about the principles. First, they might generalize their beliefs narrowly, only to the same objects in the same contexts. In this case, participants would only predict inconsistent outcomes for the same objects in the same contexts. Second, they might generalize their beliefs widely to a certain extent. In this case, participants might predict inconsistent outcomes for both the same objects and different objects, in various contexts. They might also be gradually less likely to generalize their revised beliefs as the objects and contexts become more and more different from the original objects and contexts in the counterevidence. This hypothesis is consistent with a large body of literature showing that children and adults rationally integrate new evidence with their prior beliefs and revise their beliefs (e.g., Griffiths et al., 2011; Kimura & Gopnik, 2019; Kushnir & Gopnik, 2007; Lucas & Griffiths, 2010; Lucas et al., 2014a). However, most of the past studies did not examine the extent to which children and adults generalize their revised beliefs. One study has shown that children can use new evidence to revise their higher-order beliefs, and generalize their revised beliefs to new objects (Kimura & Gopnik, 2019).

We tested adults in Experiments 1—3, and young children (4- to 6-year-olds) in Experiments 4—6. Presumably adults, by virtue of having lived longer, have observed much more evidence consistent with these principles than children. On the other hand, adults have also observed more evidence inconsistent with these principles in magic shows, science fiction, movies, etc. We will compare adults' and children's results to examine whether there are any age differences in the strength of their prior beliefs about the core physical principles, and the extent to which they are willing to revise their beliefs given counterevidence.

2.2. Experiments 1

2.2.1. Methods

Participants

Forty-seven adults (mean age = 30 years; range = 18 to 55; *SD* = 9.2; 25 females) participated on Prolific, an online platform for behavioral experiments. Participants provided written informed consent prior to the experiment, and they were paid $3.2 for a 20-minute experiment.

Design

Participants were randomly assigned to one of the two conditions, the Belief Consistent (BC) condition and the Belief Violation (BV) condition. They were tested on 3 principles, Contact, Continuity, and Solidity, in counterbalanced orders. For each principle, there were 4 familiarization trials and 4 test trials (2 easy test trials and 2 hard test trials; order counterbalanced). The familiarization trials in the BC condition displayed events that were consistent with the principle and those in the BV condition displayed events that violated the principle. In test trials, participants chose between the *Belief Consistent (BC) response* and the *Belief Violation* (*BV*) *response*. They never received feedback about whether their choices were correct or incorrect.

Stimuli and procedure

Participants were instructed to watch a set of events and make predictions about new events.

Continuity principle. In the familiarization trials, two orange screens appeared side by side, with a gap in between. An object disappeared behind one of the screens. Then, the screens were removed. The object was either at the location of the screen that the object disappeared behind (BC condition) or at the location of the other screen (BV condition) (Figure 2.1). The object was different in each trial.

In the easy test trials, a new object disappeared behind one of the orange screens. A blue triangle and a green triangle indicated the screen that the object disappeared behind (the *BC response*) and the other screen (the *BV response*) (Figure 2.1). Participants chose the location where they believed they would find the object. In the hard test trials, a red door and a yellow door appeared. A new object disappeared behind one of the doors (Figure 2.1). Participants chose the location where they believed they would find the object, either the door that the object disappeared behind (the *BC response*) or the other door (the *BV response*).

Solidity principle. In the familiarization trials, a dark grey wall appeared and rotated 180 degrees to show that there was no hole in the wall. A green screen was placed in front of the wall and occluded the lower half of the wall. An object moved behind the screen. Then, the screen was removed. The object was either on the side of the wall that it went behind (BC condition) or on the other side of the wall (BV condition) (Figure 2.1). A different object was used in each trial.

In the easy test trials, a new object moved behind the green screen. A purple heart and an orange heart indicated the side of the wall that the object went behind (the *BC response*) and the other side of the wall (the *BV response*) (Figure 2.1). Participants chose the location where they believed they would find the object. In the hard test trials, two doors (side by side, with no gap in between) were placed in front of the wall and occluded the lower half of the wall. A new object moved behind the doors (Figure 2.1). Participants chose the location where they believed they would find the object, either the side of the wall that the object went behind (the *BC response*) or the other side of the wall (the *BV response*).

Contact principle. In the familiarization trials, participants were shown a blue box that could play music. In each trial, an object was placed either on the toy (BC condition) or above the toy (BV condition), and immediately the toy lit up and played music for 5 seconds (Figure 2.1). A different object was used to activate the toy in each trial.

In the easy test trials, a new object was placed next to the blue box. Participants were told that the object could activate the toy. A red star and a yellow star indicated the location on the toy (the *BC response*) and the location above the toy (the *BV response*) (Figure 2.1). Participants chose the location where they would place the object to activate the toy. In the hard test trials, a new, brown box and a new object appeared (Figure 2.1). The participants were told that the brown box was another music toy, and the object could activate the toy. Again, participants chose the location where they would place the object to activate the toy.

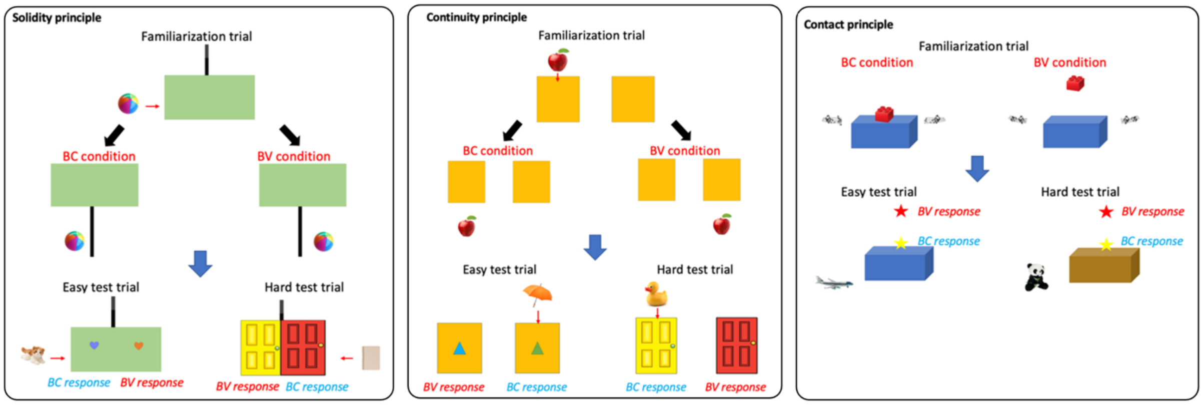

Figure 2. 1: Events shown in the familiarization trials and test trials for the Solidity principle, the Continuity principle, and the Contact principle in Experiment 1.

2.2.2. Results

The proportion of *BV response* by condition and principle is shown in Figure 2.2. We used mixed-effect logistic regression to predict participants' binary choice (*BV response* = 1, *BC response* = 0) from condition, principle, trial type, trial order, age, gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included condition as the only predictor. Participants were more likely to choose the *BV response* in the BV condition than in the BC condition ($\beta = 6.40$, $SE = 1.04$, $p < .001$).

Next, we compared participants' responses against chance. In the BC condition, participants chose the *BV response* below chance for all three principles (Exact binomial tests: *p*s < .001). In the BV condition, participants chose the *BV response* above chance for all three principles (*p*s < .001).

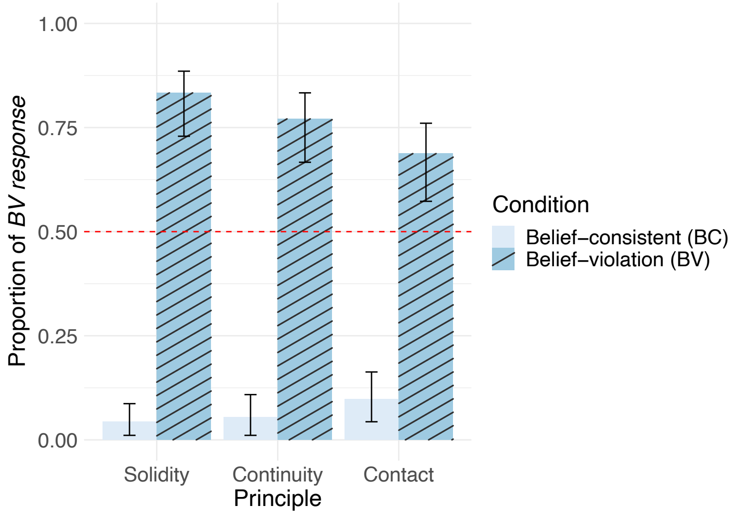

Figure 2. 2: The mean proportion of trials that adults selected the *belief-violation (BV) response* by condition and principle in Experiment 1. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

2.2.3. Discussion

Experiment 1 examined whether adults can revise their beliefs about the Contact, Continuity, and Solidity principles given counterevidence, in a specific, virtual environment. The results suggest that adults can revise their beliefs about all 3 principles, when they were given 4 pieces of counterevidence for each principle. After observing evidence consistent with these principles, adults predicted that the outcomes of new events would be consistent with the principles. However, after observing evidence violating these principles, they were more likely to predict outcomes inconsistent with the principles. Moreover, adults' performance did not differ in the easy and hard test trials, suggesting that they also generalized their revised beliefs to new situations.

In the next experiment, we aimed to replicate these findings with slightly modified stimuli. In addition, we measured adults' prior beliefs about these principles and tested the effect of belief-violating evidence on their prior beliefs. We also increased the strength of the belief-violating evidence and tested its effect on their beliefs. Lastly, we assessed participants' interpretations of the evidence to see if they accepted the counterevidence.

2.3. Experiments 2

2.3.1. Methods

Participants

Sixty adults (mean age = 33 years; range = 18 to 54; *SD* = 9.41; 35 females) participated on Prolific. The sample size was determined based on the effect sizes observed in Experiment 1. The sample size in this experiment provided us with at least 95% power (at $\alpha = .05$) to detect the effect sizes observed in Experiment 1. Participants provided written informed consent prior to the experiment, and they were paid $4 for a 25-minute experiment.

Design and procedure

The design procedure of Experiment 2 was similar to that of Experiment 1, with a few modifications. First, the events used in the Contact and the Solidity principles were slightly modified (see below for details). Second, we added a third, Baseline condition, where participants did not receive any new evidence that supported or violated the principles. Participants were randomly assigned to the Baseline condition, the BC condition, and the BV condition. Third, we increased the strength of the evidence; participants were shown 6 familiarization trials for each principle. Last, at the end of the study, participants in the BC and BV conditions were asked to explain one of the familiarization events for each principle. All adults were asked the explanation questions, but only 23 out of 36 participants were asked the explanation questions (since this measure was added to the procedure after piloting).

Continuity principle. The events used in the familiarization trials in the BC and BV conditions and the test trials in all 3 conditions were the same as in Experiment 1. In the familiarization trials of the Baseline condition, the screens were not removed after the object disappeared behind one of the screens (Figure 2.3). Thus, participants did not observe the location of the object.

For the explanation question, participants were asked to explain why the object appeared at the respective locations when the screens were removed.

Solidity principle. The events used for the Solidity principle were slightly modified so they were more similar to previous infant studies (e.g., Stahl & Feigenson, 2015). Two dark grey walls appeared and rotated 180 degrees to show that there was no hole in the walls. An object went down a ramp and moved behind a screen. In familiarization trials, when the screen was removed, the object was either stopped before the first wall (BC condition) or went past the first wall and appeared between the two walls (BV condition) (Figure 2.3). In the Baseline condition, the screen was not removed so that participants could not observe the location of the object. In the test trials, participants chose the location where they believed they would find a new object that went down the ramp. The location before the first wall was the *BC response* and the location between the two walls was the *BV response* (Figure 2.3).

For the explanation question, participants were asked to explain why the object appeared at the respective locations when the screen was removed.

Contact principle. Object launching events were used for the Contact principle, making these more similar to previous infant studies (e.g., Leslie & Keeble, 1987). In the familiarization trials, participants were told that a yellow car would launch various objects. In each trial, the yellow car moved toward an object and launched the object either by contacting it (BC condition) or at a distance (BV condition). In the Baseline condition, a screen blocked the view between the yellow car and the object so that participants could not see whether the yellow car contacted the other object or not (Figure 2.3). The yellow car launched a different object in each trial.

In the easy test trials, participants were told the yellow car could launch a new object. A red star and a yellow star indicated the location right next to the object (the *BC response*) and the location at a distance (the *BV response*) (Figure 2.3). Participants chose the location where the yellow car should stop to launch the new object. In the hard test trials, participants were told a new wheeled toy (e.g., the helicopter) could launch an object. Again, participants chose the location where the wheeled toy should stop to launch the object.

For the explanation question, participants were asked to explain why the yellow car launched the other object.

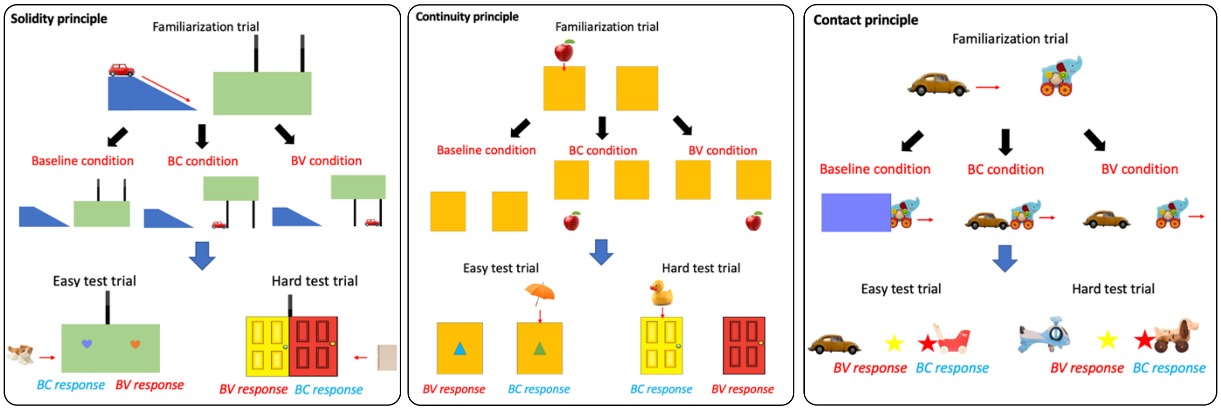

Figure 2. 3: Ev ents shown in the familiarization trials and test trials for the Solidity principle, the Continuity principle, and the Contact principle in Experiment 2.

2.3.2. Results

The proportion of *BV response* by condition and principle is shown in Figure 2.4. We used mixed-effect logistic regression to predict participants' binary choice (*BV response* = 1, *BC response* = 0) from condition, principle, trial type, trial order, age, gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included the interaction of condition and principle, the three-way interaction of condition, principle, and age, and the interaction of condition and gender as predictors. Most importantly, for all 3 principles, participants were more likely to choose the *BV response* in the BV condition than in the Baseline condition (Continuity: *β* = 9.31, *SE* = 2.52, *p* < .001; Solidity: *β* = 7.11, *SE* = 1.23, *p* < .001; Contact: *β* = 3.27, *SE* = 0.58, *p* < .001) and the BC condition (Continuity: *β* = 7.42, *SE* = 1.27, *p* < .001, Solidity: *β* = 10.28, *SE* = 1.67, *p* < .001; Contact: *β* = 4.23, *SE* = 076, *p* < .001); their choices did not differ between the Baseline and the BC conditions (Continuity: *β* = 2.14, *SE* = 2.59, *p* = .41, Solidity: *β* = -0.30, *SE* = 0.85, *p* = .72; Contact: *β* = -0.96, *SE* = 0.86, *p* = .27). In the Baseline and the BC conditions, participants chose the *BV response* below chance for all three principles (Exact binomial tests: *p*s < .04). In the BV condition, participants chose the *BV response* above chance for all three principles (*p*s < .001).

The interaction of condition and principle showed that, in the BV condition, participants were less likely to choose the *BV response* for the Contact principle compared to the Continuity principle (*β* = -2.53, *SE* = 0.57, *p* < .001) and the Solidity principle (*β* = -3.47, *SE* = 1.11, *p* < .001).

The three-way interaction of condition, principle, and age showed that, in the Baseline condition, participants were more likely to choose the *BV response* for the Contact principle with increasing age (*β* = 1.31, *SE* = 0.48, *p* = .006). No other age effect was found.
The interaction of condition and gender showed that, in the Baseline condition, males were more likely to choose the *BV response* compared to females (*β* = 1.23, *SE* = 0.54, *p* = .02). No other gender effect was found.

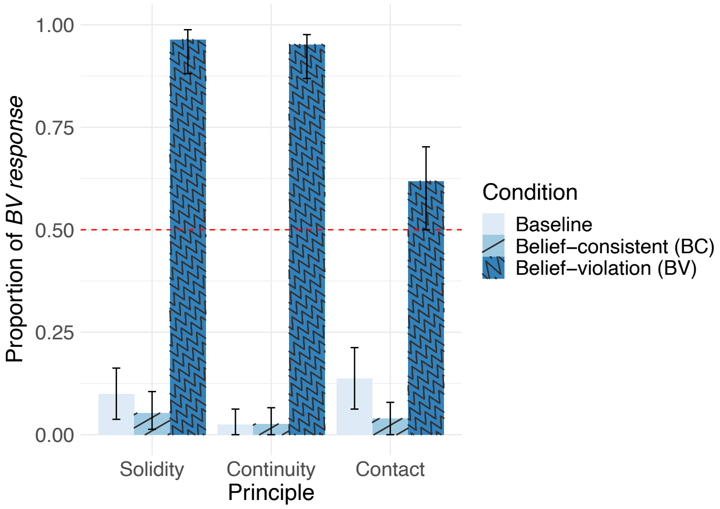

Figure 2. 4: The mean proportion of trials that adults selected the *belief-violation (BV) response* by condition and principle in Experiment 2. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

For the explanation questions, 2 researchers coded participants' responses into different categories (the interrater reliability was excellent, Cohen's Kappa = .94; disagreements were resolved through discussion). In the BC condition, most responses (98.2%) referred to the principle itself to explain the evidence (the only response not referring to the principle was incomprehensible). In the BV condition, we categorized participants' explanations into four categories. The criteria for categorization and examples are shown in Table 2.1 and Table 2.2.

Table 2.1: Coding criteria and examples for explanations

Category	Criterion
Accept Evidence	Accepted the violation of the target principle.
Explain Away	Explained the counterevidence with reasons that would not involve any violations of the target principle.
Pattern	Noted the pattern in the evidence.
Other	Explanations that cannot be categorized into the first three categories.

Table 2.2: Examples for each type of explanation in the belief-violation (BV) condition

	Adults	Children
Accept Evidence	Solidity: The ball went through the first wall. Continuity: The apple jumped to the right screen. Contact: The yellow car goes behind the purple car without touching it and it launches off.	Solidity: The ball went through the wall. Continuity: Because it went from the left side to the right side. Contact: Because it stopped there and moved the purple car.
Explain Away	Solidity: The first wall was farther behind the screen, leaving a gap for the ball to pass through. Continuity: There is an underground passage. Contact: The purple car got scared of the yellow car potentially hitting it.	Solidity: There is a hole in the first wall. Continuity: Maybe the screens are magic. Contact: Because the purple car didn't want the yellow car to bump into it.
Pattern	Solidity: Always appeared behind second wall from where it started. Continuity: That has been the pattern the whole time; for the object to show up at the opposite door. Contact: It has been the pattern the whole time for that to happen.	None.
Other	Solidity: The way it goes. Continuity: I am not sure. Contact: Physics?	Solidity: I don’t know. Continuity: I don’t know. Contact: Because there is a red light and the car can’t go.

Table 2.3 shows the number of responses coded within each category for each principle. We used mixed-effects multinomial logistic regression to predict participants' explanations from principle, while controlling for the random effects of individual participants. We found a significant effect of principle. Compared to the Contact principle, for the Continuity principle, participants were more likely to provide "explain away" explanations than "pattern" explanations (*p* = .04).

Table 2.3: Adults' Explanations by category and principle in Experiment 2

	Contact	Continuity	Solidity
Accept Evidence	6	7	7
Explain Away	9	5	6
Pattern	2	8	3
Other	4	1	5

Next, we used mixed-effects logistic regression to predict participants' binary choice in the test trials (BV response = 1, BC response = 0) from the type of explanation they provided, while controlling for the random effects of individual participants. We found that participants were more likely to choose the *BV response* for the principle if they provided "accept evidence" (β = 1.61, *SE* = 0.52, *p* = .002) or "pattern" (β = 2.06, *SE* = 0.68, *p* = .003) explanations for that principle, compared to if they provided "other" explanations. There was no difference between participants who provided "explain away" and "other" explanations (β = 0.66, *SE* = 0.44, *p* = .14). Table 2.6 shows the proportion of *BV responses* by participants who provided different types of explanations.

2.3.3. Discussion

In Experiment 2, we found that adults had strong prior beliefs about the Contact, Continuity, and Solidity principles. When they did not receive any new evidence about these principles and when they received evidence supporting these principles, they predicted that the outcomes of new events would be consistent with the principles. However, adults revised their beliefs about these principles given a few pieces of counterevidence, replicating the finding in Experiment 1. Similar to Experiment 1, adults' performance did not differ in the easy and hard test trials, suggesting that they were willing to generalize their revised beliefs to new situations.

We found an important difference across principles – given the same amount of counterevidence, participants were less likely to revise their beliefs about the Contact principle than the other two principles. We also discovered some interesting effects of age and gender on participants' prior beliefs about these principles. When given no new evidence, participants were more likely to predict outcomes inconsistent with the Contact principle with increasing age, suggesting that older adults might have weaker prior beliefs about the Contact principle. When given no new evidence, men were more likely to predict outcomes inconsistent with the principles than women, suggesting that men might have weaker prior beliefs about these principles than women. We will examine whether we can replicate these findings in the next experiment with a larger sample size.

Participants' explanations of the belief-violating evidence suggest that about a third of participants had accepted the counterevidence for each principle. Other participants simply learned from the statistical pattern of the evidence or explained away the counterevidence. However, participants who provided "explain away", "pattern" or "other" types of explanations still predicted outcomes that violated the principles in test trials most of the time, except for participants who provided "other" explanations for the Contact principle (Table 2.6). Participants were more likely to explain away the counterevidence for the Contact principle, and more likely to notice the statistical pattern of the counterevidence for the Continuity principle. More importantly, we found that participants who accepted the counterevidence or learned from the statistical pattern of the evidence were more likely to predict outcomes that violated the principles.

In the next experiment, we aim to replicate the findings of Experiments 1 and 2 with more photorealistic, three-dimensional stimuli made with Blender. Some participants explained away the counterevidence in Experiment 2 with reasons that involved perceptual ambiguity (e.g., for the solidity principle, some participants said the first wall was further towards the back, leaving a gap for the object to go through). The three-dimensional stimuli would curb these perceptual ambiguities. In Experiments 1 and 2, adults generalized their revised beliefs to slightly different contexts. In the next experiment, we further probe the extent to which adults are willing to generalize the revised beliefs by asking them to make predictions about events that are more different from the original events.

2.4. Experiments 3

2.4.1. Methods

Participants

One hundred and two adults (mean age = 20.52 years; range = 18 to 36; *SD* = 2.62; 81 females) participated on an online research platform at a university. The sample size was determined based on the effect sizes observed in Experiment 1. The sample size in this experiment provided us with at least 95% power (at $\alpha = .05$) to detect the effect sizes observed in Experiment 1. Participants provided written informed consent prior to the experiment, and they received 0.5 course credit for a 25-minute experiment.

Design and procedure

The design and procedure of Experiment 3 were similar to that of Experiment 2, with a few important modifications. First, we used photorealistic, three-dimensional stimuli made with Blender. Second, we added 2 harder test trials for each principle, where participants were asked to make predictions about completely different events (Figure 2.5). For the Continuity principle, participants saw an object move horizontally and disappear behind one of two screens; participants were asked to predict whether the object was behind the screen it went behind (*BC response*) or the other screen (*BV response*). For the Solidity principle, participants saw a vertical wall and two horizontal walls, which were then covered by a screen; an object disappeared behind the screen, and participants were asked to predict whether the object was above the first horizontal wall (*BC response*) or below the first horizontal wall (*BV response*). For the Contact principle, participants were told that a music toy can be activated by objects; they chose whether they would activate the toy by placing an object directly on top of the toy (*BC response*) or hovering above the toy (*BV response*).

Third, participants in the Baseline condition were also asked the explanation questions. Instead of explaining an event in the familiarization trial, they were asked to explain their own predictions in an easy test trial. For example, for the Continuity principle, participants were asked, “You predicted that the [object] is behind [participants’ response]. Why is that the case? Why is the [object] behind [participants’ response]?”

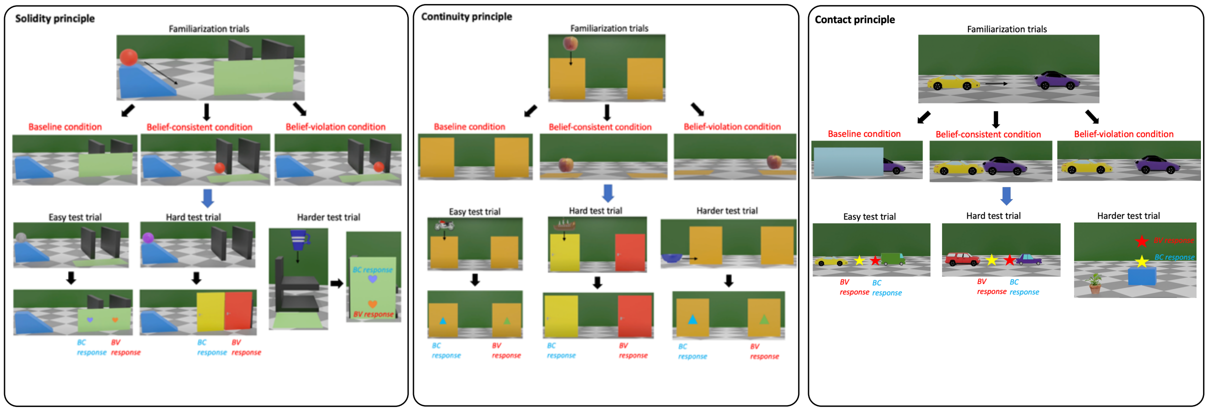

Figure 2. 5: Events shown in the familiarization trials and test trials for the Solidity principle, the Continuity principle, and the Contact principle in Experiment 3.

2.4.2. Results

The proportion of *BV response* by condition and principle is shown in Figure 2.6. We used mixed-effect logistic regression to predict participants’ binary choice (*BV response* = 1, *BC response* = 0) from condition, principle, trial type, trial order, age, gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included the interaction of condition and principle, the interaction of condition and test trial type, and the interaction of principle and age as predictors.

Most importantly, for all 3 principles, participants were more likely to choose the *BV response* in the BV condition than in the Baseline condition (Continuity: $\beta = 9.02$, $SE = 1.08$, $p < .001$; Solidity: $\beta = 6.61$, $SE = 0.84$, $p < .001$; Contact: $\beta = 5.17$, $SE = 0.81$, $p < .001$) and the BC condition (Continuity: $\beta = 8.52$, $SE = 0.93$, $p < .001$, Solidity: $\beta = 7.70$, $SE = 0.88$, $p < .001$; Contact: $\beta = 6.21$, $SE = 0.83$, $p < .001$); their choices did not differ between the Baseline and the BC conditions (Continuity: $\beta = 0.45$, $SE = 1.09$, $p = .68$; Solidity: $\beta = -1.06$, $SE = 0.83$, $p = .20$; Contact: $\beta = -1.12$, $SE = 0.80$, $p = .16$). In the Baseline and the BC conditions, participants chose the *BV response* below chance for all three principles (Exact binomial tests: *p*s < .001). In the BV condition, participants chose the *BV response* above chance for all three principles (*p*s < .001).

The interaction of condition and principle showed that, in the Baseline condition, participants were more likely to choose the *BV response* for the Contact principle ($\beta = 2.57$, $SE = 0.79$, $p = .001$) and the Solidity principle ($\beta = 2.18$, $SE = 0.79$, $p = .006$) compared to the Continuity principle. In the BC condition, participants were more likely to choose the *BV response* for the Contact principle than for the Continuity principle ($\beta = 0.98$, $SE = 0.47$, $p = .04$). In the BV condition, participants were less likely to choose the *BV response* for the

Contact principle compared to the Continuity principle (β = -1.26, *SE* = 0.31, $p < .001$) and the Solidity principle (β = -1.03, *SE* = 0.30, $p < .001$).

The interaction of condition and test trial type showed that, in the Baseline condition, (β = -1.19, *SE* = 0.52, $p = .02$). In the BV condition, participants were less likely to choose the *BV response* in harder test trials than in the easy test trials (β = -1.45, *SE* = 0.31, $p < .001$). But participants still chose the *BV response* in harder test trials above chance ($p < .001$).

The interaction of principle and age showed that, participants were more likely to choose the *BV response* for the Contact principle with increasing age (β = 0.68, *SE* = 0.27, $p = .01$).

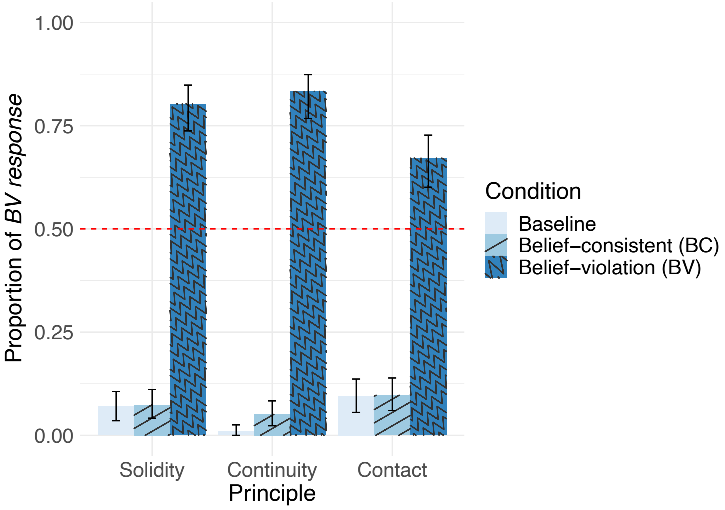

Figure 2. 6: The mean proportion of trials that adults selected the *belief-violation (BV) response* by condition and principle in Experiment 3. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

For the explanation questions, 3 researchers coded participants' responses into different categories (the interrater reliability was very good, Light's Kappa = .86; disagreements were resolved through discussion). In the Baseline and the BC condition, most responses (98.9% in the Baseline condition and 96.3% in the BC condition) referred to the principle itself to explain the evidence (the other responses were irrelevant to the principle or were incomprehensible). In the BV condition, we categorized participants' explanations into four categories. The criteria for categorization and examples are shown in Table 2.1 and Table 2.2.

Table 2.4 shows the number of responses coded within each category for each principle. Mixed-effects multinomial logistic regression did not reveal a significant effect of principle on participants' explanations, suggesting that the distribution of the types of explanations participants provided did not differ between principles. Mixed-effects logistic regression found a significant effect of the type of explanations on participants' choice in the test trials: Participants were more likely to choose the *BV response* for the principle if they provided "accept evidence" compared to if they provided "explain away" (β = 0.70, *SE* = 0.35, $p = .048$) or "other" (β = 1.03, *SE* = 0.42, $p = .014$) explanations for that principle. Table 2.6 shows the proportion of *BV responses* by participants who provided different types of explanations.

Table 2.4: Adults' Explanations by category and principle in Experiment 3

	Contact	Continuity	Solidity
Accept Evidence	5	9	9
Explain Away	20	12	18
Pattern	2	3	2
Other	6	9	4

2.4.3. Discussion

Experiment 3 replicated the main findings of Experiments 1 and 2 in a photorealistic, three-dimensional environment. Adults had strong prior beliefs about the Contact, Continuity, and Solidity principles, and they revised their beliefs given counterevidence. In addition, participants' performance differed across test trial types. Both when participants did not receive any new evidence and after they received evidence violating the principles, they were less likely to predict outcomes inconsistent with the principles in harder test trials than in the easy test trials. This suggests that adults had stronger prior beliefs for the 3 principles in the context of the events used in harder test trials, and they were less likely to generalize their revised beliefs to completely different contexts.

We also discovered some important differences across principles. Participants were more likely to predict outcomes inconsistent with the principles for the Contact principle than for the Continuity principle, both when they did not receive any new evidence and after receiving evidence consistent with the principles. Participants were also more likely to predict outcomes inconsistent with the principles for the Solidity principle than for the Continuity principle, when they did not receive any new evidence. Thus, adults had weaker prior beliefs for the Contact and Solidity principles than for the Continuity principle. After receiving evidence violating the principles, participants were least likely to predict outcomes inconsistent with the principles for the Contact principle. This replicates the finding in Experiment 2, and suggests that the Contact principle might be the hardest to revise. We also found an interesting age effect. Across conditions, participants were more likely to predict outcomes inconsistent with the principles for the Contact principle with increasing age. Similar to the finding in Experiment 2, this suggests that older adults might have weaker prior beliefs about the Contact principle than younger adults.

Participants' explanations of the belief-violating evidence suggest that only a small portion of participants (23%) had accepted the counterevidence. Most participants explained away the counterevidence. It is possible that the photorealistic stimuli made participants more skeptical about the counterevidence and they came up with alternative interpretations to explain away the counterevidence. However, participants who provided "explain away", "pattern", or "other" explanations still predicted outcomes inconsistent with the principles in test trials most of the time, except in harder test trials for the Contact principle for participants who provided "explain away" and "other" explanations (Table 2.6). We again found that participants who accepted the counterevidence were indeed more likely to predict outcomes that violated the principles, replicating the finding in Experiment 2.

2.5. Experiments 1—3: Combined results and discussion

Next, we analyzed the combined results of Experiments 1—3. We used mixed-effect logistic regression to predict participants' binary choice (*BV response* = 1, *BC response* = 0) from condition, principle, test trial type, experiment, age, gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included the interaction of condition and principle, and the interaction of condition and test trial type as predictors.

Most importantly, for all 3 principles, participants were more likely to choose the *BV response* in the BV condition than in the Baseline condition (Continuity: $\beta = 8.91$, $SE = 0.79$, $p < .001$; Solidity: $\beta = 7.00$, $SE = 0.66$, $p < .001$; Contact: $\beta = 4.98$, $SE = 0.63$, $p < .001$) and the BC condition (Continuity: $\beta = 8.57$, $SE = 0.67$, $p < .001$; Solidity: $\beta = 8.21$, $SE = 0.65$, $p < .001$; Contact: $\beta = 6.20$, $SE = 0.62$, $p < .001$). For the Continuity and the Solidity principles, participants' responses did not differ between the BC condition and the Baseline condition (Continuity: $\beta = 0.42$, $SE = 0.79$, $p = .59$; Solidity: $\beta = -1.20$, $SE = 0.64$, $p = .062$). For the Contact principle, they were less likely to choose the *BV response* in the BC condition than in the Baseline condition ($\beta = -1.25$, $SE = 0.63$, $p = .046$). In the Baseline and the BC conditions, participants chose the *BV response* below chance for all three principles (Exact binomial tests: *p*s < .001). In the BV condition, participants chose the *BV response* above chance for all three principles (*p*s < .001).

The interaction between condition and principle showed that, in the Baseline condition, compared to the Continuity principle, participants were more likely to choose the *BV response* for the Solidity ($\beta = 1.98$, $SE = 0.57$, $p < .001$) and the Contact principle ($\beta = 2.40$, $SE = 0.56$, $p < .001$). In the BC condition, participants were more likely to choose the *BV response* for the Contact principle than for the Continuity principle ($\beta = 0.87$, $SE = 0.35$, $p = .012$). In the BV condition, participants were more likely to choose the *BV response* for the Continuity principle ($\beta = 1.59$, $SE = 0.23$, $p < .001$) and the Solidity principle ($\beta = 1.63$, $SE = 0.23$, $p < .001$) compared to the Contact principle.

The interaction between condition and test trial type showed that, while participants' performance did not differ across test trial types in the Baseline condition, in the BC condition, they were more likely to choose the *BV response* in the harder test trials than in the easy test trials ($\beta = 0.88$, $SE = 0.40$, $p = .028$). In the BV condition, adults were less likely to choose the *BV response* in the hard test trials than in the easy test trials ($\beta = -0.44$, $SE = 0.21$, $p = .038$), and less likely to choose the *BV response* in the harder test trials than in the easy test trials ($\beta = -1.47$, $SE = 0.28$, $p < .001$) and the hard test trials ($\beta = -1.03$, $SE = 0.27$, $p < .001$). Importantly, participants still chose the *BV response* above chance in all 3 types of trials in the BV condition (*p*s < .001).

Overall, there was no significant effect of experiment, suggesting that the different stimuli sets did not affect participants' choices in the test trials. To examine the effect of the amount of evidence on participants' choices, we also analyzed the BC condition data across the 3 experiments and the BV condition data across the 3 experiments, respectively. The amount of evidence did not affect adults' *BV response* in the BC condition or the BV condition.

We next analyzed the combined explanation data of Experiments 2 and 3. In the BV condition, participants were more likely to choose the *BV response* for the principle if they provided "accept evidence" or "pattern" explanations compared to if they provided "explain away" or "other" explanations for that principle ("accept evidence" vs. "explain away": $\beta = 0.74$, $SE = 0.32$, $p = .02$; "accept evidence" vs. "other": $\beta = 1.00$, $SE = 0.36$, $p = .005$; "pattern" vs.

"explain away": *β* = 1.50, *SE* = 0.61, *p* = .02; "accept evidence" vs. "other": *β* = 1.75, *SE* = 0.65, *p* = .007).

In conclusion, across three experiments, we found consistent evidence that adults can revise their beliefs about all three principles given counterevidence. Adults might have weaker prior beliefs for the Contact and Solidity principles than the Continuity principle, and the Contact principle was the hardest to revise. I will return to these differences across principles in the general discussion. Moreover, participants who accepted the counterevidence or learned from the statistical pattern of the evidence were more likely to revise their beliefs. Furthermore, adults were gradually less likely to generalize their revised beliefs to objects and contexts that are more and more different from the original objects and contexts in the counterevidence.

2.6. Experiments 4

2.6.1. Methods

Participants

Twenty-four 4- to 6-year-olds (mean age = 5.04; range = 4.08 to 6.92; SD = 0.82; 11 females) participated over Zoom. Children's parents provided written informed consent prior to the experiment, and they received electronic certificates.

Design and Procedure

The design and procedure of Experiment 4 were the same as Experiment 1. An experimenter showed the stimuli videos to children via Zoom, and recorded their verbal responses in the test trials.

2.6.2. Results

The proportion of *BV response* by condition and principle is shown in Figure 2.7. We used mixed-effect logistic regression to predict participants' binary choice (*BV response* = 1, *BC response* = 0) from condition, principle, trial type, trial order, age, gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included the interaction of condition and principle as predictors. Most importantly, for the Continuity and Solidity principles, children were more likely to choose the *BV response* in the BV condition than in the BC condition (Continuity: *β* = 3.99, *SE* = 0.72, *p* < .001; Solidity: *β* = 2.09, *SE* = 0.59, *p* < .001). For the Contact principle, children were marginally more likely to choose the *BV response* in the BV condition than in the BC condition (*β* = 1.09, *SE* = 0.57, *p* = .056).

The interaction of condition and principle showed that, in the BC condition, compared to the Continuity principle, children were more likely to choose the *BV response* for the Solidity principle (*β* = 1.25, *SE* = 0.59, *p* = .033) and the Contact principle (*β* = 1.25, *SE* = 0.59, *p* = .033). In the BV condition, children are less likely to select the *BV response* for the Contact principle than for the Continuity principle (*β* = -1.64, *SE* = 0.51, *p* = .001) and the Solidity principle (*β* = -1.00, *SE* = 0.47, *p* = .033).

Next, we compared participants' responses against chance. In the BC condition, participants chose the *BV response* below chance for all three principles (Exact binomial tests: *p*s < .003). In the BV condition, participants chose the *BV response* above chance for the Solidity (*p* = .008) and the Continuity principle (*p* < .001), and chose the *BV response* at chance for the Contact principle (*p* = 1).

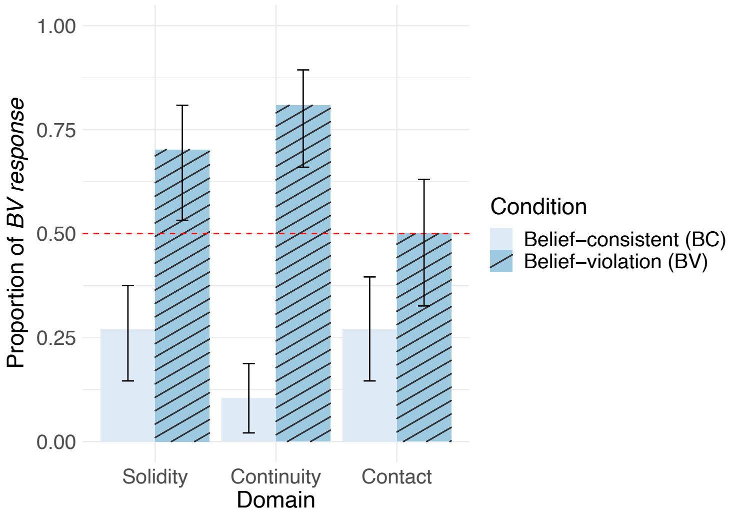

Figure 2. 7: The mean proportion of trials that children selected the *belief-violation (BV) response* by condition and principle in Experiment 4. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

2.6.3. Discussion

The results of Experiment 4 showed that children can revise their beliefs about the Solidity and the Continuity principles given counterevidence, but they did not reliably revise their beliefs about the Contact principle. After observing evidence consistent with these principles, children predicted that the outcomes of new events would be consistent with the principles. After observing evidence violating the Solidity and the Continuity principles, they were more likely to predict outcomes inconsistent with these two principles. Children did not reliably revise their beliefs about the Contact principle, possibly because they needed more counterevidence about this principle. Children's performance did not differ in the easy and hard test trials, suggesting that they also generalized their revised beliefs to new situations.

In the next experiment, we aimed to replicate these findings with slightly modified stimuli. In addition, we measured children's prior beliefs about these principles and tested the effect of belief-violating evidence on their prior beliefs. We also increased the strength of the belief-violating evidence and tested its effect on their beliefs. Lastly, we assessed participants' interpretations of the evidence to see if they accepted the counterevidence.

2.7. Experiments 5

2.7.1. Methods

Participants

Thirty-six 4- to 6-year-olds (mean age = 4.85; range = 4 to 6.83; SD = 0.80; 18 females, 15 males, and 3 of unknown gender) participated over Zoom. The sample size was determined based on the effect sizes observed in Experiment 4. The sample size in this experiment provided us with at least 85% power (at $\alpha = .05$) to detect the effect sizes observed in Experiment 4. Children's parents provided written informed consent prior to the experiment, and they received electronic certificates.

Design and Procedure

The design and procedure of Experiment 5 were the same as Experiment 2. Thirteen out of the 36 participants were not asked the explanation questions (since this measure was added to the procedure after piloting). An experimenter showed the stimuli videos to children via Zoom, and recorded their verbal responses in the test trials.

2.7.2. Results

The proportion of *BV response* by condition and principle is shown in Figure 2.8. We used mixed-effect logistic regression to predict participants' binary choice (*BV response* = 1, *BC response* = 0) from condition, principle, trial type, trial order, age, gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included the interaction between condition and principle as predictors. Most importantly, for the Continuity and Contact principles, participants were more likely to choose the *BV response* in the BV condition than in the Baseline condition (Continuity: $\beta = 4.51$, $SE = 0.68$, $p < .001$; Contact: $\beta = 2.13$, $SE = 0.84$, $p = .011$) and the BC condition (Continuity: $\beta = 5.83$, $SE = 1.25$, $p < .001$; Contact: $\beta = 3.91$, $SE = 1.01$, $p < .001$); their choices did not differ between the Baseline and the BC conditions (Continuity: $\beta = 0.12$, $SE = 1.52$, $p = .94$; Contact: $\beta = -1.77$, $SE = 0.96$, $p = .064$). For the Solidity principle, children were less likely to choose the *BV response* in the BC condition than in the Baseline condition ($\beta = -5.61$, $SE = 1.41$, $p < .001$), they were more likely to choose the *BV response* in the BV condition than in the BC condition ($\beta = 6.94$, $SE = 1.47$, $p < .001$), and their choices did not differ between the Baseline and the BV conditions ($\beta = 1.31$, $SE = 0.84$, $p = .12$). In the Baseline condition, participants chose the *BV response* below chance for the Continuity (Exact binomial test: $p < .001$) and the Contact principles ($p = .01$), and chose the *BV response* at chance for the Solidity principle ($p = .31$). In the BC condition participants chose the *BV response* below chance for all three principles (*p*s < .001). In the BV condition, participants chose the *BV response* above chance for all three principles (*p*s < .03).

The interaction of condition and principle showed that, in the Baseline condition, children were more likely to choose the *BV response* for the Contact principle than for the Continuity principle ($\beta = 3.39$, $SE = 1.09$, $p = .002$); they were more likely to choose the *BV response* for the Solidity principle than the Continuity principle ($\beta = 4.91$, $SE = 1.11$, $p < .001$) and the Contact principle ($\beta = 1.51$, $SE = 0.51$, $p = .003$). Their responses did not differ across principles in the BC and the BV conditions.

Since only a small number of children were asked the explanation questions in Experiment 5, we analyzed the combined explanation data from Experiments 5 and 6 and reported the results in the results section of Experiment 6 below.

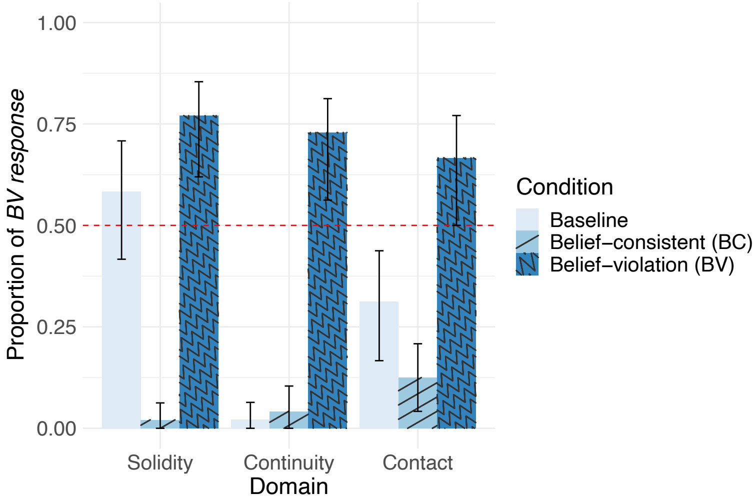

Figure 2. 8: The mean proportion of trials that children selected the *belief-violation (BV) response* by condition and principle in Experiment 5. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

2.7.3. Discussion

In Experiment 5, we found that children had strong prior beliefs about the Continuity and the Contact principles. When they did not receive any new evidence about these principles and when they received evidence supporting these principles, they predicted that the outcomes of new events would be consistent with the principles. In contrast, children were equally likely to predict outcomes consistent and inconsistent with the Solidity principle when they were not given any new evidence. This might suggest that children have weaker prior beliefs about the solidity principle. However, it might also be because of perceptual ambiguities in the stimuli (e.g., children might think the first wall was further toward the back, leaving a gap for the object to pass through). After receiving new evidence supporting the Solidity principle, children predicted that the outcomes of new events would be consistent with the principle. More importantly, children revised their beliefs about all 3 principles given counterevidence – they predicted that the outcomes of new events would be inconsistent with the principles after receiving new evidence violating the principles. Similar to Experiment 4, children's performance did not differ in the easy and hard test trials, suggesting that they were willing to generalize their revised beliefs to new situations.

We also found that children have weaker prior beliefs for the Solidity principle than the other 2 principles, possibly because of perceptual ambiguities in the stimuli. In addition, like adults, children also have weaker prior beliefs about the Contact principle than the Continuity principle.

In the next experiment, we aim to replicate the findings of Experiments 4 and 5 with more photorealistic, three-dimensional stimuli made with Blender. The three-dimensional stimuli would rule out the perceptual ambiguities in the stimuli for the Solidity principle. We also further probe the extent to which children are willing to generalize the revised beliefs.

2.8. Experiments 6

2.8.1. Methods

Participants

Thirty-six 4- to 6-year-olds (mean age = 5.41; range = 4 to 6.92; SD = 0.88; 14 females and 22 males) participated over Zoom. The sample size was determined based on the effect sizes observed in Experiment 4. The sample size in this experiment provided us with at least 85% power (at $\alpha = .05$) to detect the effect sizes observed in Experiment 4. Children's parents provided written informed consent prior to the experiment, and they received electronic certificates.

Design and Procedure

The design and procedure of Experiment 6 were the same as Experiment 3. An experimenter showed the stimuli videos to children via Zoom, and recorded their verbal responses in the test trials.

2.8.2. Results

The proportion of *BV response* by condition and principle is shown in Figure 2.9. We used mixed-effect logistic regression to predict participants' binary choice (*BV response* = 1, *BC response* = 0) from condition, principle, trial type, trial order, age, gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included the interaction between condition and principle as predictors. For all 3 principles, participants were more likely to choose the *BV response* in the BV condition than in the Baseline condition (Continuity: $\beta = 4.51$, $SE = 0.68$, $p < .001$; Solidity: $\beta = 1.46$, $SE = 0.47$, $p = .002$; Contact: $\beta = 1.74$, $SE = 0.49$, $p < .001$) and the BC condition (Continuity: $\beta = 4.84$, $SE = 0.74$, $p < .001$, Solidity: $\beta = 2.39$, $SE = 0.51$, $p < .001$; Contact: $\beta = 1.49$, $SE = 0.48$, $p = .002$); their choices did not differ between the Baseline and the BC conditions (Continuity: $\beta = -0.32$, $SE = 0.85$, $p = .70$, Solidity: $\beta = -0.93$, $SE = 0.50$, $p = .06$; Contact: $\beta = 0.25$, $SE = 0.51$, $p = .62$). In the Baseline and the BC conditions, participants chose the *BV response* below chance for all three principles (Exact binomial tests: *p*s < .05). In the BV condition, participants chose the *BV response* above chance for the Solidity ($p = .001$) and the Continuity principle ($p < .001$), and chose the *BV response* at chance for the Contact principle ($p = .41$).

The interaction of condition and principle showed that, in the Baseline and the BC conditions, compared to the Continuity principle, participants were more likely to choose the *BV response* for the Contact principle (Baseline: $\beta = 1.59$, $SE = 0.60$, $p = .008$; BC: $\beta = 2.17$, $SE = 0.66$, $p < .001$) and the Solidity principle (Baseline: $\beta = 2.51$, $SE = 0.59$, $p < .001$; BC: $\beta = 1.91$, $SE = 0.66$, $p = .004$). In the BV condition, participants were less likely to choose the *BV response* for the Contact principle compared to the Continuity principle ($\beta = -1.18$, $SE = 0.38$, $p = .002$).

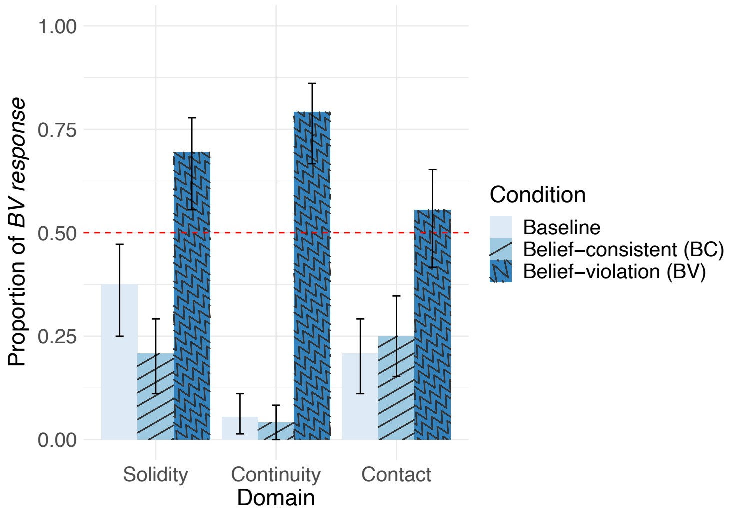

Figure 2. 9: The mean proportion of trials that children selected the *belief-violation (BV) response* by condition and principle in Experiment 6. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

For the combined children's explanation data in Experiments 5 and 6, 2 researchers coded participants' responses into different categories (the interrater reliability was excellent, Cohen's Kappa = .93; disagreements were resolved through discussion). In the Baseline and the BC condition, most responses (58.3% in the Baseline condition and 69.8% in the BC condition) referred to the principle itself to explain the evidence (the other responses were irrelevant to the principle or were incomprehensible). In the BV condition, we categorized participants' explanations into four categories. The criteria for categorization and examples are shown in Table 2.1 and Table 2.2.

Table 2.5 shows the number of responses coded within each category for each principle. We used mixed-effects multinomial logistic regression to predict participants' explanations from principle, while controlling for the random effects of individual participants. We did not find a significant effect of principle, suggesting that the distribution of the types of explanations participants provided did not differ between principles.

Next, we used mixed-effects logistic regression to predict participants' binary choice in the test trials (BV response = 1, BC response = 0) from the type of explanation they provided, while controlling for the random effects of individual participants. We did not find a significant effect of the type of explanations, suggesting that the type of explanation did not significantly predict participants' choices in the test trials.

Table 2.5: Children’s Explanations by category and principle in Experiments 5 & 6

	Contact	Continuity	Solidity
Accept Evidence	9	5	8
Explain Away	9	11	9
Pattern	0	0	0
Other	1	3	2

Table 2.6: Proportion of *BV response* by Experiment, principle, trial type, and explanation type

Experiment	Principle	Trial Type	Explanation Type			
			Accept Evidence	Explain Away	Pattern	Other
Experiment 2	Contact	easy	0.75	0.67	1.00	0.25
		hard	0.75	0.56	1.00	0.25
	Continuity	easy	0.93	1.00	1.00	1.00
		hard	1.00	1.00	0.81	1.00
	Solidity	easy	1.00	1.00	1.00	0.90
		hard	1.00	0.83	1.00	1.00
Experiment 3	Contact	easy	0.67	0.69	1.00	0.80
		hard	0.67	0.65	1.00	0.70
		harder	0.67	0.38	1.00	0.40
	Continuity	easy	1.00	1.00	1.00	0.90
		hard	0.93	0.93	1.00	0.89
		harder	0.81	0.63	0.75	0.90
	Solidity	easy	1.00	0.82	1.00	0.67
		hard	1.00	0.73	0.75	1.00
		harder	1.00	0.77	1.00	0.67
Experiments 5 & 6	Contact	easy	0.50	0.50	N/A	1.00
		hard	0.63	0.43	N/A	1.00
		harder	0.75	0.57	N/A	0.00
	Continuity	easy	0.83	0.79	N/A	1.00
		hard	0.33	0.86	N/A	1.00
		harder	0.67	0.86	N/A	0.75
	Solidity	easy	0.50	0.71	N/A	1.00
		hard	0.83	0.86	N/A	0.75
		harder	0.50	0.57	N/A	0.50

2.8.3. Discussion

In Experiment 6, we found that like adults, children had strong prior beliefs about the Contact, Continuity, and Solidity principles, and they revised their beliefs given counterevidence. Children's performance did not differ in the easy, hard, and harder test trials, suggesting that they were willing to generalize their revised beliefs to new situations. Similar to adults, children also had weaker prior beliefs for the Contact and the Solidity principles than for the Continuity principle, and they were least likely to revise their beliefs about the Contact principle.

Participants' explanations of the belief-violating evidence suggest that a group of children (39%) had accepted the counterevidence and a larger group of children (51%) explained away the counterevidence. However, participants who provided "explain away", "pattern", or "other" explanations still predicted outcomes inconsistent with the principles in test trials most of the time, except in hard test trials for the Continuity principle for participants who provided "accept evidence" explanations, and in hard test trials for the Contact principle for participants who provided "explain away" explanations (Table 2.6). Unlike adults, the types of explanations children provided did not predict their behaviors in the test trials. One possible reason is that the sample size is much smaller for children than for adults. Another possible reason is that children have poorer verbal abilities, and their explanations might not accurately reflect their interpretation of the belief-violating evidence.

2.9. Experiments 4—6: Combined results and discussion

Next, we analyzed the combined results of Experiments 4—6. We used mixed-effect logistic regression to predict participants' binary choice (*BV response* = 1, *BC response* = 0) from condition, principle, test trial type, experiment, age, gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included the three-way interaction of condition, principle, and age as predictors.

Most importantly, for all 3 principles, participants were more likely to choose the *BV response* in the BV condition than in the Baseline condition (Continuity: $\beta = 5.24$, $SE = 0.63$, $p < .001$; Solidity: $\beta = 1.47$, $SE = 0.39$, $p < .001$; Contact: $\beta = 2.12$, $SE = 0.45$, $p < .001$) and the BC condition (Continuity: $\beta = 4.84$, $SE = 0.50$, $p < .001$; Solidity: $\beta = 3.04$, $SE = 0.39$, $p < .001$; Contact: $\beta = 1.90$, $SE = 0.36$, $p < .001$). For the Continuity and the Contact principles, participants' responses did not differ between the BC condition and the Baseline condition (Continuity: $\beta = 0.36$, $SE = 0.68$, $p = .60$; Contact: $\beta = 0.23$, $SE = 0.46$, $p = .62$). For the Solidity principle, participants were less likely to choose the *BV response* in the BC condition than in the Baseline condition ($\beta = -1.58$, $SE = 0.41$, $p < .001$). In the Baseline condition, children chose the *BV response* at chance for the Solidity principle (Exact binomial tests: $p = .41$) and below chance for the Continuity and Contact principles (*p*s < .001). In the BC condition, they chose the *BV response* below chance for all three principles (*p*s < .001). In the BV condition, they chose the *BV response* above chance for the Solidity and the Continuity principle (*p*s < .001), and marginally below chance for the Contact principle ($p = .07$).

In the Baseline and the BC conditions, compared to the Continuity principle, participants were more likely to choose the *BV response* for the Solidity (Baseline condition: $\beta = 3.23$, $SE = 0.56$, $p < .001$; BC condition: $\beta = 1.25$, $SE = 0.42$, $p = .003$) and the Contact principle (Baseline condition: $\beta = 1.80$, $SE = 0.60$, $p = .003$; BC condition: $\beta = 1.63$, $SE = 0.41$, $p < .001$). In the BV condition, participants were less likely to choose the *BV response* for the Contact principle compared to the Continuity principle (β = -1.32, *SE* = 0.29, *p* < .001) and the Solidity principle (β = -0.78, *SE* = 0.26, *p* = .002).

For the Contact principle, children were less likely to choose the *BV response* with increasing age in the Baseline condition (β = -1.39, *SE* = 0.49, *p* = .005), in the BC condition (β = -0.57, *SE* = 0.26, *p* = .03), and in the BV condition (β = -0.49, *SE* = 0.24, *p* = .045). This suggests that older children might have stronger prior beliefs about the Contact principle than younger children. For the Continuity principle, children were more likely to choose the *BV response* in the BV condition with increasing age (β = 0.97, *SE* = 0.32, *p* = .002), suggesting that older children were more likely to revise their beliefs given counterevidence than younger children. For the Solidity principle, children were less likely to choose the *BV response* in the BC condition with increasing age (β = -0.67, *SE* = 0.29, *p* = .02), suggesting that older children's beliefs about the Solidity principle were affected more by the belief-consistent evidence than younger children's.

There was no significant effect of test trial type (easy vs. hard vs. harder test trials), suggesting that children generalized their revised beliefs to new objects and events. There was no significant effect of experiment, suggesting that the different stimuli sets did not affect participants' choices in the test trials. To examine the effect of the amount of evidence on participants' choices, we analyzed the BC condition data across the 3 experiments and the BV condition data across the 3 experiments, respectively. The amount of evidence did not affect children's *BV response* in the BC condition or the BV condition.

In conclusion, across three experiments, we found consistent evidence that children can revise their beliefs about all three principles given counterevidence. Like adults, children might also have weaker prior beliefs for the Contact and Solidity principles than the Continuity principle, and they were also least likely to revise the Contact principle. I will return to these differences across principles in the general discussion. Moreover, children generalized their revised beliefs about these principles to new objects and contexts.

2.10. Experiments 1—6: Combined results and discussion

Lastly, we analyzed the combined results of all six experiments to compare adults' and children's performances (Figure 2.10). We used mixed-effect logistic regression to predict participants' binary choice (*BV response* = 1, *BC response* = 0) from condition, principle, test trial type, age group (adults vs. children), gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included the three-way interaction of condition, principle, and age group, and the two-way interaction of condition and test trial type as predictors.

Most importantly, for all 3 principles, participants were more likely to choose the *BV response* in the BV condition than in the Baseline condition (Continuity: β = 8.08, *SE* = 0.68, *p* < .001; Solidity: β = 6.14, *SE* = 0.51, *p* < .001; Contact: β = 4.30, *SE* = 0.48, *p* < .001) and the BC condition (Continuity: β = 7.36, *SE* = 0.52, *p* < .001; Solidity: β = 6.87, *SE* = 0.49, *p* < .001; Contact: β = 5.10, *SE* = 0.46, *p* < .001); their choices did not differ between the Baseline and the BC conditions (Continuity: β = -0.98, *SE* = 0.53, *p* = .065, Solidity: β = -0.98, *SE* = 0.53, *p* = .065; Contact: β = -0.82, *SE* = 0.50, *p* = .10).

The three-way interaction of condition, principle, and age group (adult vs. child) showed that, in the Baseline and BC conditions, children were more likely than adults to choose the *BV response* for the Solidity principle (Baseline: β = 3.16, *SE* = 0.56, *p* < .001; BC: β = 1.64, *SE* = 0.51, *p* = .001) and the Contact principle (Baseline: β = 1.57, *SE* = 0.56, *p* = .005; BC: β = 1.69, *SE* = 0.49, *p* < .001). In the BV condition, children were less likely than adults to choose the *BV response* for the Continuity principle (*β* = -1.01, *SE* = 0.45, *p* = .024) and the Solidity principle (*β* = -1.32, *SE* = 0.44, *p* = .003).

The interaction of condition and test trial type showed that, in the Baseline condition, participants were less likely to choose the *BV response* in the harder test trials than in the easy test trials (*β* = -0.76, *SE* = 0.34, *p* = .024) and the hard test trials (*β* = -0.67, *SE* = 0.33, *p* = .044). In the BC condition, participants were more likely to choose the *BV response* in the harder test trials than in the easy test trials (*β* = 0.63, *SE* = 0.29, *p* = .029). In the BV condition, participants were less likely to choose the *BV response* in the harder test trials than in the easy test trials (*β* = -0.95, *SE* = 0.21, *p* < .001) and the hard test trials (*β* = -0.78, *SE* = 0.21, *p* < .001). Importantly, participants still chose the *BV response* above chance in all 3 types of trials in the BV condition (*p*s < .001).

To examine the effect of the amount of evidence on participants' choices, we analyzed the BC condition data across the 6 experiments and the BV condition data across the 6 experiments, respectively. The amount of evidence did not affect participants' *BV response* in the BC condition or the BV condition.

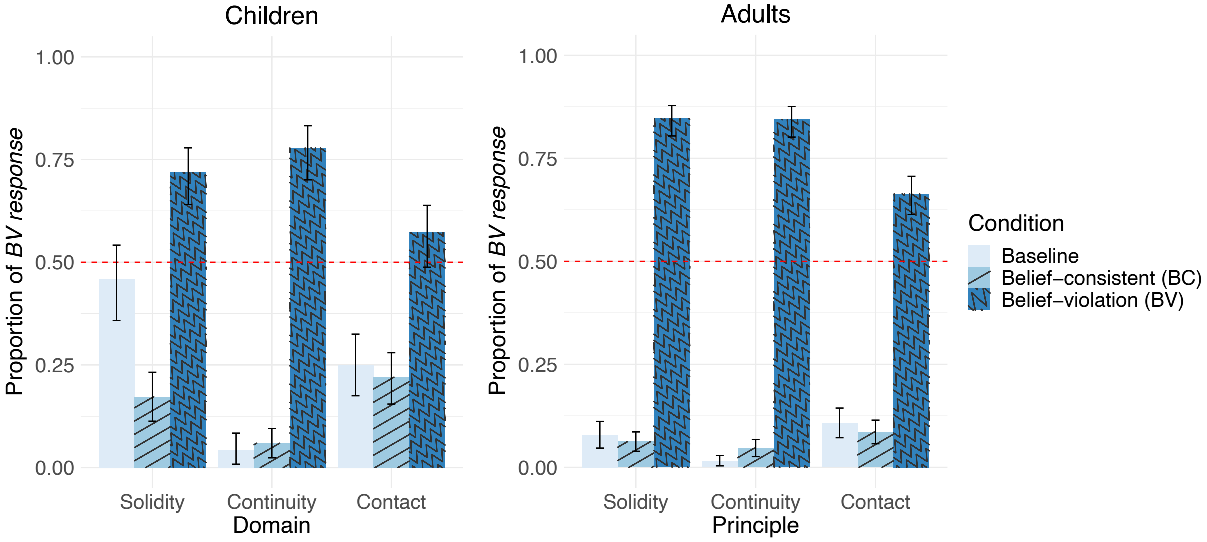

Figure 2. 10: The mean proportion of trials that children (left) and adults (right) selected the *belief-violation (BV) response* by condition and principle in Experiments 1—6. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

We next analyzed the combined explanation data for adults and children. Participants who provided "pattern" explanations were more likely to choose the *BV response* for the principle compared to participants who provided "explain away" explanations (*β* = 1.55, *SE* = 0.59, *p* = .008) or "other" explanations (*β* = 1.69, *SE* = 0.62, *p* = .01). Participants who provided "accept evidence" explanations were marginally more likely to choose the *BV response* for the principle compared to participants who provided "explain away" (*β* = 0.40, *SE* = 0.23, *p* = .09) or "other" explanations (*β* = 0.54, *SE* = 0.29, *p* = .07) for that principle. When we grouped "accept evidence" or "pattern" explanations into a single category, and grouped "explain away" and "other" explanations into a single category, we found that participants who provided "accept evidence" or "pattern" explanations were more likely to choose the *BV response* for the principle compared to participants who provided “explain away” or “other” explanations ($\beta = 0.57$, $SE = 0.21$, $p = .006$).

In conclusion, across six experiments, we found consistent evidence that both adults and children can revise their beliefs about the Solidity, Continuity, and Contact principles given counterevidence. Children had weaker prior beliefs about the Solidity and Contact principles than adults, and they were less likely to revise their beliefs about the Solidity and Continuity principles than adults. Moreover, while adults and children generalized their revised beliefs to new objects and new contexts, they were gradually less likely to generalize the revised beliefs to more different contexts. Adults’ and children’s explanations of the belief-violating evidence showed that they were more likely to revise their beliefs if they accepted the counterevidence or if they learned from the statistical pattern in the counterevidence, compared to if they explained away the counterevidence or provided irrelevant explanations.

2.11. General Discussion

In six experiments, we used a novel paradigm to measure adults’ and children’s prior beliefs about three core knowledge principles in the domain of object – Solidity, Continuity, and Contact – and we investigated whether they could revise these most fundamental beliefs about objects in a specific, virtual environment. We found that both adults and children have strong prior beliefs about these three principles. When they were not given any new evidence about these principles and when they observed a few events consistent with these principles, they predicted that objects would behave in accordance with these principles. The only exception is that children did not have strong prior beliefs about the Solidity principle, which might be due to perceptual ambiguities, as explained below. However, after observing just a few events of objects violating these principles, adults and children predicted that objects would behave inconsistently with the principles, and they generalized their inconsistent predictions to new objects and new contexts in this environment.

We had 3 hypotheses for whether children and adults would revise their beliefs given counterevidence about core principles about objects. Our first hypothesis was that the object representation and reasoning system would “shut down” upon observing violations of the core principles. Our second hypothesis was that children and adults would try to come up with alternative interpretations to explain away the counterevidence, and not revise their beliefs about these principles. Our third hypothesis was that children will accept the counterevidence and revise their beliefs, and they will generalize their revised beliefs to a certain extent (narrowly or widely). The finding that children and adults predicted inconsistent outcomes after receiving counterevidence provides clear evidence against our first hypothesis. To further examine whether the present findings support the second and third hypotheses, we will discuss the explanation data and the generalization trials below.

The explanation data suggest that not all adults and children accepted the belief-violating evidence completely. When asked to explain the belief-violating evidence, a relatively small portion of adults (27% on average) and children (39% on average) stated that they had actually accepted the counterevidence. These participants were indeed more likely to predict outcomes inconsistent with the principles than participants who explained away the counterevidence or provided other irrelevant responses. Thus, about a third of the participants genuinely accepted the counterevidence and revised the principles in this specific, virtual environment, providing partial support for our third hypothesis. A group of adults (12% on average) acknowledged the statistical pattern in the counterevidence, but did not state that they had accepted the counterevidence. These adults were also more likely to predict outcomes inconsistent with the principles than participants who explained away the counterevidence or provided irrelevant responses. One possibility is that they responded based on statistical learning mechanisms – they noticed the pattern that objects in this environment behave in ways that were inconsistent with the principles, and therefore predicted that other objects would behave according to this pattern. However, it is less clear whether these adults have genuinely revised their beliefs and whether they would generalize the revised principles to objects and events that are completely different from the ones involved in the counterevidence. A larger proportion of adults (43% on average) and children (51% on average) did not accept the counterevidence. Instead, they explained away the counterevidence with reasons that did not involve any violations of the principles, so that they did not have to revise the principles. They were indeed less likely to predict outcomes inconsistent with the principles for new events. This group of participants lends support to our second hypothesis. Would they accept the belief-violating evidence if they observed more evidence or a more diverse set of evidence? This is an interesting question for future research. Lastly, a group of adults (18% on average) and children (11% on average) said "I don't know" or referred to irrelevant aspects of the events when they were asked to explain the counterevidence. They were probably reluctant to accept the belief-violating evidence, but they also could not find good reasons to explain the counterevidence. These participants were also less likely to predict outcomes inconsistent with the principles for new events, suggesting they were less likely to have revised their beliefs.

We also examined the extent to which adults and children were willing to generalize their revised beliefs. We found that they generalized the revised beliefs to new objects and new contexts in this virtual environment, but they were conservative in their generalizations. As the objects and the contexts of the events became more and more different from the objects and contexts involved in demonstrating the belief violations, participants were gradually less likely to predict that the objects would behave inconsistently with the principles. This is rational, given that both children and adults have observed much more evidence supporting each principle in various contexts. It would not be rational for them to generalize the revised principle to all contexts without reservation after observing just 4 or 6 pieces of counterevidence in a specific context. Thus, the generalization data supported our third hypothesis, and showed that learners were willing to generalize their revised beliefs to wide contexts (albeit conservatively). One limitation of the present study is that we only provided violations in a limited context and tested a limited generalization context in our experiments. In future work, we can provide learners with more counterevidence and more diverse sets of counterexamples to see if more learners will revise their beliefs and generalize the revised beliefs to broader contexts.

We also discovered some differences across principles. Both adults and children had weaker prior beliefs about the Contact and Solidity principles than the Continuity principle, and they were less likely to revise their beliefs about the Contact principle than the other two principles. One possible reason that adults and children had weaker prior beliefs about the Solidity principle is because of perceptual ambiguities in the stimuli – in the particular event we used, there were a few ways that the solid object could have passed the first wall without violating the principle. For example, there could be a gap between the first wall and the occluder, or there could be a hole in the wall (even though these possibilities were clearly ruled out in the three-dimensional stimuli in Experiments 3 and 6, some participants still mentioned these reasons when asked to explain the belief-violating evidence). For the Contact principle, there might be more violations of this principle in modern society, for example, magnets, remote controls, light switches, etc. Infants, children, and adults appear to learn these new technologies without much difficulty, even when they do not understand the underlying physics. People may believe that the Contact principle is more probabilistic (i.e., there are more exceptions to this principle) compared to the other physical principles. As a result, when given no new evidence, people appear to have weaker prior beliefs about this principle – they were more likely to predict outcomes inconsistent with the Contact principle than the other principles. After observing a few events violating the Contact principle, participants might treat these events as exceptions to the principle, just like magnets and remote controls, and therefore they were less likely to revise the Contact principle based on the counterevidence. Future studies could examine whether there are indeed more violations of the Contact principle in real life by asking people to come up with exceptions for each physical principle.

The current study makes several important contributions to the study of reasoning about objects and cognitive development more generally. Past research has shown that when learners observe an object that violates core knowledge principles, they expect that there is something peculiar about that particular object that they should learn about (Stahl & Feigenson, 2015; 2017). The current study focused on a different question: if learners observe multiple violations of the core knowledge principles, can they revise their higher-level beliefs about the abstract principles governing object reasoning? In our study, adults and children's most common response was to explain away the violations, which is reasonable given that we have strong prior beliefs about these principles. However, a small group of adults and children accepted the violations and genuinely revised their beliefs about these principles, which demonstrates that learners are capable of revising the abstract, core knowledge principles about objects given multiple violations. In addition, learners generalized their revised beliefs to new objects and new contexts (albeit conservatively), suggesting that their learning and belief revision is at the level of the abstract principles, rather than at the level of the individual objects involved in the counterevidence.

Past research has also shown that children and adults rationally learn from new evidence to update their beliefs in various domains (e.g., Griffiths et al., 2011; Kimura & Gopnik, 2019; Kushnir & Gopnik, 2007; Lucas & Griffiths, 2010; Lucas et al., 2014a). The current study provides another strong piece of evidence that children and adults have powerful learning mechanisms, and even our most strongly held beliefs about objects are subject to revision given counterevidence.

Lastly, children and adults often immerse themselves in fictional worlds (e.g., movies, novels, magic shows, pretense), where they suspend their beliefs about the core physical principles. Is it possible that participants in our study treated the experiments as an imaginary world different from the real world, and therefore only revised the core knowledge principles in the imaginary world (e.g., Chandler & Lalonde, 1994; Johnson & Harris, 1994)? We believe that this is unlikely for several reasons. First, there are examples of violations of these principles in the real world. Remote controls, light switches, and magnets are all examples of violations of the contact principle. There are also some examples that appear to violate the solidity principle – toddlers enjoy putting objects in pots (where the object and the pot appear to occupy the same space simultaneously), and objects can be placed in water (where it is not always obvious that the water level rises). Infants, children, and adults seem to readily accept these violations in the real world, and learn to interact with objects in ways that might appear to violate the physical principles without much difficulty. Therefore, it is entirely possible for us to revise physical principles in the real world if we observe enough counterevidence. Second, we asked our participants to explain the violations of the core physical principles. Only a small portion of responses (3% of adults' responses and 12% of children's responses) mentioned magic (e.g., "It is a magic trick", "The car is magic"), and these explanations were coded as "explain away". Thus, only a small portion of adults and children thought that the violations happened in a magical/imaginary world, and possibly did not revise their beliefs about these principles in the real world. A larger portion of adults and children accepted the violations and stated that objects can behave in ways that are inconsistent with the principles (e.g., an object can go through a wall, or jump from one location to the other instantaneously). Lastly, our reasoning about how objects behave in fictional worlds is still connected to our intuitive theories about physics in the real world. In one study (McCoy & Ullman, 2019), adults were asked to estimate the efforts required to cast spells that cause various physical violations (e.g., transform a frog into a mouse, teleport a frog 100 feet forward). Adults' judgments were guided by intuitive physics – their judgments were affected by the extent of the spell (e.g., teleporting a frog by 100 feet is more effortful than teleporting a frog by 1 foot); and their judgments were consistent across individuals, and correlated with the age when the corresponding principles are acquired in development (e.g., the spells that were judged as most effortful were violations of object permanence and cohesion, which are also the earliest developing, most fundamental principles). Thus, it is unlikely that adults and children adopted a set of entirely different physical principles to reason about the violations they observed in our experiments, without revising their beliefs about the core physical principles in the real world.

In conclusion, the current research shows that adults and 4- to 6-year-olds have powerful statistical learning mechanisms, and they can revise their most fundamental beliefs about objects when provided with small amounts of statistical evidence.

Chapter 3: Children and Adults Revise Core Knowledge Principles About Agents

3.1. Introduction

Another well-studied core knowledge system guides how we represent and reason about agents. Between the ages of 6 to 12 months, infants understand that agents' intentional actions are directed to goals (Woodward, 1998), agents choose efficient means to achieve their goals (Gergely & Csibra, 2003), and agents' preferences can be inferred based on violations of random sampling (Wellman et al., 2016). These principles support further learning in the psychological domain (e.g., Jara-Ettinger et al., 2015; Kushnir et al., 2010; Sodian et al., 2016). They also persist into adulthood – these principles underlie adults' mental state reasoning and action understanding in complex scenes (Baker et al., 2017, Jara-Ettinger et al., 2020). Are these most fundamental core principles about agents subject to revision once we acquire them? If children and adults are given enough evidence that violates these principles, will they rationally update their beliefs?

Past research has shown that children and adults are sensitive to evidence that violates the core knowledge principles about agents and use this evidence to update their beliefs. Studies using the violation of expectation paradigm (VOE) have demonstrated that infants are surprised and look longer at events that violate the Efficiency, Goal, and Sampling principles than events that are consistent with these principles (Gergely & Csibra, 2003; Wellman et al., 2016; Woodward, 1998). Adults are also surprised by apparent violations of the core psychological principles. When they observed events that violated the Goal and Efficiency principles, they rated these as more surprising than events that did not violate the principles (Shu et al., 2021). Furthermore, children update their beliefs about the agents who violated the core knowledge principles. Toddlers expected a neutral observer to approach an agent who behaved efficiently instead of an agent who behaved inefficiently (Colomer et al., 2020). In addition, toddlers were less likely to learn novel words from an inefficient agent compared to an efficient agent (Colomer & Woodward, 2023). These findings suggest that observing violations of core psychological principles provides an opportunity for learners to learn something about the particular agents who violated the principles (e.g., their knowledgeability and third-party preferences for these agents). However, no past studies have investigated whether learners can revise the fundamental, abstract principles that govern agent reasoning since infancy, given multiple violations of the principles.

In the present chapter, we examine this question with 3 core principles in the agent system: the Efficiency principle – agents choose efficient means to achieve their goals, the Goal principle – agents' intentional actions are directed to goals, and the Sampling principle – agents' preferences can be inferred based on violations of random sampling. The methods of the experiments were similar to the methods of the experiments for the object domain in Chapter 2. In 6 experiments, participants observed events that supported or violated these principles, or they did not receive any new evidence about these principles. Then, they were asked to make predictions about the outcomes of new events that varied in the extent to which they were different from the original events. For each new event, participants were asked to predict whether agents would behave in ways consistent with the principles or inconsistent with the principles.

We investigated this question with geometric-shaped agents in virtual environments. Past research suggests that children and adults expect geometric-shaped agents to behave similarly to humans in the real world in terms of the core psychological principles (Baker et al., 2017; Jara-

Ettinger et al., 2015; Jara-Ettinger et al., 2020; Shu et al., 2021). Thus, we expect our results to generalize to children's and adults' reasoning about humans in the real world.

Unlike the object system (Chapter 2), we do not hypothesize that the system responsible for representation and reasoning about agents would be disrupted upon encountering belief-violating evidence. Past research has shown that when children observe an agent violate the Efficiency principle, they continue to represent and reason about that agent, and they use that information to update their beliefs about that agent (Colomer et al., 2020; Colomer & Woodward, 2023). In addition, we are more flexible in reasoning about agents in real life – when we see a person take a detour to get to her goal, we can think of various reasons to explain her behavior, for example, she might prefer a more scenic path, or she might want to get some exercise. Thus, it is unlikely that the system responsible for reasoning about agents would completely shut down when participants observe belief-violating evidence in our experiments.

We compare 2 hypotheses for how children and adults would respond to the belief-violating evidence and whether they would revise their beliefs about the core psychological principles. These hypotheses parallel the second and the third hypotheses we tested in Chapter 2 for the object domain. Our first hypothesis is that the core knowledge principles are relatively robust in the sense that children and adults would try to come up with alternative interpretations to explain away the violations, and not revise their beliefs about the principles. This hypothesis predicts that when children and adults are asked to make predictions about the same agents, they will predict outcomes inconsistent with the principles, since the alternative interpretations may still apply. However, when they are asked to make predictions about different agents, the alternative interpretations no longer apply, and they will predict outcomes consistent with the principles. In addition, when asked to explain the belief-violating evidence, participants would be more likely to provide alternative interpretations of the evidence instead of stating that they accept the evidence.

Our second hypothesis is that children and adults will accept the evidence violating the principles, and use this evidence to revise their beliefs about the principles. This hypothesis predicts that participants who saw the belief-violating evidence would be more likely to predict outcomes that are inconsistent with the principles, compared to those who saw the belief-consistent evidence and those who did not receive new evidence. In addition, when asked to explain the belief-violating evidence, participants would state that they have accepted the violations of the principles in the belief-violating evidence. After children and adults accept the counterevidence and revise their beliefs about the principles, they might generalize their beliefs narrowly, only to the same agents, and only predict inconsistent outcomes for the same agents. Alternatively, they might generalize their beliefs widely, to different agents, and predict inconsistent outcomes for both the same agents and different gents. They might also be gradually less likely to generalize their revised beliefs as the agents become more and more different from the original agents in the counterevidence.

We tested adults in Experiments 1—3, and young children (4- to 6-year-olds) in Experiments 4—6. Since adults might have observed both more evidence consistent with these principles and more evidence inconsistent with these principles than children in real life, we will compare adults' and children's results to examine whether there are any age differences in the strength of their prior beliefs about the core psychological principles, and the extent to which they are willing to revise their beliefs given counterevidence.

3.2. Experiment 1

3.2.1. Methods

Participants

Forty-seven adults (mean age = 30 years; range = 18 to 55; *SD* = 9.2; 25 females) participated on Prolific, an online platform for behavioral experiments. Participants provided written informed consent prior to the experiment, and they were paid $3.2 for a 20-minute experiment.

Design

Participants were randomly assigned to one of the two conditions, the Belief Consistent (BC) condition and the Belief Violation (BV) condition. They were tested on 3 principles, Efficiency, Goal, and Sampling, in counterbalanced orders. For each principle, there were 3 familiarization trials and 4 test trials (2 easy test trials and 2 hard test trials; order counterbalanced). The familiarization trials in the BC condition displayed events that were consistent with the principle and those in the BV condition displayed events that violated the principle. In test trials, participants chose between the *Belief Consistent (BC) response* and the *Belief Violation* (*BV*) *response*. They never received feedback about whether their choices were correct or incorrect.

Stimuli and procedure

Participants were instructed to watch some events and make predictions about new events.

Efficiency principle. In the familiarization trials, a grey wall and 2 agents (i.e., geometric shapes with eyes) appeared. One agent went toward the other agent by jumping over the wall. Then, the wall was moved to the side. The agent went toward the other agent by taking a straight path (BC condition) or a jumping path (BV condition) (Figure 3.1). The goal was a different agent in each familiarization trial.

In the easy test trials, the same agent went toward a new geometric-shaped agent by jumping over a wall. Then, the wall was moved to the side. A red path and a blue path indicated the straight path (the *BC response*) and the jumping path (the *BV response*). Participants were asked, “The red kid wants to play with the purple kid. Which path do you think the red kid will take to get to the purple kid? The blue path or the red path?” Participants chose their responses, either the blue, straight path (the *BC response*) or the red, jumping path (the *BV response*). In the hard test trials, a new geometric-shaped agent went toward another geometric-shaped agent by jumping over a wall. Then, the wall was moved to the side. Participants chose which path the agent would choose to get to the other agent, either the blue, straight path (the *BC response*) or the red, jumping path (the *BV response*) (Figure 3.1).

Goal principle. In the familiarization trials, an agent and 2 objects appeared. The agent went toward one of two objects and took the object 3 times. Then, the two objects switched locations. The agent took the old object at the new location (BC condition) or the new object at the old location (BV condition) (Figure 3.1). A different pair of objects was used in each familiarization trial.

In the easy test trials, a new pair of objects appeared. The same agent took one of the objects 3 times. Then the two objects switched locations. Participants were asked, “Which toy do you think the pink kid will take? The drum or the fox?” Participants chose their responses, either the old object at the new location (*BC response*) or the new object at the old location (*BV* *response*). In the hard test trials, a new geometric-shaped agent and a new pair of objects appeared. The agent took one of the objects 3 times. Then the two objects switched locations. Participants chose which object the agent would take, either the old object at the new location (*BC response*) or the new object at the old location (*BV response*) (Figure 3.1).

Sampling principle. In the familiarization trials, an agent and a box of objects appeared. The box contained 7 objects of one type and 31 objects of the other type. The agent picked out 4 objects of the minority type from the box, and put them into a small box in front of the agent. Then, an object of the minority type and an object of the majority type appeared, equidistant from the agent. The agent went toward the minority type (BC condition) or the majority type (BV condition) (Figure 3.1). A different toy box was used in each familiarization trial.

In the easy test trials, the same agent sampled 4 objects of the minority type from a new toy box. Then, an object of the minority type and an object of the majority type appeared, equidistant from the agent. Participants were asked, "Which toy do you think the green kid likes better, the black toy or the blue toy?" Participants chose their responses, either the minority-type object (*BC response*) or the majority-type object (*BV response*). In the hard test trials, a new geometric-shaped agent sampled 4 objects of the minority type from a new toy box. Then, an object of the minority type and an object of the majority type appeared, equidistant from the agent. Participants chose which object the agent liked better, either the minority-type object (*BC response*) or the majority-type object (*BV response*) (Figure 3.1).

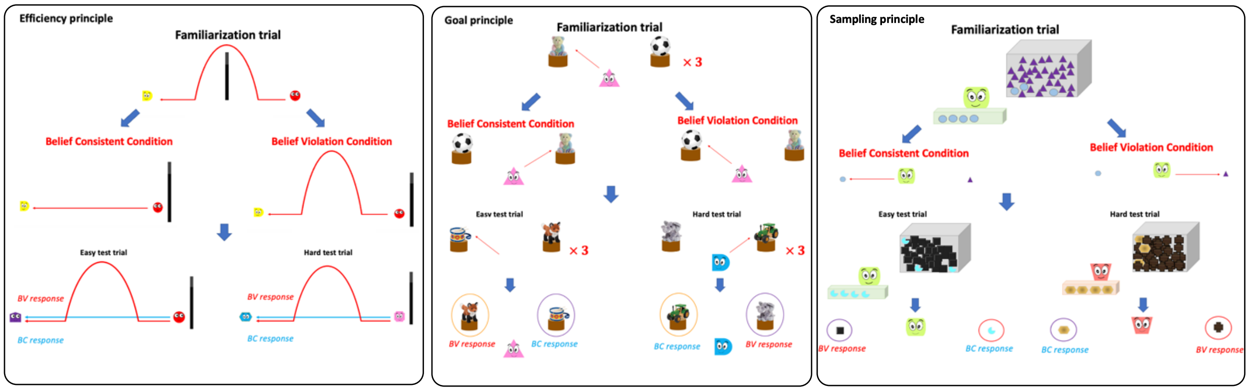

Figure 3. 1: Events shown in the familiarization trials and test trials for the Efficiency principle, the Goal principle, and the Sampling principle in Experiment 1.

3.2.2. Results

The proportion of *BV response* by condition and principle is shown in Figure 3.2. We used mixed-effect logistic regression to predict participants' binary choice (*BV response* = 1, *BC response* = 0) from condition, principle, test trial type, age, and gender, while controlling for the random effects of individual participants. The best-fitting model included condition and principle as predictors. Most importantly, participants were more likely to choose the *BV response* in the BV condition than in the BC condition for all 3 principles (Efficiency: $\beta = 4.48$, $SE = 0.92$, $p < .001$; Goal: $\beta = 4.35$, $SE = 0.94$, $p < .001$; Sampling: $\beta = 6.55$, $SE = 1.35$, $p < .001$). In the BC condition, participants were more likely to choose the *BV response* for the Efficiency and Goal principles than for the Sampling principle (Efficiency: $\beta = 3.08$, $SE = 1.10$, $p = .005$; Goal: $\beta = 2.46$, $SE = 1.12$, $p = .03$). In the BV condition, participants were more likely to choose the *BV* *response* for the Efficiency than for the Goal principles ($\beta = 0.75$, $SE = 0.38$, $p = .045$), and the Sampling principle ($\beta = 1.01$, $SE = 0.38$, $p = .007$).

Next, we compared participants' responses against chance. In the BC condition, participants chose the *BV response* below chance for all three principles (Exact binomial tests: *p*s < .001). In the BV condition, participants chose the *BV response* above chance for the Efficiency principle ($P_{Efficiency} = .69$ [.58, .78], $p < .001$), and at chance for the Goal and the Sampling principles ($P_{Goal} = .57$ [.47, .67], $p = .18$; $P_{Sampling} = .53$ [.43, .63], $p = .61$).

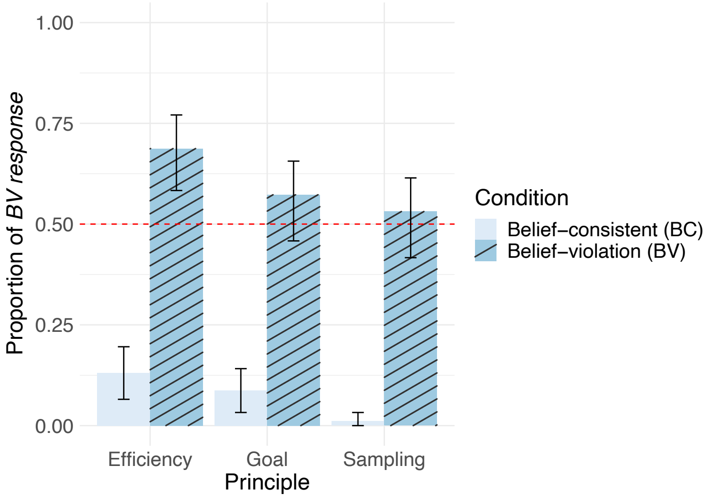

Figure 3. 2: The mean proportion of trials that adults selected the *belief-violation (BV) response* by condition and principle in Experiment 1. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

3.2.3. Discussion

Experiment 1 examined whether adults can revise their beliefs about the Efficiency, Goal, and Sampling principles given counterevidence, in a specific, virtual environment. The results showed that adults can revise their beliefs about all 3 principles, when they were given 3 pieces of counterevidence for each principle. After observing evidence consistent with these principles, adults predicted that the outcomes of new events would be consistent with the principles. However, after observing evidence violating these principles, they were more likely to predict outcomes inconsistent with the principles. Moreover, adults' performance did not differ in the easy and hard test trials, suggesting that they also generalized their revised beliefs to new situations.

We also found an important effect of principle. Regardless of the evidence they observed, adults were more likely to predict inconsistent outcomes for the Efficiency principle than for the Sampling principle. This suggests that they might have weaker prior beliefs for the Efficiency principle than the Sampling principle.

In the next experiment, we aimed to replicate these findings. In addition, we measured adults' prior beliefs about these principles and tested the effect of belief-violating evidence on their prior beliefs. We also increased the strength of the belief-violating evidence and tested its effect on their beliefs. Lastly, we assessed participants' interpretations of the evidence to see if they accepted the counterevidence.

3.3. Experiment 2

3.3.1. Methods

Participants

Sixty adults (mean age = 33 years; range = 18 to 54; *SD* = 9.41; 35 females) participated on Prolific. The sample size was determined based on the effect sizes observed in Experiment 1. The sample size in this experiment provided us with at least 95% power (at $\alpha = .05$) to detect the effect sizes observed in Experiment 1. Participants provided written informed consent prior to the experiment, and they were paid $4 for a 25-minute experiment.

Design and Procedure

The procedure of Experiment 2 was similar to that of Experiment 1, with a few modifications. First, we added a third, Baseline condition, where participants did not receive any new evidence that supported or violated the principles. Participants were randomly assigned to the Baseline condition, the BC condition, and the BV condition. Second, we increased the strength of the evidence; participants were shown 6 familiarization trials for each principle. Thus, for each principle, there were 6 familiarization trials and 4 test trials (2 easy test trials and 2 hard test trials; order counterbalanced). Last, participants in the BC and BV conditions were asked to explain one of the familiarization events for each principle after they completed the test trials for that principle.

Efficiency principle. The events used in the familiarization trials in the BC and BV conditions and the test trials in all 3 conditions were the same as in Experiment 1. In the familiarization trials of the Baseline condition, after the wall was moved to the side, the agent did not go toward the other agent (Figure 3.3).

For the explanation question, participants were asked to explain why the agent took the jumping path/the straight path to get to the other agent.

Goal principle. The events used in the familiarization trials in the BC and BV conditions and the test trials in all 3 conditions were the same as in Experiment 1. In the familiarization trials of the Baseline condition, after the objects switched locations, the agent did not take either object (Figure 3.3).

For the explanation question, participants were asked to explain why the agent took the respective object during the last time.

Sampling principle. The events used in the familiarization trials in the BC and BV conditions and the test trials in all 3 conditions were the same as in Experiment 1. In the familiarization trials of the Baseline condition, when an object of the minority type and an object of the majority type appeared, the agent did not go toward either object (Figure 3.3).

For the explanation question, participants were asked to explain why the agent went toward the respective toy.

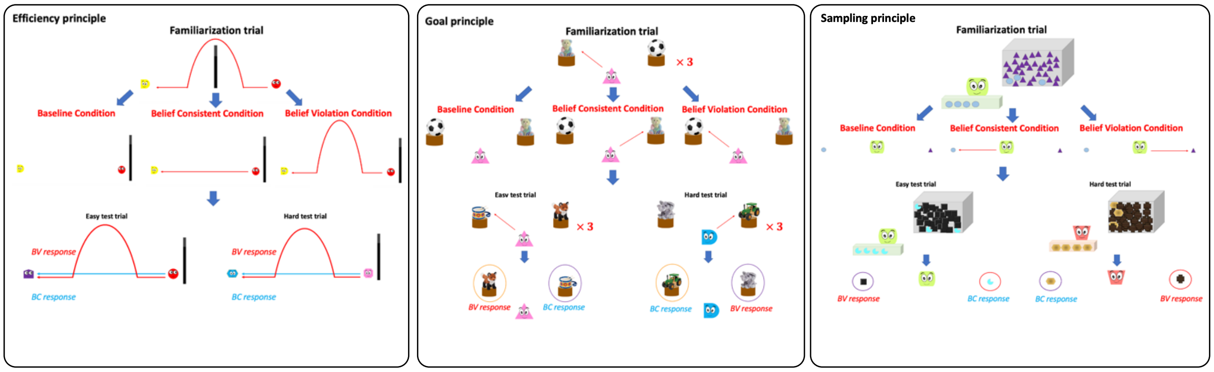

Figure 3. 3: Events shown in the familiarization trials and test trials for the Efficiency principle, the Goal principle, and the Sampling principle in Experiment 2.

3.3.2. Results

The proportion of BV response by condition and principle is shown in Figure 3.4. We used mixed-effect logistic regression to predict participants' binary response (*BV response* = 1, *BC response* = 0) from condition, principle, test trial type, age, and gender, while controlling for the random effects of individual participants. The best-fitting model included condition and principle as predictors. Mostly importantly, across principles, participants were more likely to choose the *BV response* in the BV condition than in the Baseline condition (β = 5.39, *SE* = 0.88, *p* < .001) and the BC condition (β = 7.81, *SE* = 1.10, *p* < .001), and they were less likely to choose the *BV response* in the BC condition than in the Baseline condition (β = - 2.42, *SE* = 0.93, *p* = .009). In the Baseline and the BC conditions, participants chose the *BV response* below chance for all three principles (Exact binomial tests: *p*s < .001). In the BV condition, participants chose the *BV response* above chance for the Efficiency and the Goal principles ($P_{Efficiency}$ *= .93 [.85, .97], p < .001;* P_{Goal} *= .83 [.74, .91], p < .001*), and at chance for the Sampling principle ($P_{Sampling}$ = .55 [.44, .66], *p* = .45).

The effect of principle showed that, compared to the Goal principle, participants were more likely to choose the *BV response* for the Efficiency principle (β = 1.26, *SE* = 0.34, *p* < .001), and less likely to choose the *BV response* for the Sampling principle (β = -2.43, *SE* = 0.41, *p* < .001).

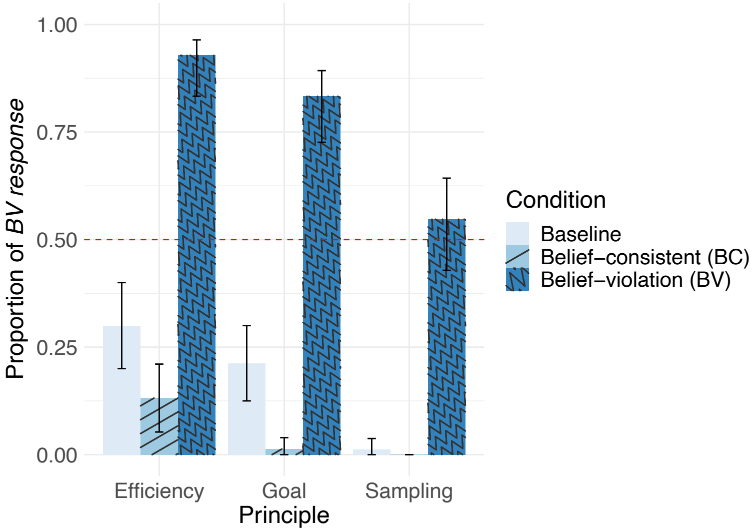

Figure 3. 4: The mean proportion of trials that adults selected the *belief-violation (BV) response* by condition and principle in Experiment 2. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

Table 3.1: Coding criteria and examples for explanations

Category	Criterion
Accept Evidence	Accepted the violation of the target principle in the counterevidence.
Explain Away	Explained the counterevidence with reasons that would not involve any violations of the target principle.
Pattern	Noted the pattern in the evidence.
Other	Explanations that cannot be categorized into the first three categories.

For the explanation questions, 2 researchers coded participants' responses into different categories (interrater reliability was excellent, Cohen's Kappa = .89; disagreements were resolved through discussion). In the BC condition, most responses (89.5%) referred to the principle itself to explain the evidence (other responses were irrelevant to the principle or incomprehensible). In the BV condition, we coded participants' explanations into 4 categories based on the criteria in Table 3.1.

Table 3.2: Examples for each type of explanation in the belief-violation (BV) condition

	Adults	Children
Accept Evidence	Efficiency: The kid continues to take the jump because he enjoys it more. Goal: The kid only wanted toys from the left side. Sampling: The kid likes to weed out the toys he does not like from the big box.	Efficiency: Because it was fun to jump. Goal: Because it liked that side more than that side. Sampling: Because he likes it better.
Explain Away	Efficiency: The kid thought there was still a wall. Goal: Because the bear moved places and the cartoon could only see one toy based on how his eyes were drawn. Sampling: Got bored of the other toy.	Efficiency: There is an invisible wall in the middle. Goal: Because he was done playing with the soccer ball. Sampling: Because he wants to play with another type of toy.
Pattern	Efficiency: Because that is what the red kid had previously done. Goal: Three times of one toy, the fourth time picks the other toy. Sampling: Every fifth toy they pick the opposite.	None.
Other	Efficiency: When there is no wall the parabola is smaller. Goal: I have no idea. Sampling: Not as many.	Efficiency: When there is no wall the parabola is smaller. Goal: I have no idea. Sampling: Not as many.

Table 3.3 shows the number of responses in each category for each principle. We used mixed-effects multinomial logistic regression to predict participants' explanations from principle, while controlling for the random effects of individual participants. We found an effect of principle. Participants were more likely to provide "accept evidence" explanations than any other types of explanations for the Efficiency and Goal principles than for the Sampling principles (*p*s < .02).

Table 3.3: Adults' Explanations by category and principle in Experiment 2

	Efficiency	Goal	Sampling
Accept Evidence	17	15	3
Explain Away	2	1	6
Pattern	1	2	6
Other	1	3	6

Next, we used mixed-effects logistic regression to predict participants' binary choice in the test trials (*BV response* = 1, *BC response* = 0) from the type of explanation they provided, while controlling for the random effects of individual participants. Participants were more likely to choose the *BV response* for the principle if they provided "accept evidence" explanations, compared to any other types of explanations ("explain away": $\beta = 2.78$, $SE = 0.62$, $p < .001$; "pattern": $\beta = 1.41$, $SE = 0.68$, $p = .040$; "other": $\beta = 1.93$, $SE = 0.60$, $p = .001$). Table 3.6 shows the proportion of *BV responses* by participants who provided different types of explanations.

3.3.3. Discussion

Experiment 2 assessed adults' prior beliefs about the Efficiency, Goal, and Sampling principles, and examined whether adults can revise their prior beliefs about these principles given counterevidence. We found that adults had strong prior beliefs about these principles. Most adults who did not receive any new evidence expected agents to behave in ways consistent with these principles. After observing evidence supporting these principles, their prior beliefs were strengthened. Furthermore, we found that adults can revise their beliefs about these principles given counterevidence. After observing evidence violating these principles, they were more likely to predict that agents would behave in ways inconsistent with the principles. Moreover, adults' performance did not differ in the easy and hard test trials, suggesting that they generalized their revised beliefs to new agents.

We also found an important effect of principle. Across conditions, adults' likelihood of choosing the inconsistent outcome was higher for the Efficiency principle than the Goal principle, which was in turn higher than the Sampling principle. This suggests that adults had stronger prior beliefs for the Sampling principle than the Goal principle, and stronger prior beliefs for the Goal principle than the Efficiency principle. There was no interaction between principle and condition, suggesting that across the 3 principles, the belief-violating evidence had similar effects on participants' prior beliefs.

Participants' explanations for the evidence showed that a majority of participants (56%) accepted the belief-violating evidence. Other participants simply learned from the statistical pattern of the evidence or explained away the counterevidence. However, participants who provided "explain away", "pattern" or "other" types of explanations still predicted outcomes that violated the principles in test trials most of the time, except participants who provided "explain away" explanations for the Goal principle, or those who provided "other" explanations for the Sampling principle (Table 3.6). Participants were more likely to accept the belief-violating evidence for the Efficiency and Goal principles than the Sampling principle. This suggests that the counterevidence for the Sampling principle might not be as compelling as the counterevidence for the other 2 principles. More importantly, participants who had accepted the counterevidence were indeed more likely to predict outcomes that violated the principles.

In the next experiment, we aim to replicate these findings with more realistic, three-dimensional stimuli, and investigate whether adults can generalize their revised beliefs to agents that are not geometric shapes.

3.4. Experiment 3

3.4.1. Methods

Participants

Eighty-two undergraduate Psychology students (mean age = 20.28 years; range = 18 to 36; *SD* = 2.54; 65 females, 15 males, 2 of unknown gender) participated on an online research platform at a university. The sample size was determined based on the effect sizes observed in Experiment 1. The sample size in this experiment provided us with at least 95% power (at α

= .05) to detect the effect sizes observed in Experiment 1. Participants provided written informed consent prior to the experiment, and they received 0.5 course credit for a 25-minute experiment.

Design and Procedure

The stimuli and procedure of Experiment 3 were similar to that of Experiment 2, with a few important modifications. First, we used photorealistic, three-dimensional stimuli made with Blender. Second, we added 2 harder test trials for each principle, where participants were asked to predict how animals would behave in the same situation (Figure 3.5). Third, participants in the Baseline condition were also asked the explanation questions. Instead of explaining an event in the familiarization trial, they were asked to explain their predictions in an easy test trial. For example, for the Efficiency principle, participants were asked, "You predicted that [the agent] would take [participants' response] to get to [the other agent] when there is no wall. Why is that the case? Why would [the agent] take [participants' response]?"

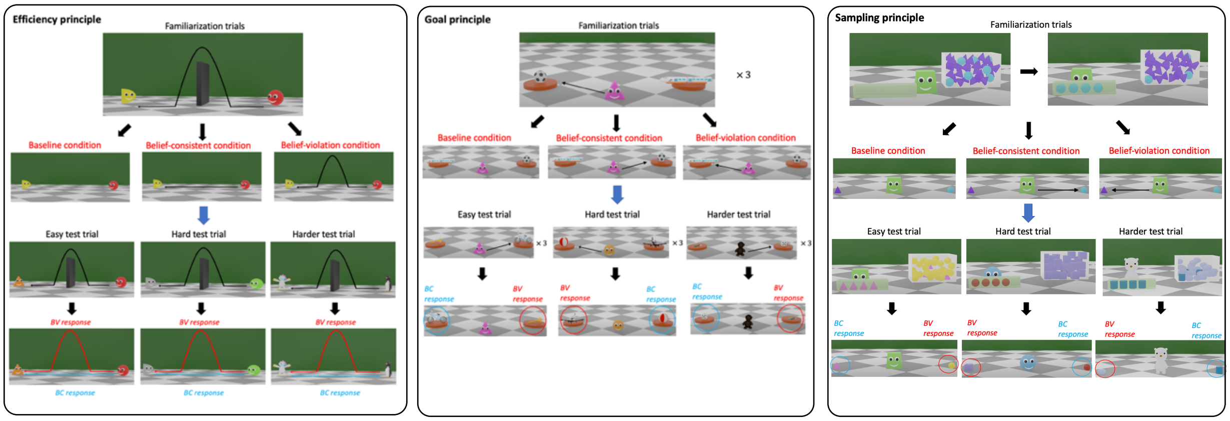

Figure 3. 5: Events shown in the familiarization trials and test trials for the Efficiency principle, the Goal principle, and the Sampling principle in Experiment 3.

3.4.2. Results

The proportion of *BV response* by condition and principle is shown in Figure 3.6. We used mixed-effect logistic regression to predict participants' binary response (*BV response* = 1, *BC response* = 0) from condition, principle, test trial type, age, and gender, while controlling for the random effects of individual participants. The best-fitting model included condition and principle as predictors. Across 3 principles, participants were more likely to choose the *BV response* in the BV condition than in the Baseline condition ($\beta = 4.92$, $SE = 0.73$, $p < .001$) and the BC condition ($\beta = 5.82$, $SE = 0.79$, $p < .001$); their choices did not differ between the Baseline and the BC conditions ($\beta = -0.90$, $SE = 0.70$, $p = .20$). In the Baseline and the BC conditions, participants chose the *BV response* below chance for all three principles (Exact binomial tests: *p*s < .001). In the BV condition, participants chose the *BV response* above chance for the Efficiency and the Goal principles ($P_{Efficiency} = .86$ *[.79, .91], p < .001;* $P_{Goal} = .74$ *[.66, .81], p < .001*), and at chance for the Sampling principle ($P_{Sampling} = .49$ [.40, .57], $p = .80$).

Compared to the goal principle, participants were more likely to choose the *BV response* for the Efficiency principle ($\beta = 0.55$, $SE = 0.21$, $p = .009$), and less likely to choose the *BV response* for the Sampling principle ($\beta = -1.52$, $SE = 0.24$, $p < .001$).

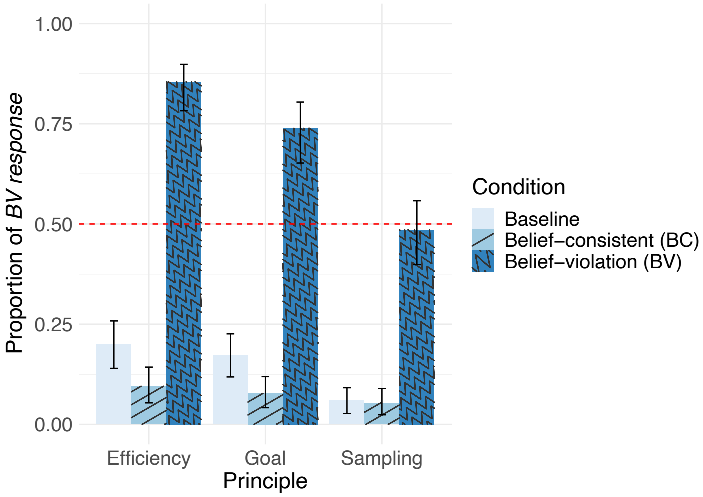

Figure 3. 6: The mean proportion of trials that adults selected the *belief-violation (BV) response* by condition and principle in Experiment 3. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

For the explanation questions, 3 researchers coded participants' responses (interrater reliability was good, Light's Kappa = .72; disagreements were resolved through discussion). In the Baseline and the BC condition, most responses (97.6% and 94.4%) referred to the principle itself to explain the evidence or their predictions (other responses were irrelevant to the principle or incomprehensible). In the BV condition, we categorized participants' explanations into four categories. The criteria for categorization and examples are shown in Table 3.1 and Table 3.2.

Table 3.4 shows the number of responses in each category for each principle in the BV condition. Mixed-effects multinomial logistic regression revealed that participants were more likely to provide "accept evidence" explanations than "explain away" explanations for the Goal principle than for the Efficiency principle ($\beta = 1.79$, $SE = 0.83$, $p = .03$); they were more likely to provide "accept evidence" explanations than "other" explanations for the Goal principle than for the Sampling principle ($\beta = 2.24$, $SE = 1.13$, $p = .047$).

Table 3.4: Adults' Explanations by category and principle in Experiment 3

	Efficiency	Goal	Sampling
Accept Evidence	16	19	10
Explain Away	4	1	4
Pattern	1	2	3
Other	2	1	6

Mixed-effects logistic regression revealed that participants were more likely to choose the *BV response* if they provided "accept evidence" explanations for that principle, compared to if they provided "explain away" (β = 1.30, *SE* = 0.46, *p* = .004) or "other" (β = 1.54, *SE* = 0.44, *p* < .001) explanations. Table 3.6 shows the proportion of *BV responses* by participants who provided different types of explanations.

3.4.3. Discussion

Experiment 3 replicated the main findings of Experiments 1 and 2 in a photorealistic, three-dimensional environment. Adults had strong prior beliefs about the Efficiency, Goal, and Sampling principles. Most adults who did not receive any new evidence expected agents to behave in ways consistent with the principles. Unlike Experiment 2, we did not find statistically significant evidence that observing evidence supporting the principles further strengthened their prior beliefs. After observing evidence violating these principles, adults revised their prior beliefs in this specific context. Their performance did not differ in the easy, hard, and harder test trials, suggesting that they generalized their revised beliefs even to new agents that were not geometric shapes.

We also replicated the effect of principle in Experiments 1 and 2. Adults' prior beliefs were stronger for the Sampling principle than the Goal principle, which were in turn stronger than their prior beliefs for the Efficiency principle. We will return to this effect of principle in the general discussion.

A majority of participants (65%) accepted the counterevidence. Other participants who provided "explain away", "pattern" or "other" types of explanations still predicted outcomes that violated the principles in test trials most of the time, except participants who provided "explain away" explanations for the Goal principle, or those who provided "other" explanations for the Sampling principle (Table 3.6). There was also some evidence that participants were more likely to accept the counterevidence for the Goal principle than the other 2 principles. It is possible that the counterevidence for the Goal principle was more compelling than the other 2 principles, or that it was less likely to come up with alternative explanations to explain away the counterevidence for the Goal principle. We again found that participants who accepted the counterevidence were indeed more likely to predict outcomes that violated the principles, replicating the finding in Experiment 2.

3.5. Experiments 1—3: Combined results and discussion

Next, we analyzed the combined results of Experiments 1—3. We used mixed-effect logistic regression to predict participants' binary choice (BV = 1, BC = 0) from condition, principle, trial type, experiment, age, gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included condition, the interaction of principle and age, and the interaction of principle and gender as predictors.

Most importantly, for all 3 principles participants were more likely to choose the *BV response* in the BV condition than in the Baseline (β = 5.00, *SE* = 0.55, *p* < .001) and the BC conditions (β = 6.15, *SE* = 0.53, *p* < .001); they were less likely to choose the *BV response* in the BC condition than in the Baseline condition (β = -1.15, *SE* = 0.53, *p* = .03). There was no significant effect of test trial type (easy vs. hard vs. harder test trials), suggesting that adults generalized their revised beliefs to new agents. In the Baseline and the BC conditions, participants chose the *BV response* below chance for all three principles (Exact binomial tests: *p*s < .001). In the BV condition, participants chose the *BV response* above chance for the Efficiency and Goal principles (*p*s < .001), and at chance for the Sampling principle (*p* = .61).

The interaction of principle and experiment showed that adults were more likely to choose the *BV response* for the Efficiency principle in Experiment 2 than in Experiment 1 (β = 1.55, *SE* = 0.61, *p* = .01). Since we varied the amount of evidence across Experiments 1 and 2, it is possible that this difference between experiments was driven by the difference in the BV condition – adults were more likely to choose the *BV response* in the BV condition when given 6 pieces of counterevidence (Experiment 2) than when given 3 pieces of counterevidence (Experiment 1).

The interaction of principle and age showed that adults were less likely to choose the *BV response* for the Goal principle with increasing age, compared to the Efficiency principle (β = -0.54, *SE* = 0.23, *p* = .02) and the Sampling principle (β = -0.57, *SE* = 0.24, *p* = .02). This suggests that older adults have stronger prior beliefs about the Goal principle compared to the other 2 principles.

The interaction of principle and gender showed that females were more likely than males to choose the *BV response* for the Efficiency principle (β = 1.01, *SE* = 0.48, *p* = .03), suggesting that males might have stronger beliefs for the Efficiency principle than females.

To examine the effect of the amount of evidence on participants' choices, we analyzed the BC condition data across the 3 experiments and the BV condition data across the 3 experiments, respectively. In the BC condition, the amount of evidence did not affect participants' *BV response*. In the BV condition, we used mixed-effect logistic regression to predict participants' binary choice (BV = 1, BC = 0) from the amount of evidence (3 pieces of evidence in Experiment 1; 6 pieces of evidence in Experiment 2, and 3), while controlling for the random effects of individual participants. We found a significant effect of the amount of evidence. In the BV condition, when participants were given 6 pieces of counterevidence, they were more likely to choose the *BV response* in the test trials, compared to when they were given 3 pieces of counterevidence (β = 0.83, *SE* = 0.42, *p* = .046).

We next analyzed the combined explanation data of Experiments 2 and 3. In the BV condition, participants were more likely to choose the *BV response* for the principle if they provided "accept evidence" explanations compared to if they provided "pattern" (β = 1.09, *SE* = 0.44, *p* = .01), "explain away" (β = 1.61, *SE* = 0.34, *p* < .001), or "other" explanations (β = 2.01, *SE* = 0.35, *p* < .001) for that principle.

In conclusion, across three experiments, we found consistent evidence that adults can revise their beliefs about all three principles given counterevidence. They were more likely to revise their beliefs given a larger amount of counterevidence. Adults who accepted the counterevidence were more likely to revise their beliefs. Moreover, adults readily generalized their revised beliefs to new geometric-shaped agents and animals. Their prior beliefs might be stronger for the Sampling principle than the Goal principle, which is in turn stronger for the Efficiency principle. I will return to these differences across principles in the general discussion.

3.6. Experiment 4

3.6.1. Methods

Participants

Twenty-four 4- to 6-year-olds (mean age = 5.04; range = 4.08 to 6.92; SD = 0.82; 11 females) participated over Zoom. Children's parents provided written informed consent prior to the experiment, and they received electronic certificates.

Design and Procedure

The design and procedure of Experiment 4 were the same as Experiment 1. An experimenter showed the stimuli videos to children via Zoom, and recorded their verbal responses in the test trials.

3.6.2. Results

The proportion of *BV response* by condition and principle is shown in Figure 3.7. We used mixed-effect logistic regression to predict participants' binary choice (BV = 1, BC = 0) from condition, principle, trial type, trial order, age, and gender, while controlling for the random effects of individual participants. The best-fitting model included the interaction of condition and principle as predictors. For the Goal and Sampling principles, children were more likely to choose the *BV response* in the BV condition than in the BC condition (Goal: $\beta = 2.15$, $SE = 0.57$, $p < .001$; Sampling: $\beta = 2.58$, $SE = 0.63$, $p < .001$). For the Efficiency principle, children were equally likely to choose the *BV response* in the BV condition and the BC condition ($\beta = 0.07$, $SE = 0.53$, $p = .90$).

In the BC condition, children were more likely to choose the *BV response* for the Efficiency principle than the Goal principle ($\beta = 1.33$, $SE = 0.46$, $p = .004$) and the Sampling principle ($\beta = 2.29$, $SE = 0.52$, $p < .001$).

Next, we compared participants' responses against chance. In the BC condition, participants chose the *BV response* at chance for the Efficiency principle (Exact binomial tests: $P_{Efficiency} = .64\ [.49, .77], p = .64$), and below chance for the Goal and the Sampling principles ($P_{Goal} = .34\ [.21, .49], p = .03; P_{Sampling} = .18\ [.08, .31], p < .001$). In the BV condition, participants chose the *BV response* at chance for the Efficiency and the Sampling principles ($P_{Efficiency} = .65\ [.49, .79], p = .05; P_{Sampling} = .54\ [.39, 691], p = .66$), and above chance for the Goal principle ($P_{Goal} = 834\ [.69, .92], p < .001$).

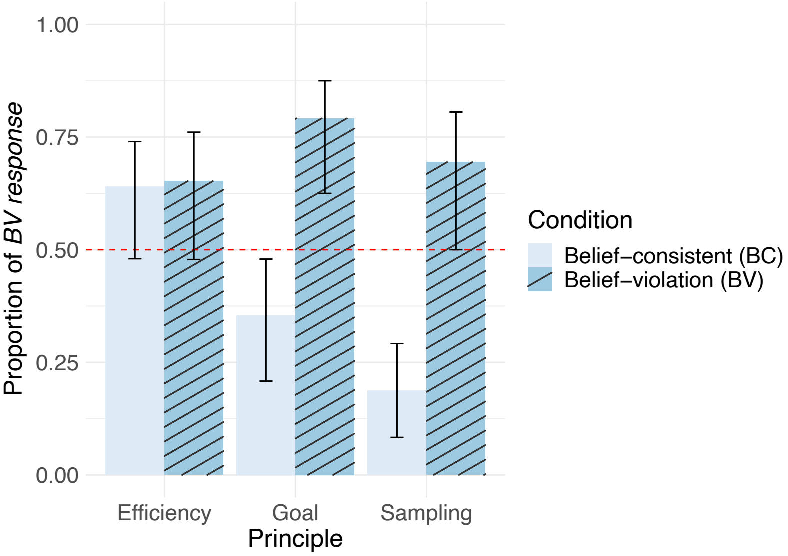

Figure 3. 7: The mean proportion of trials that children selected the *belief-violation (BV) response* by condition and principle in Experiment 4. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

3.6.3. Discussion

The results of Experiment 4 showed that 4- to 6-year-olds can revise their beliefs about the Goal and the Sampling principles given counterevidence. After observing evidence consistent with these 2 principles, children predicted that the outcomes of new events would be consistent with the principles. After observing evidence violating these two principles, they were more likely to predict outcomes inconsistent with these two principles. Children's performance did not differ in the easy and hard test trials, suggesting that they also generalized their revised beliefs to new agents.

For the Efficiency principle, children were equally likely to choose consistent and inconsistent outcomes for new events both after observing evidence consistent with the principle and after observing evidence inconsistent with the principle. Thus, children were less affected by evidence about the Efficiency principle compared to the other two principles.

In the next experiment, we aimed to replicate these findings. In addition, we measured children's prior beliefs about these principles and tested the effect of belief-violating evidence on their prior beliefs. We also increased the strength of the belief-violating evidence and tested its effect on their beliefs. Lastly, we assessed participants' interpretations of the evidence to see if they accepted the counterevidence.

3.7. Experiment 5

3.7.1. Methods

Participants

Thirty-six children between the ages of 4 and 6 years (mean age = 5.12; range = 4 to 6.75; SD = 0.92; 11 females, 23 males, and 2 of unknown gender) participated over Zoom. The sample size was determined based on the effect sizes observed in Experiment 4. The sample size in this experiment provided us with at least 80% power (at $\alpha = .05$) to detect the effect sizes observed in Experiment 4. Children's parents provided written informed consent prior to the experiment, and they received electronic certificates.

Design and Procedure

The design and procedure of Experiment 5 were the same as Experiment 2. An experimenter showed the stimuli videos to children via Zoom, and recorded their verbal responses in the test trials.

3.7.2. Results

The proportion of *BV response* by condition and principle is shown in Figure 3.8. We used logistic regression to predict participants' binary choice (*BV response* = 1, *BC response* = 0) from condition, principle, trial type, age, and gender, with random intercepts for participants. The best-fitting model included the interaction of condition and principle as predictors. For the Goal and Sampling principles, children were more likely to choose the *BV response* in the BV condition than in the Baseline condition (Goal: $\beta = 3.17$, $SE = 0.80$, $p < .001$; Sampling: $\beta = 4.36$, $SE = 0.95$, $p < .001$) and the BC condition (Goal: $\beta = 4.77$, $SE = 0.93$, $p < .001$; Sampling: $\beta = 4.34$, $SE = 0.95$, $p < .001$); and there was no difference between the Baseline and the BC conditions (Goal: $\beta = -1.60$, $SE = 0.90$, $p = .075$; Sampling: $\beta = 0.02$, $SE = 1.08$, $p = .99$). For the Efficiency principle, children were equally likely to choose the *BV response* in the Baseline, BC, and BV conditions (Baseline vs. BC: $\beta = 0.96$, $SE = 0.76$, $p = .20$; Baseline vs. BV: $\beta = 1.37$, $SE = 0.75$, $p = .069$). In the Baseline condition, participants chose the *BV response* below chance for all 3 principles (Exact binomial tests: *p*s < .005). In the BC condition, participants chose the *BV* *response* at chance for the Efficiency principle (Exact binomial tests: $P_{Efficiency} = .08\ [.02, .20], p = .67$), and below chance for the Goal and the Sampling principles ($P_{Goal} = .34\ [.21, .49], p = .03; P_{Sampling} = .06\ [.01, .17], p < .001$). In the BV condition, participants chose the *BV response* at chance for the Efficiency principle (Exact binomial tests: $P_{Efficiency} = .52\ [.37, .67], p = .89$), and above chance for the Goal and the Sampling principles ($P_{Goal} = .79\ [.65, .90], p < .001; P_{Sampling} = .67\ [.52, .80], p = .03$).

In the Baseline condition, compared to the Sampling principle, children were more likely to choose the *BV response* for the Efficiency ($\beta = 2.23$, $SE = 0.75$, $p = .003$) and Goal principles ($\beta = 1.95$, $SE = 0.75$, $p = .010$). In the BC condition, children were more likely to choose the *BV response* for the Efficiency principle than the Goal ($\beta = 2.84$, $SE = 0.69$, $p < .001$) and the Sampling principles ($\beta = 3.18$, $SE = 0.76$, $p < .001$). In the BV condition, children were less likely to choose the *BV response* for the Efficiency principle than the Goal principle ($\beta = -1.53$, $SE = 0.51$, $p = .003$).

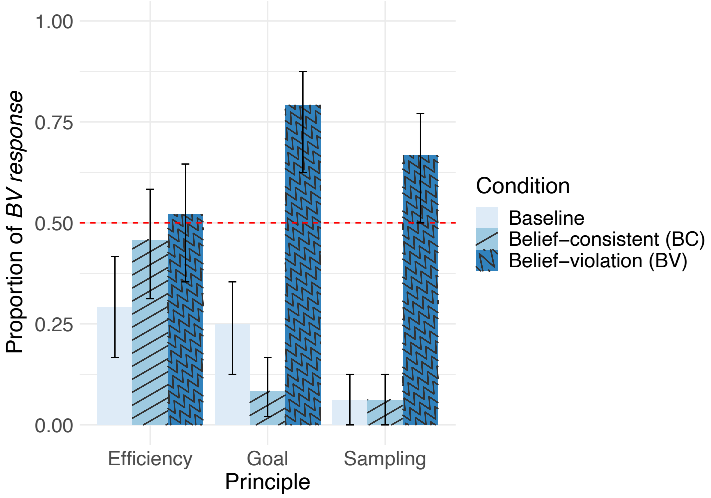

Figure 3. 8: The mean proportion of trials that children selected the *belief-violation (BV) response* by condition and principle in Experiment 5. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

We analyzed the combined explanation data from Experiments 5 and 6 and reported the results in the results section of Experiment 6 below, due to the small sample size in each condition in each experiment.

3.7.3. Discussion

Experiment 5 assessed children's prior beliefs about the Efficiency, Goal, and Sampling principles, and examined whether they can revise their prior beliefs about these principles given counterevidence. We found that children had strong prior beliefs about the Goal and Sampling principles. When they did not receive any new evidence about these principles and when they received evidence supporting these principles, they predicted that the outcomes of new events would be consistent with the principles. Moreover, they can revise their prior beliefs when given counterevidence – after receiving new evidence violating these principles, they predicted that the outcomes of new events would be inconsistent with the principles.

For the Efficiency principle, children also had strong prior beliefs about this principle – they predicted consistent outcomes for new events when given no new evidence. However, we replicated the finding in Experiment 4 that children were less affected by evidence about the Efficiency principle – observing evidence consistent with the principle or inconsistent with the principle did not change their predictions for new events.

We also discovered some important differences across principles. Children had weaker prior beliefs about the Efficiency and Goal principles than the Sampling principle. They were less likely to revise their beliefs about the Efficiency principle than the Goal principle.

In the next experiment, we aim to replicate these findings with more realistic, three-dimensional stimuli, and investigate whether children can generalize their revised beliefs to agents that are not geometric shapes.

3.8. Experiment 6

3.8.1. Methods

Participants

Thirty-six children between the ages of 4 and 6 years (mean age = 4.99; range = 4 to 6.83; SD = 0.87; 22 females and 14 males) participated over Zoom. The sample size was determined based on the effect sizes observed in Experiment 4. The sample size in this experiment provided us with at least 80% power (at $\alpha = .05$) to detect the effect sizes observed in Experiment 4. Children's parents provided written informed consent prior to the experiment, and they received electronic certificates.

Design and Procedure

The design and procedure of Experiment 6 were the same as Experiment 3. An experimenter showed the stimuli videos to children via Zoom, and recorded their verbal responses in the test trials.

3.8.2. Results

The proportion of *BV response* by condition and principle is shown in Figure 3.9. We used logistic regression to predict participants' binary choice (*BV response* = 1, *BC response* = 0) from condition, principle, test trial type, age, and gender, with random intercepts for participants. The best-fitting model included the interaction between condition and principle, and the interaction between condition and age as predictors.

Children were more likely to choose the *BV response* in the BV condition than in the Baseline condition (Goal: $\beta = 3.58$, $SE = 0.87$, $p < .001$; Sampling: $\beta = 1.55$, $SE = 0.54$, $p = .004$; marginally significant for Efficiency: $\beta = 0.85$, $SE = 0.54$, $p = .11$) and the BC condition (Goal: $\beta = 4.84$, $SE = 0.89$, $p < .001$; Sampling: $\beta = 1.75$, $SE = 0.55$, $p = .002$; Efficiency: $\beta = 1.98$, $SE = 0.56$, $p < .001$); they were less likely to choose the *BV response* in the BC condition than in the Baseline condition for the Goal and the Efficiency principles (Goal: $\beta = -1.26$, $SE = 0.55$, $p = .02$; Efficiency: $\beta = -1.14$, $SE = 0.55$, $p = .04$), and their responses did not differ in the BC and the Baseline conditions for the Sampling principle ($\beta = -0.20$, $SE = 0.55$, $p = .71$). In the Baseline condition, participants chose the *BV response* at chance for the Efficiency and the Goal principles (Exact binomial tests: $P_{Efficiency} = .56$ *[.43, .67], p = .41*; $P_{Goal} = .60$ *[.47, .71], p = .12*), and below chance for the Sampling principle *(*$P_{Sampling} = .36$ *[.25, .48], p = .02*). In the BC condition, participants chose the *BV response* below chance for all 3 principles (*p*s < .05). In the BV condition, participants chose the *BV response* above chance for all 3 principles (*p*s < .006).

In the Baseline condition, compared to the Sampling principle, children were more likely to choose the *BV response* for the Efficiency (β = 0.93, *SE* = 0.37, *p* = .01) and the Goal principle (β = 1.13, *SE* = 0.37, *p* = .002). In the BV condition, children were more likely to choose the *BV response* for the Goal principle than for the Efficiency (β = 2.94, *SE* = 0.79, *p* < .001) and the Sampling principle (β = 3.17, *SE* = 0.78, *p* < .001).

The interaction of condition and age showed that, in the BC condition, children were less likely to choose the *BV response* with increasing age (β = -1.12, *SE* = 0.35, *p* = .001).

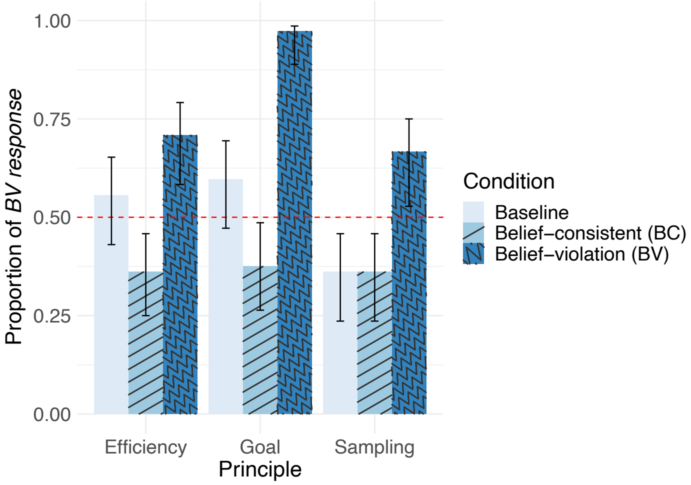

Figure 3. 9: The mean proportion of trials that children selected the *belief-violation (BV) response* by condition and principle in Experiment 6. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

For the combined children's explanation data in Experiments 5 and 6, 2 researchers coded participants' responses into different categories (the interrater reliability was good, Cohen's Kappa = .81; disagreements were resolved through discussion). In the Baseline and the BC condition, most responses (66.7% in the Baseline condition and 73.3% in the BC condition) referred to the principle itself to explain the evidence (the other responses were irrelevant to the principle or were incomprehensible). In the BV condition, we categorized participants' explanations into four categories. The criteria for categorization and examples are shown in Table 3.1 and Table 3.2.

Table 3.5 shows the number of explanations coded within each category for each principle. We used mixed-effects multinomial logistic regression to predict participants' explanations from principle, while controlling for the random effects of individual participants. There was no significant effect of principle, suggesting that the distribution of explanations did not differ across principles.

Next, we used mixed-effects logistic regression to predict participants' binary choice in the test trials (BV response = 1, BC response = 0) from the type of explanation they provided, while controlling for the random effects of individual participants. We found that children who provided "accept evidence" explanations were more likely to choose the *BV response* for the principle compared to those who provided "other" explanations (β = 0.95, *SE* = 0.34, *p* = .01). Those who provided "explain away" explanations were more likely to choose the *BV response* for the principle compared to those who provided "accept evidence" explanations (β = 0.89, *SE* =

0.43, p = .04) and “other” explanations (β = 1.84, *SE* = 0.43, p < .001). Table 3.6 shows the proportion of *BV responses* by participants who provided different types of explanations.

Table 3.5: Children’s Explanations by category and principle in Experiments 5 & 6

	Efficiency	Goal	Sampling
Accept Evidence	12	10	10
Explain Away	4	9	9
Pattern	0	0	0
Other	8	5	5

Table 3.6: Proportion of *BV response* by Experiment, principle, trial type, and explanation type

Experiment	Principle	Trial Type	Explanation Type			
			Accept Evidence	Explain Away	Pattern	Other
Experiment 2	Efficiency	easy	0.91	0.75	1.00	1.00
		hard	0.94	1.00	1.00	1.00
	Goal	easy	0.97	0	1.00	0.33
		hard	0.83	0.50	1.00	0.83
	Sampling	easy	0.67	0.67	0.83	0.33
		hard	0.50	0.50	0.58	0.33
Experiment 3	Efficiency	easy	0.94	0.88	1.00	1.00
		hard	0.91	1.00	0	0.75
		harder	0.75	0.63	1.00	1.00
	Goal	easy	0.74	0	0.50	1.00
		hard	0.84	0	0.75	0
		harder	0.82	0.50	0.50	0.50
	Sampling	easy	0.55	0.50	0.67	0.25
		hard	0.50	0.50	1.00	0.25
		harder	0.50	0.38	1.00	0.25
Experiment 5 & 6	Efficiency	easy	0.50	1.00	N/A	0.68
		hard	0.38	0.75	N/A	0.68
		harder	0.75	1.00	N/A	0.67
	Goal	easy	0.94	0.83	N/A	0.80
		hard	0.94	0.67	N/A	0.80
		harder	0.95	N/A	N/A	1.00
	Sampling	easy	0.76	0.67	N/A	0.38
		hard	0.74	0.50	N/A	0.63
		harder	0.61	0.75	N/A	0

3.8.3. Discussion

In Experiment 6, we found that children had strong prior beliefs about the Sampling principle, but they did not have strong prior beliefs about the Efficiency and Goal principles. However, for all 3 principles, they reliably predicted consistent outcomes for new events after observing evidence consistent with the principles, and predicted inconsistent outcomes for new events after observing evidence violating the principles. For the Goal and Sampling principles, they were more likely to predict inconsistent outcomes given counterevidence than given no new evidence, suggesting that they reliably revised their prior beliefs about these 2 principles.

We also found that children had weaker prior beliefs for the Efficiency and Goal principles than for the Sampling principle. They were more likely to revise their beliefs for the Goal principle than for the other two principles. There was an interesting age effect: given evidence consistent with the principles, children were more likely to predict consistent outcomes for new events with increasing age, suggesting that older children were more affected by the belief-consistent evidence compared to younger children.

Children's explanations of the belief-violating evidence showed that a large group of children (44%) accepted the counterevidence. Other participants who provided "explain away", "pattern" or "other" types of explanations still predicted outcomes that violated the principles in test trials most of the time, except participants who provided "other" explanations for the Sampling principle (Table 3.6). In addition, they were more likely to accept the belief-violating evidence for the Goal and Sampling principles than the Efficiency principle, suggesting that for children, the counterevidence of the Efficiency principle might be less compelling than the counterevidence of the other 2 principles. Importantly, across principles, children who had accepted the counterevidence were indeed more likely to predict outcomes that violated the principles than those who provided irrelevant explanations. However, children's explanation data were noisier – those who explained away the counterevidence were also more likely to choose the *BV response* for the principle compared to those who accepted the counterevidence or provided irrelevant explanations.

3.9. Experiments 4—6: Combined results and discussion

Next, we analyzed the combined results of Experiments 4—6. We used mixed-effect logistic regression to predict participants' binary choice (BV = 1, BC = 0) from condition, principle, trial type, experiment, age, gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included the three-way interaction of condition, principle, and experiment, and the three-way interaction of condition, principle, and age as predictors.

Most importantly, for the Goal principle, children were more likely to choose the *BV response* in the BV condition than in the Baseline ($\beta = 3.05$, $SE = 0.71$, $p < .001$) and the BC ($\beta = 5.40$, $SE = 0.93$, $p < .001$) conditions, and less likely to choose the *BV response* in the BC condition than in the Baseline condition ($\beta = -2.35$, $SE = 0.92$, $p = .01$). For the Sampling principle, children were more likely to choose the *BV response* in the BV condition than in the Baseline ($\beta = 4.40$, $SE = 0.88$, $p < .001$) and the BC ($\beta = 4.35$, $SE = 0.88$, $p < .001$) conditions, and they were equally likely to choose the *BV response* in the BC and the Baseline conditions ($\beta = -.05$, $SE = 1.05$, $p = 0.96$). For the Efficiency principle, children were more likely to choose the *BV response* in the BV condition than in the Baseline condition ($\beta = 1.29$, $SE = 0.64$, $p = .045$); their responses did not differ between the BC and the BV conditions ($\beta = 0.38$, $SE = 0.63$, $p = .55$), or between the Baseline and the BC conditions ($\beta = 0.91$, $SE = 0.65$, $p = .16$). In the

Baseline condition, children chose the *BV response* below chance for the Sampling principle ($p < .001$), and at chance for the Efficiency ($p = .32$) and Goal ($p = .41$) principles. In the BC condition, children chose the *BV response* below chance for the Sampling and the Goal principles (*p*s < .001), and at chance for the Efficiency ($p = .49$). In the BV condition, participants chose the *BV response* above chance for all three principles (*p*s < .001).

The three-way interaction of condition, principle, and experiment showed that, in the Baseline condition, children were more likely to choose the *BV response* for the Goal and the Efficiency principles than for the Sampling principle in Experiment 5 (Goal: $\beta = 2.08$, $SE = 0.76$, $p = .006$; Efficiency: $\beta = 2.38$, $SE = 0.75$, $p = .002$) and Experiment 6 (Goal: $\beta = 1.17$, $SE = 0.38$, $p = .002$; Efficiency: $\beta = 0.97$, $SE = 0.39$, $p = .01$), and more likely to choose the *BV response* for the Efficiency principles than for the Sampling principle in Experiment 4 ($\beta = 2.86$, $SE = 1.02$, $p = .005$). This suggests that children had stronger prior beliefs for the Sampling principle than the other 2 principles. In the BC condition, children were more likely to choose the *BV response* for Efficiency principle than for the other 2 principles in Experiment 4 (Goal: $\beta = 2.12$, $SE = 0.55$, $p < .001$; Sampling: $\beta = 2.77$, $SE = 0.56$, $p < .001$), and in Experiment 5 (Goal: $\beta = 3.56$, $SE = 0.79$, $p < .001$; Sampling: $\beta = 3.24$, $SE = 0.76$, $p < .001$). Thus, children's beliefs about the Efficiency principle were affected less by the belief-consistent evidence compared to the other two principles. In the BV condition in Experiment 5, children were more likely to choose the *BV response* for the Goal principles than for the Efficiency principle ($\beta = 1.46$, $SE = 0.50$, $p = .003$); in the BV condition in Experiment 6, children were more likely to choose the *BV response* for the Goal principles than for the other 2 principles (Efficiency: $\beta = 2.95$, $SE = 0.79$, $p < .001$; Sampling: $\beta = 3.20$, $SE = 0.79$, $p < .001$). Thus, children were more likely to revise their beliefs about the Goal principle given counterevidence compared to the other principles. However, these differences across principles might be less robust, since these differences were not shown in all experiments.

The three-way interaction of condition, principle, and age showed that children were less likely to choose the *BV response* for the Sampling principle with age in the Baseline ($\beta = -0.66$, $SE = 0.31$, $p = .04$) and the BC conditions ($\beta = -1.24$, $SE = 0.39$, $p = .001$). This suggests that older children might have stronger prior beliefs about the Sampling principle than younger children. Children were less likely to choose the *BV response* for the Goal principle with age in the BC condition ($\beta = -2.36$, $SE = 0.52$, $p < .001$), suggesting that older children's beliefs about the Goal principle were affected more by the belief-consistent evidence than younger children's. There were no other age effects.

There was no significant effect of test trial type (easy vs. hard vs. harder test trials), suggesting that children generalized their revised beliefs to new agents. There was no significant effect of experiment, suggesting that the different stimuli sets did not affect participants' choices in the test trials.

To examine the effect of the amount of evidence on participants' choices, we analyzed the BC condition data across the 3 experiments and the BV condition data across the 3 experiments, respectively. The amount of evidence (3 pieces of evidence vs. 6 pieces of evidence) did not affect children's *BV response* in the BC condition or the BV condition.

In conclusion, across three experiments, we found that children can revise their beliefs about all three principles given counterevidence. Moreover, children generalized their revised beliefs to new geometric-shaped agents and animals. Similar to adults, children might also have stronger prior beliefs for the Sampling principle than for the other two principles. I will return to these differences across principles in the general discussion.

3.10. Experiments 1—6: Combined results and discussion

Next, we analyzed the combined results of all six experiments to compare adults' and children's performances (Figure 3.10). We used mixed-effect logistic regression to predict participants' binary choice (BV = 1, BC = 0) from condition, principle, test trial type, age group (adults vs. children), gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included the three-way interaction of condition, principle, and age group as predictors.

Most importantly, adults were more likely to choose the *BV response* in the BV condition than in the Baseline (Efficiency: $\beta = 4.18$, $SE = 0.43$, $p < .001$; Goal: $\beta = 3.68$, $SE = 0.42$, $p < .001$; Sampling: $\beta = 4.36$ $SE = 0.49$, $p < .001$) and the BC conditions (Efficiency: $\beta = 5.27$, $SE = 0.43$, $p < .001$; Goal: $\beta = 5.30$, $SE = 0.46$, $p < .001$; Sampling: $\beta = 5.03$, $SE = 0.52$, $p < .001$); they were less likely choose the *BV response* in the BC condition than in the Baseline condition (Efficiency: $\beta = -1.21$, $SE = 0.44$, $p = .006$; Goal: $\beta = -1.61$, $SE = 0.47$, $p < .001$; not significant for Sampling: $\beta = -0.71$, $SE = 0.60$, $p = .24$). Children were also more likely to choose the *BV response* in the BV condition than in the Baseline (Efficiency: $\beta = 1.25$, $SE = 0.53$, $p = .019$; Goal: $\beta = 2.80$, $SE = 0.56$, $p < .001$; Sampling: $\beta = 2.66$, $SE = 0.55$, $p < .001$) and the BC (Efficiency: $\beta = 0.96$, $SE = 0.47$, $p = .040$; Goal: $\beta = 3.63$, $SE = 0.51$, $p < .001$; Sampling: $\beta = 2.65$, $SE = 0.49$, $p < .001$) conditions, and their responses did not differ between the Baseline and the BC conditions (Efficiency: $\beta = 0.31$, $SE = 0.53$, $p = .56$; Goal: $\beta = -0.85$, $SE = 0.54$, $p = .12$; Sampling: $\beta = -.001$, $SE = 0.56$, $p = 1.00$).

In the Baseline and the BC conditions, children were more likely than adults to choose the *BV response* for all 3 principles (Baseline: Efficiency: $\beta = 1.41$, $SE = 0.52$, $p = .006$; Goal: $\beta = 1.82$, $SE = 0.52$, $p = .005$; Sampling: $\beta = 2.36$, $SE = 0.60$, $p < .001$; BC: Efficiency: $\beta = 2.83$, $SE = 0.46$, $p < .001$; Goal: $\beta = 2.55$, $SE = 0.50$, $p < .001$; Sampling: $\beta = 3.12$, $SE = 0.57$, $p < .001$). Thus, children have weaker prior beliefs about the principles than adults. In the BV condition, children were less likely than adults to choose the *BV response* for the Efficiency principle ($\beta = -1.51$, $SE = 0.43$, $p < .001$), and more likely than adults to choose the *BV response* for the Goal principle ($\beta = 0.95$, $SE = 0.46$, $p = .037$). Thus, children were less likely than adults to revise their beliefs about the Efficiency principle, but more likely than adults to revise their beliefs about the Goal principle.

There was no significant effect of test trial type (easy vs. hard vs. harder test trials), suggesting that children and adults generalized their revised beliefs to new agents.

To examine the effect of the amount of evidence on participants' choices, we analyzed the BC condition data across the 6 experiments and the BV condition data across the 6 experiments, respectively. The amount of evidence did not affect participants' *BV response* in the BC condition or the BV condition.

We next analyzed the combined explanation data of adults and children. Participants were more likely to choose the *BV response* if they provided "accept evidence" explanations for that principle, compared to if they provided "explain away" ($\beta = 0.64$, $SE = 0.25$, $p = .01$), or "other" explanations ($\beta = 1.62$, $SE = 0.24$, $p < .001$).

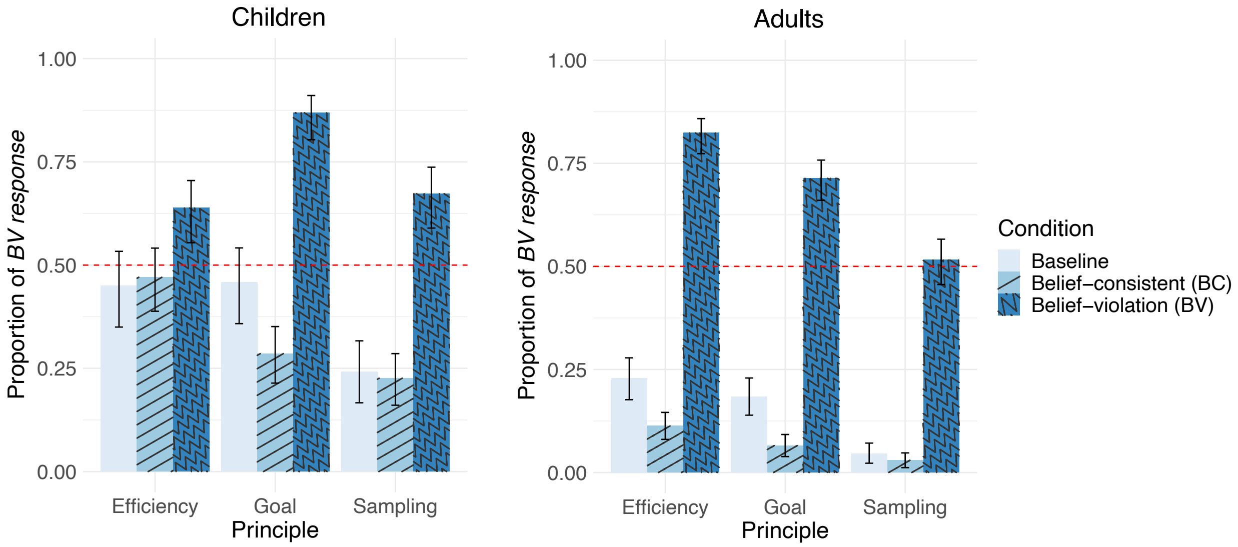

Figure 3. 10: The mean proportion of trials that children (left) and adults (right) selected the *belief-violation (BV) response* by condition and principle in Experiments 1—6. The dashed line indicates chance performance (.5), and the error bars indicate bootstrapped 95% CIs.

In conclusion, across six experiments, we found consistent evidence that both adults and children can revise their beliefs about the Efficiency, Goal, and Sampling principles given counterevidence. Children had weaker prior beliefs about all three principles than adults. Children were less likely to revise their beliefs about the Efficiency principle than adults, but they were more likely to revise their beliefs about the Goal principle than adults. Moreover, adults and children generalized their revised beliefs to new geometric-shaped agents and animals, and they did not differ in the extent to which they were willing to generalize the revised beliefs to different types of agents. Adults' and children's explanations of the belief-violating evidence showed that they were more likely to revise their beliefs if they accepted the counterevidence than if they explained away the counterevidence or provided irrelevant explanations.

3.11. Comparison across the object domain (Chapter 2) and the agent domain (Chapter 3): Combined results and discussion

Lastly, we analyzed the combined data of Chapter 2 and Chapter 3, since the two chapters used similar methods to investigate whether adults and children revise their beliefs about the core principles in the object domain vs. the agent domain.

We used mixed-effect logistic regression to predict participants' binary choice (BV = 1, BC = 0) from domain, condition, test trial type, age group (adults vs. children), gender, and their interactions, while controlling for the random effects of individual participants. The best-fitting model included the interaction of condition and test trial type, the interaction of condition and domain, the interaction of condition and age group, and the interaction of domain and age group as predictors.

Most importantly, both adults and children were more likely to choose the *BV response* in the BV condition than in the Baseline condition in the physical domain (adults: $\beta = 5.28$, $SE = 0.35$, $p < .001$; children: $\beta = 2.96$, $SE = 0.39$, $p < .001$) and the psychological domain (adults: $\beta = 3.97$, $SE = 0.33$, $p < .001$; children: $\beta = 1.65$, $SE = 0.38$, $p < .001$); they were more likely to choose the *BV response* in the BV condition than in the BC condition in the physical domain (adults: *β* = 6.04, *SE* = 0.33, *p* < .001; children: *β* = 2.47, *SE* = 0.35, *p* < .001) and the psychological domain (adults: *β* = 4.97, *SE* = 0.33, *p* < .001; children: *β* = 3.68, *SE* = 0.42, *p* < .001); adults were less likely to choose the *BV response* in the BC condition than in the Baseline condition in the physical domain (*β* = -0.76, *SE* = 0.36, *p* = .03) and the psychological domain (*β* = -1.00, *SE* = 0.35, *p* = .004); children were equally likely to choose the *BV response* in the BC and the Baseline conditions in the physical domain (*β* = -0.58, *SE* = 0.40, *p* = .15), and less likely to choose the *BV response* in the BC condition than in the Baseline condition in the psychological domain (*β* = -0.82, *SE* = 0.39, *p* = .04).

The interaction of condition and test trial type showed that in the BV condition, participants were less likely to choose the *BV response* in harder test trials than in easy test trials (*β* = -0.64, *SE* = 0.15, *p* < .001) and hard test trials (*β* = -0.45, *SE* = 0.17, *p* = .002). Participants' choices did not differ across test trial types in the Baseline and the BC conditions. Thus, participants were less likely to generalize their revised beliefs to completely different contexts.

The interaction of condition and domain showed that participants were more likely to choose the *BV response* in the psychological domain than in the physical domain in both the Baseline (*β* = 1.27, *SE* = 0.36, *p* < .001) and the BC conditions (*β* = 1.03, *SE* = 0.33, *p* = .002). This suggests that adults and children had weaker prior beliefs for the psychological principles than the physical principles.

The interaction of condition and age group showed that children were more likely than adults to choose the *BV response* in the Baseline (*β* = 1.54, *SE* = 0.36, *p* < .001) and the BC conditions (*β* = 1.72, *SE* = 0.33, *p* < .001), but they were less likely than adults to choose the *BV response* in the BV condition (*β* = -0.78, *SE* = 0.30, *p* = .008). Thus, children had weaker prior beliefs about the physical and the psychological principles than adults, and they were less likely to revise their beliefs about these principles given counterevidence.

The interaction of domain and age group showed that children were more likely than adults to choose the *BV response* in both the physical (*β* = 1.54, *SE* = 0.36, *p* < .001) and the psychological domains (*β* = 2.23, *SE* = 0.35, *p* < .001), and the difference between age groups was larger in the psychological domain than in the physical domain (*β* = 0.69, *SE* = 0.32, *p* = .03). This suggests that children might have weaker prior beliefs than adults in both domains, and the difference in the strengths of their beliefs is larger in the psychological domain.

Next, we compared the explanation data across the two domains. We used mixed-effects multinomial logistic regression to predict participants' explanations from domain, while controlling for the random effects of individual participants. We found a significant effect of domain. Participants were more likely to provide "accept evidence" explanations for agents than "explain away" explanations (*β* = 1.48, *SE* = 0.28, *p* < .001) and "pattern" explanations (marginally significant: *β* = 0.79, *SE* = 0.41, *p* = .052), compared to objects, and they were less likely to provide "explain away" explanations for agents than "other" explanations, compared to objects (*β* = -0.95, *SE* = 0.34, *p* = .005).

In conclusion, the comparison across domain showed that adults and children have weaker prior beliefs for the psychological principle than the physical principles. Adults and children were more likely to accept the belief-violating evidence for the psychological principle than the physical principles, and they were more likely to explain away the belief-violating evidence for the physical principles than for the psychological principles. We will return to these domain differences in the general discussion. In addition, children have weaker prior beliefs in both domains than adults, and the difference in the strength of their prior beliefs is larger in the agent domain than in the object domain. Children were also less likely to revise their beliefs than adults in both domains. We will return to these age differences in the general discussion.

3.12. General Discussion

In six experiments, we used a novel paradigm to measure adults' and children's prior beliefs about three core knowledge principles in the domain of agents – Efficiency, Goal, and Sampling – and we investigated whether they could revise these most fundamental beliefs about agents in a virtual environment. We found that adults have strong prior beliefs about these three principles. When they were not given any new evidence about these principles and when they observed a few events consistent with these principles, they predicted that agents would behave in accordance with these principles. For children, we found that they had strong prior beliefs about the Sampling principle, but they did not have strong prior beliefs about the Efficiency and Goal principles. For the Sampling principle, when they were not given any new evidence and when they observed a few events consistent with the principle, children reliably predicted that agents would behave in accordance with the principle. For the Efficiency principle, children were equally likely to predict that agents would behave consistently or inconsistently with the principle, both when given no new evidence and when given a few pieces of evidence consistent with the principle. For the Goal principle, children were equally likely to predict that agents would behave consistently or inconsistently with this principle when given no new evidence, and they predicted that agents would behave in accordance with the principle given a few pieces of evidence consistent with the principle. Importantly, we found that both adults and children revised their beliefs about all 3 principles given just a few pieces of counterevidence. After observing a few events of agents violating these principles, children and adults predicted that agents were more likely to behave inconsistently with the principles, and they generalized their inconsistent predictions to new geometric-shaped agents and animals in this environment.

We had 2 hypotheses for whether children and adults would revise their beliefs given counterevidence about core principles about agents. Our first hypothesis was that children and adults would try to come up with alternative interpretations to explain away the counterevidence, and not revise their beliefs about these principles. Our second hypothesis was that children will accept the counterevidence and revise their beliefs, and they will generalize their revised beliefs to a certain extent (narrowly or widely). We will discuss the explanation data and the generalization trials to further examine which hypothesis is supported by the present findings.

The explanation data suggest that most adults (61% on average) and children (44% on average) stated that they had accepted the counterevidence. These participants were indeed more likely to predict outcomes inconsistent with the principles than participants who explained away the counterevidence or provided other irrelevant responses. Thus, about half of the participants genuinely accepted the counterevidence and revised the principles in this specific, virtual environment, providing support for our first hypothesis. A small group of adults (11% on average) noticed the statistical pattern that agents in this environment behave in ways that were inconsistent with the principles, and therefore predicted that other agents would behave according to this pattern. These adults did not predict outcomes inconsistent with the principles more than adults who provided other types of explanations, suggesting that they probably did not genuinely revise their beliefs. A small group of adults (14% on average) and children (31% on average) explained away the counterevidence with reasons that did not involve any violations of the principles, so that they did not have to revise the principles. This group of participants lends partial support to our first hypothesis. Lastly, a small group of adults (14% on average) and children (25% on average) said “I don’t know” or referred to irrelevant aspects of the events when they were asked to explain the counterevidence. These participants were also less likely to predict outcomes inconsistent with the principles for new events, suggesting they were less likely to have revised their beliefs.

We also found that adults and children were willing to generalize their revised beliefs to new geometric-shaped agents and animals in this virtual environment. Moreover, they generalize the revised beliefs to different types of agents to the same extent. These findings further supported our second hypothesis, and showed that learners were willing to generalize their revised beliefs widely, even to agents who were completely different from the agents they observed in the belief-violating evidence. Why would they expect new agents to also violate the principles, even though they had not observed any behaviors of the new agents? One possibility is that participants interpreted the behaviors that violated the principles as “norms” in this virtual environment. For example, after observing 2 agents jump to get to their goals when there is no obstacle, participants might think that the norm in this virtual environment is to jump to get to goals. Future studies can examine this possibility by asking participants to explain why they expected the new agents to violate the principles. It would also be interesting to examine whether participants would generalize their revised beliefs to new agents entering this world and agents in a completely different virtual environment. One limitation of the present study is that we used the same type of events in all the belief-violating evidence and the test trials. In future work, we can provide learners with counterevidence in more diverse types of events and test their generalization in different types of events to see if learners will generalize the revised beliefs to broader contexts. For example, would they expect an agent who always jumps to also take a detour when walking on the ground?

We also discovered some interesting differences across principles. Both adults and children had weaker prior beliefs for the Efficiency and Goal principles than the Sampling principle. One possible reason is that adults and children have observed more violations of the Efficiency and Goal principles in the real world. For instance, people do not always take the shortest path to achieve their goals – I might take a detour when I walk to the grocery store because I need exercise. It is also common for people to change their goals – after eating three chocolates, they might want to have some pretzels. Future studies could examine whether there are indeed more violations of the Efficiency and Goal principles than violations of the Sampling principle in real life by asking people to come up with exceptions for each principle. Another possible reason that adults and children have stronger prior beliefs for the Sampling principle is that violation of random sampling is a really strong indication of preference. Within the Naïve Utility Calculus framework, when two types of objects are randomly distributed in space, reaching the rarer objects requires more effort. Thus, only if the rarer objects are more rewarding to an agent, the agent should selectively draw those objects at a higher cost (Jara-Ettinger et al., 2016). In contrast, when an agent repeatedly chooses one object from two objects that are equidistant from her (the event used in the Goal principle), the agent is not incurring a greater cost to choose the preferred object. Thus, adults and children might have stronger prior beliefs about the Sampling principle because violation of random sampling is a stronger indication of preference.

The current study makes several important contributions to the study of reasoning about agents and cognitive development more generally. Past research has shown that when learners observe an agent violate core knowledge principles, they learn something special about the particular agents who violated the principles (e.g., their knowledgeability and third-party preferences for these agents) (Colomer et al., 2020; Colomer & Woodward, 2023). The current study focused on a different question: if learners observe multiple violations of the core knowledge principles, can they revise their higher-level beliefs about the abstract principles governing agent reasoning? In our study, most adults and children accepted the violations and genuinely revised their beliefs about these principles, and they generalized their revised beliefs to new geometric-shaped agents and animals. Thus, adults and children revised the abstract principles governing agent reasoning given multiple pieces of counterevidence. These findings also provide another strong piece of evidence that children and adults have powerful statistical learning mechanisms (e.g., Griffiths et al., 2011; Kimura & Gopnik, 2019; Kushnir & Gopnik, 2007; Lucas & Griffiths, 2010; Lucas et al., 2014), and even our most strongly held beliefs about agents are subject to revision given counterevidence.

We also compared the data from Chapter 2 and Chapter 3 and discovered some important differences across the domains of objects and agents. First, adults and children had strong prior beliefs about the principles governing object reasoning. Adults also had strong prior beliefs about the psychological principles governing agent reasoning, however, children did not have strong prior beliefs about two of the three principles – the Efficiency and the Goal principles. In addition, both adults and children have stronger prior beliefs for the physical principles than the psychological principles. Past research has also shown that 1-year-old infants and children older than 4-year-olds expect agents to take efficient paths to achieve their goals, but 3-year-olds fail to show this expectation (Gergely & Csibra, 2003; Gönül & Paulus, 2021). Thus, the current findings and the past findings might suggest that in the agent system, the development of the principles governing agent reasoning might be discontinuous, contradicting the argument of the Core Knowledge view (Spelke, 2022).

Second, compared with the psychological principles, the physical principles are harder to revise given the same amount of counterevidence in two ways: learners are less likely to generalize the new principles to new objects and new events; they are also less likely to accept the counterevidence and more likely to try to explain it away (e.g., "there is a gap between the wall and the screen so the ball can go through"). In contrast, learners readily generalize the new principles to new agents in the psychological domain, and they accept the counterevidence and generate plausible reasons for the agent's unusual behavior (e.g., "the red child just likes to jump", instead of taking the most efficient path to reach her goal). What explains these domain differences? One possibility is that infants are born with stronger prior beliefs about objects (i.e., the object principles are more hard-wired to begin with); another possibility is that children and adults have observed more counterevidence about agents in everyday life, and therefore have weaker and more flexible beliefs about them.

Taken together, Chapters 2 and 3 show that adults and 4- to 6-year-olds have powerful statistical learning mechanisms, and they can revise their most fundamental beliefs about objects and agents when provided with small amounts of statistical evidence.

Chapter 4: Statistically Representative Counterevidence Changes Children's Intergroup Biases

4.1. Introduction

Another core knowledge system guides how we reason about people as social beings who interact with other individuals. Unlike the object and agent systems which have been extensively studied, the system of social beings is more recently identified, and there is no strong consensus on the set of core principles that accompany this system (Spelke, 2022; Spelke & Kinzler, 2007). A candidate for the core principles in the social being system is our tendency to categorize ourselves and others into social groups and prefer members of our own social groups (the ingroup) over members of other social groups (the outgroup) (Spelke & Kinzler, 2007).

Ingroup biases emerge early in development and underlie our later-developing stereotypes and prejudice toward social groups. Infants prefer others who share trivial similarities with themselves (Mahajan & Wynn, 2012), and they expect individuals who share similarities to show ingroup preferences toward each other (Bian et al., 2022). Infants preferentially look at or engage with individuals based on their race (Kelly et al., 2005; Bar-Haim et al., 2006), gender (Quinn et al., 2002), and the languages they speak (Kinzler et al., 2007). By 3-5 years of age, children show explicit and implicit preferences for their own gender (Dunham et al., 2016; Yee & Brown, 1994), and North American White children show an explicit preference for their own race, as well as an implicit pro-White/anti-Black bias (Dunham et al., 2006). By 5-6 years of age, children have also formed specific stereotypes about gender groups (e.g., girls are gentle, and boys are adventurous) (Halim & Ruble, 2010), and North American White children are more likely to attribute positive traits (e.g., nice, smart, clean) to Whites and negative traits (e.g., mean, stupid, dirty) to Blacks (Augoustinos & Rosewarne, 2001). In addition to attributing specific traits to social groups, children also believe that these traits are inherited and stable, and that members of social groups are homogeneous (i.e., essentialist beliefs, Gelman, 2004). Holding essentialist beliefs about social groups leads to negative attitudes, stereotyping, and prejudice toward the outgroup (Rhodes & Mandalaywala, 2017). By early to middle childhood, children across the globe hold essentialist beliefs about culturally salient categories (e.g., race in the U.S., ethnicity in Israel; Hirschfeld, 1995; Diesendruck & haLevi, 2006).

How do these biases and stereotypes develop in childhood? The Developmental Intergroup Theory (DIT, Bigler & Liben, 2006) proposes that stereotyping and prejudice emerge as children actively construct their understanding of the social categories that are salient in their social environment. The formation of stereotypes and prejudice are affected by both externally driven processes (e.g., observing the covariation between social categories and attributes in the environment) and internally driven processes, including ingroup bias and essentialist beliefs. A crucial question then is how ingroup bias and essentialist beliefs develop in children. This question can be analyzed within the Rational Constructivist framework (Xu, 2019). Under this framework, learning and belief formation depend on both prior knowledge or biases and statistical information from environmental input. The basic forms of both ingroup bias and essentialist beliefs emerge in infancy: Infants prefer their ingroup members over outgroup members (Mahajan & Wynn, 2012), and they expect members of social groups to behave similarly (Powell & Spelke, 2013). These early forms of ingroup bias and essentialist beliefs constitute their prior knowledge or bias. The statistical information from children's social environment often reinforces these intergroup attitudes and essentialist beliefs. Children learn from observing their parents' and teachers' attitudes and behaviors toward different social groups (Sinclair et al., 2005; Vezzali et al., 2012). Children also learn from linguistic inputs. For instance, generic language, which attributes traits to the category in general (e.g., "Girls do not like math"), leads to essentialist beliefs about these categories (Rhodes, et al., 2012).

Thus, children's initial bias and early environmental input work in tandem, allowing children to develop strong intergroup attitudes that are hard to revise. However, an important aspect of the Rational Constructivist framework is that *re-learning and belief revision is always possible given the right kind of counterevidence, even when we have strong prior beliefs and biases*. Would the same powerful statistical learning mechanisms that allowed children and adults to revise their fundamental beliefs about objects and agents also allow children to change their biases about social groups?

A prevalent method that past studies used to change intergroup biases is exposure to exemplars that are inconsistent with prior biases. For example, showing non-Black adults vivid, second-person stories with counterstereotypic exemplars (e.g., a White man assaults the participant, and a Black man rescues the participant) had small to medium effects in reducing adults' implicit pro-White/anti-Black biases. However, this method did not have any effect on changing adults' explicit racial bias, even though participants in the studies started with moderately strong explicit preferences for White individuals over Black individuals (Lai et al., 2014). In a study with White and Asian 5- to 12-year-olds, reading stories about 4 Black adult exemplars accompanied by positive facts (e.g., the Black individual is an excellent firefighter) had a moderate effect on reducing 9- to 12-year-olds' implicit pro-White biases. But this method was ineffective for 5- to 8-year-olds (Gonzalez et al., 2017). In another study, reading stories about countersterotypic child exemplars (e.g., two Black children engaging in 3 prosocial behaviors and 2 White children engaging in 3 antisocial behaviors) had small effects on changing 9- to 12-year-olds' implicit pro-White biases. However, this method showed mixed results with younger children (5- to 8-year-olds) (Gonzalez et al., 2021). Another study examined the effect of counterstereotypic exemplars on the implicit math = male stereotype. The results showed that reading 4 counterstereotypic exemplars (e.g., a female character engaging in and preferring math-related activities in childhood and then growing up to be a math professor) had small effects on reducing the implicit gender bias in 6- to 11-year-olds (Block et al., 2022). Thus, past research showed mixed findings regarding the effectiveness of exposure to counterstereotypic exemplars in reducing intergroup biases in adults and children.

One reason that disconfirming exemplars might fail to reduce bias is that children's processing of new information is still filtered by their preexisting biases (Bigler & Liben, 2006). For instance, children prefer to hear positive information about ingroups and negative information about outgroups than vice-versa (Over et al., 2018). As another example, showing children mean outgroup members decreased their liking of the outgroup, but showing them nice outgroup members did not increase their liking of the outgroup (Schug et al., 2013). Another reason that disconfirming exemplars might be ineffective is because of a process called subtyping (Richards & Hewstone, 2001; Hayes et al., 2003). Disconfirming exemplars can be mentally clustered into a subtype, allowing the exemplars to be seen as exceptions and therefore not representative of the entire group.

Thus, changing children's biases and stereotypes might require more exemplars that are shown to be representative of the group, in order to prevent subtyping. In the present chapter, we assess whether exposing children to counterevidence that is *statistically representative* of the entire social group, might change their attitudes and beliefs about the group. Specifically, we showed children a randomly drawn sample from the group, with information about the distribution of nice vs. mean traits in this sample, and examined whether the trait distribution in the sample can change children's attitudes and beliefs about the group as a whole. Infants and children are sensitive to statistical information, and they understand that a randomly drawn sample is representative of the group (see Denison & Xu, 2019, for a review). Thus, when children observe a randomly drawn, mostly nice sample from the outgroup, it is unlikely that they will discount the sample as an exception, and more likely to take it into account in forming more positive attitudes and beliefs about the outgroup. Given the strong sensitivity to statistical information even in infants, we tested whether this paradigm would be effective in changing young children's (5- to 6-year-olds') attitudes and beliefs about social groups.

We conducted 2 experiments to examine this question. In Experiment 1, to control for any prior biases children might have about particular social groups, we examined whether statistically representative counterevidence can change 5- to 6-year-olds' attitudes and beliefs about minimal groups. Children show the same forms of ingroup biases for real social groups and minimally defined social groups (Dunham, 2018), although their biases are weaker for minimal groups than for real groups (Mullen et al., 1992). In Experiment 2, we assessed this paradigm with real social groups – we examined whether statistically representative counterevidence can change children's attitudes and beliefs about racial groups. Originally, we planned to test both White and Black children. However, we were only able to recruit 2 Black children during the past six months for this experiment in Berkeley. Thus, we only report the results from White children in Experiment 2. We plan to continue putting more effort into recruiting Black children in this experiment, using other recruitment methods in a larger geographic region (e.g., recruiting at museums and schools in Oakland, where there is a larger Black population).

We hypothesized that a priori, children would show an ingroup bias – they would show more positive attitudes and beliefs toward the ingroup than the outgroup. Children's ingroup biases would be stronger for racial groups than for minimal groups. Crucially, their attitudes and beliefs would be changed by the trait distribution of the sample they observe. Their attitudes and beliefs toward both the ingroup and the outgroup would become more positive after observing a mostly nice sample, and more negative after observing a mostly mean sample. We further hypothesized that children might process the information in a biased way, such that the mostly nice sample would have a larger positive effect on children's attitudes and beliefs about the ingroup than the outgroup, and the mostly mean sample would have a larger negative effect on attitudes and beliefs about the outgroup than the ingroup. Their processing of the information about racial groups would be more biased than their processing of the information about minimal groups.

4.2. Experiment 1

4.2.1. Methods

Participants

One hundred and seventy-one 5- to 6-year-olds (93 females; mean age = 5.95; range = 5.00 to 6.97; *SD* = 0.60) participated in the experiment. The sample size was determined based on typical sample sizes and the effect sizes reported in similar published studies (e.g., Bigler et al., 1997; Dunham et al. 2011). Our sample size provided us with at least 90% power (at $\alpha = .05$) to detect the effect sizes observed in a similar past study (Baron & Dunham, 2015). Participants were tested in a lab room or at children's museums. Parents of the participants provided written informed consent prior to the experiment session.

Design and Procedure

The study employed a 2 (Group condition: Ingroup vs. Outgroup) × 2 (Trait Distribution condition: Majority nice vs. Majority mean) between-subject design. We used a between-subject design to avoid carry-over effect and to prevent the procedure from being too long. A visual schematic of the procedure is shown in Figure 4.1.

Room Introduction. Participants were shown two social groups – in two rooms on the computer screen – each filled with pictures of 50 children. Children in one room all wore yellow shirts, and children in the other room all wore blue shirts. Participants were told that some children in the rooms were nice, and some children were mean, and they could find out whether a child was nice or mean when they turned around the picture and saw the expression on the child's face (smiling or angry, respectively).

Room Assignment. Participants were shown 2 cups on the screen. They were told that a blue coin was hidden in one cup and a yellow coin was hidden in the other cup. The experimenter asked the participant to choose a cup and revealed the coin in the cup. Depending on the color of the coin, the experimenter told the participant, "You belong to the blue/yellow room!" Half of the participants were assigned to the blue room, and half to the yellow room. Then, the experimenter gave participants a blue/yellow hat and a blue/yellow sticker to reinforce their group membership.

Prior Measurements. We measured participants' attitudes and essentialist beliefs about the ingroup and the outgroup, as well as their expectations about the likelihood of drawing a nice child and a smart child from the groups. Participants in the Ingroup condition were asked the following questions about the room they were assigned to, and participants in the Outgroup condition were asked the following questions about the room they were not assigned to.

Attitudes. The experimenter showed participants pictures of 4 gender-matched children from the room corresponding to their group condition (Ingroup or Outgroup). For each child, the experimenter assessed whether participants were willing to interact with the child (e.g., "Would you like to play with this child?"), and to what extent they wanted or did not want to interact with the child (e.g., "Do you sort of want to or really want to?"). Each answer received a score ranging from 1 to 4, with higher scores indicating more positive attitudes (really don't want to = 1, sort of don't want to = 2, sort of want to = 3, really want to = 4). The participant's *prior attitude score* was the average of the scores for the 4 questions.

Essentialism. To measure the homogeneity component of essentialism, the experimenter assessed whether participants believed that members in the room were similar or different in terms of biological properties (e.g., "Do you think all the children in the blue room have the same kind of blood or different kinds of blood?") and psychological properties (e.g., "Do you think all the children in the blue room like the same things?"), and the extent they thought they were similar (e.g., "Do you think their blood are exactly the same or kind of the same?") or different ("Do you think their blood are very different or kind of different?"). For each question, children's responses will receive a score ranging from 1 to 4, with higher scores indicating stronger essentialist beliefs ("very different" –1; "kind of different"– 2; "kind of the same" – 3; "exactly the same" – 4).

Then, the experimenter measured the heritability and stability components of essentialism. The experimenter showed participants pictures of another 2 gender-matched children from this room. Participants were asked questions about the trait “nice” for one child, and questions about the trait “mean” for the other child. The experimenter told participants to imagine that the child had a smiling face or an angry face. The experimenter asked participants whether or not the child was born with the trait, and how sure they were. Each response received a score ranging from 1 to 4, with higher scores indicating stronger essentialist beliefs (“No; I’m sure” – 1; “No; I’m not sure” – 2; “Yes; I’m not sure” – 3; “Yes; I’m sure” – 4). Then, participants were asked whether the child’s trait can change in the future (e.g., “Do you think the nice child can become mean when she grows up?”), and how sure they were. Each response received a score ranging from 1 to 4, with higher scores indicating stronger essentialist beliefs (“Yes; I’m sure” – 1; “Yes; I’m not sure” – 2; “No; I’m not sure” – 3; “No; I’m sure” – 4).

The average of the scores for the above 6 questions will render a *prior essentialism score* ranging from 1 to 4.

Expectation. To assess participants’ expectations about the distribution of nice and mean traits in the rooms, they were shown all the children from the room, and were asked, “If we were to check one child from this room, do you think this child would be nice or mean?”

Over-hypothesis. To assess participants’ expectations about the distribution of a different trait, they were asked, “If we were to check another child from this room, do you think this child would be smart or not smart?”

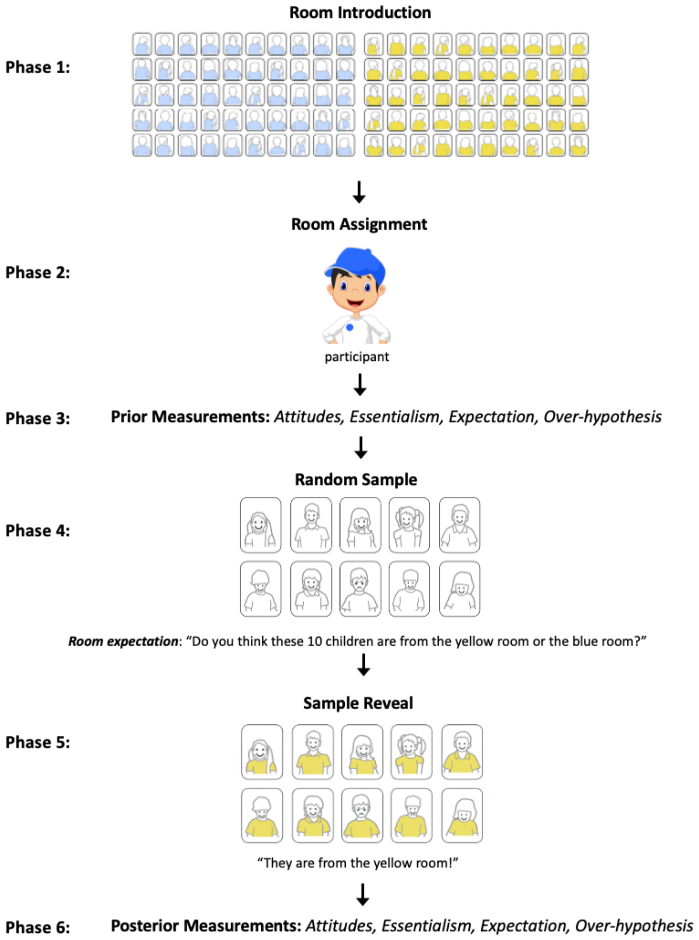
Room Introduction
Phase 1:
Room Assignment
Phase 2:
participant
Phase 3:
Prior Measurements: Attitudes, Essentialism, Expectation, Over-hypothesis
Random Sample
Phase 4:
Room expectation: “Do you think these 10 children are from the yellow room or the blue room?”
Sample Reveal
Phase 5:
“They are from the yellow room!”
Phase 6:
Posterior Measurements: Attitudes, Essentialism, Expectation, Over-hypothesis

Figure 4. 1: A visual schematic of the procedure in Experiment 1.

Random Sample. Next, the experimenter told participants that the computer would randomly pick a sample of 10 children from one of the rooms without telling them from which. Participants saw a picture of the sample of 10 children, with each child showing a smiling or an angry face. In this picture, all children were wearing "white" shirts, denoting that we still did not know from which room they had been drawn. Depending on the participant's Trait Distribution condition, the sample of children was either mostly nice (9 nice and 1 mean children) or mostly mean (1 nice and 9 mean children). The experimenter described the distribution of nice and mean children in the sample, and asked participants to repeat the distribution.

Room Expectation. Participants were asked to guess from which room the sample was drawn.

Sample Reveal. Then, the experimenter revealed from which room the sample was drawn, by showing that the 10 children either wore yellow shirts or blue shirts. Participants in the Ingroup condition observed a sample from the room they were assigned to, and participants in the Outgroup condition observed a sample from the room they were not assigned to.

Posterior Measurements. Finally, participants were asked the attitudes, essentialism, expectation, and over-hypothesis questions again.

4.2.2. Results

Room Expectation

Table 4.1 shows the number of 5- to 6-year-olds who expected the Majority Nice sample and the Majority Mean sample to be from the ingroup or the outgroup. We used logistic regression to predict children's room expectations (ingroup = 1, outgroup = 0) from trait distribution condition, age (z-scored), gender, and their interactions. The best-fitting model included trait distribution condition as the only predictor. Children in the Majority Nice condition were more likely to expect the sample to be from the ingroup rather than from the outgroup, compared to children in the Majority Mean condition ($\beta = 1.07$, $SE = 0.32$, $p < .001$)[1].

Table 4.1: Room expectation results in Experiment 1

Trait distribution condition	Room expectation	
	Ingroup	Outgroup
Majority Nice	52	42
Majority Mean	23	54

Attitudes

The distribution of children's prior and posterior attitude scores is shown in Figure 4.2. We used mixed-effects ANOVAs to predict children's attitude scores from group condition, trait distribution condition, time of measurement, age (z-scored), gender, and their interactions, with random intercepts for participants. The best-fitting model included group condition, and the interaction of trait distribution condition and time of measurement as predictors. Children had more positive attitudes toward the ingroup than toward the outgroup ($\beta = 0.31$, $SE = 0.11$, p

[1] For all subsequent measures, we analyzed the data from all children (including children who did not make the expected predictions in the Room Expectation measure). However, we found similar results if we only analyzed children who made the expected predictions in the Room Expectation measure.

= .005). For both the ingroup and the outgroup, children's attitudes became more negative after observing the mostly mean sample (β = -0.34, *SE* = 0.08, *p* < .001), and became more positive after observing the mostly nice sample, although this trend was marginally significant (β = 0.13, *SE* = 0.07, *p* = .058).

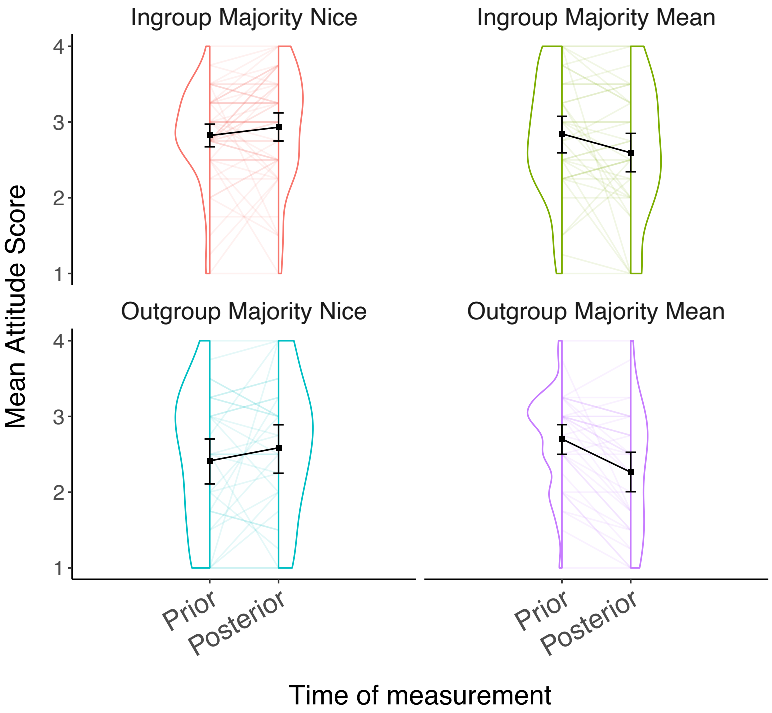

Figure 4. 2: Distribution of children's prior and posterior attitude scores by condition in Experiment 1. The error bars indicate bootstrapped 95% CIs.

Essentialism

The distribution of children's prior and posterior essentialism scores is shown in Figure 4.3. We used mixed-effects ANOVAs to predict children's essentialism scores from group condition, trait distribution condition, time of measurement, age (z-scored), gender, and their interactions, with random intercepts for participants. The best-fitting model included age as the only predictor. Older children had weaker essentialist beliefs than younger children (β = -0.09, *SE* = 0.04, *p* = .04). There were no other effects of group condition, trait distribution condition, time of measurement, or gender.

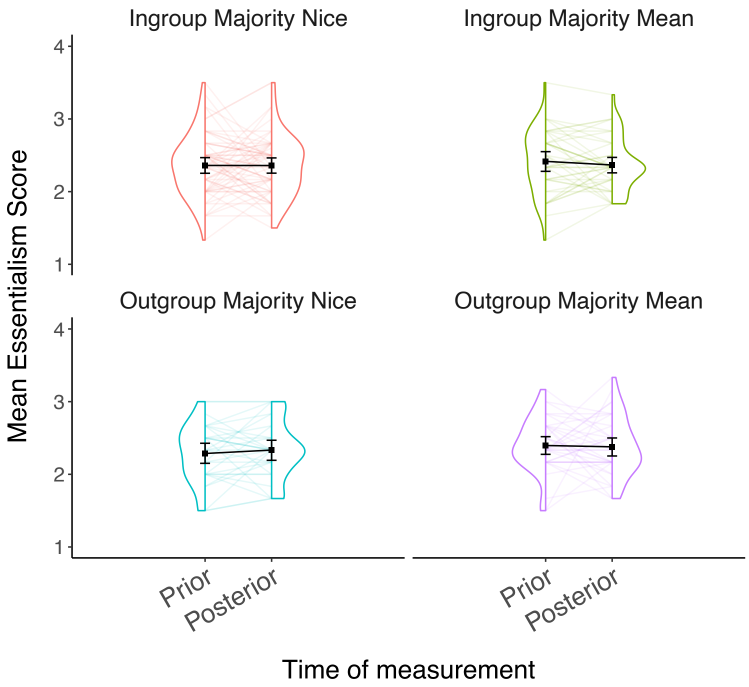

Figure 4. 3: Distribution of children's prior and posterior essentialism scores by condition in Experiment 1. The error bars indicate bootstrapped 95% CIs.

Expectation

The proportion of children who expected a randomly drawn child from the room to be nice is shown in Figure 4.4. We used mixed-effects logistic regression to predict children's expectations (nice = 1, mean = 0) from group condition, trait distribution condition, time of measurement, age (z-scored), gender, and their interactions, with random intercepts for participants. The best-fitting model included the interaction of trait distribution condition and time of measurement and the interaction of group condition and trait distribution condition as predictors. Observing the mostly mean sample led children to become less likely to expect a random child to be nice (β = -2.17, *SE* = 0.58, *p* < .001), and observing the mostly nice sample did not significantly change children's expectations (β = -0.43, *SE* = 0.42, *p* = .30). Children in the Majority Nice condition had more positive expectations for the ingroup than the outgroup (β = 2.21, *SE* = 0.58, *p* < .001), but children in the Majority Mean condition did not differ in their expectations for the ingroup and the outgroup (β = 0.35, *SE* = 0.56, *p* = .53).

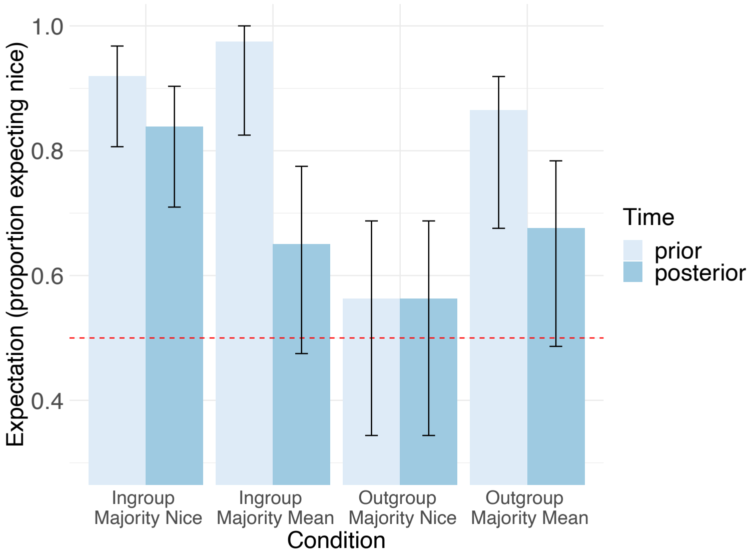

Figure 4. 4: Proportion of children who expected a randomly drawn child from the room to be nice, by condition and time in Experiment 1. The dashed line indicates chance selection (.5), and the error bars indicate bootstrapped 95% CIs.

Over-hypothesis

The proportion of children who expected a randomly drawn child from the room to be smart is shown in Figure 4.5. We used mixed-effects logistic regression to predict children's over-hypothesis (smart = 1, not smart = 0) from group condition, trait distribution condition, time of measurement, age (z-scored), gender, and their interactions, with random intercepts for participants. The best-fitting model included the three-way interaction of group condition, trait distribution condition and time of measurement as predictors. In the Ingroup condition, children's expectations that a random child from the room was smart decreased after observing a mostly nice sample (β = -2.72, *SE* = 1.18, *p* = .02), but did not change after observing a mostly mean sample (β = 0.00, *SE* = 1.06, *p* = .99). In the Outgroup condition, children's expectations that a random child from the room was smart increased non-significantly after observing a mostly nice sample (β = 1.47, *SE* = 1.31, *p* = .26), and decreased after observing a mostly mean sample (β = -6.14, *SE* = 1.94, *p* = .002).

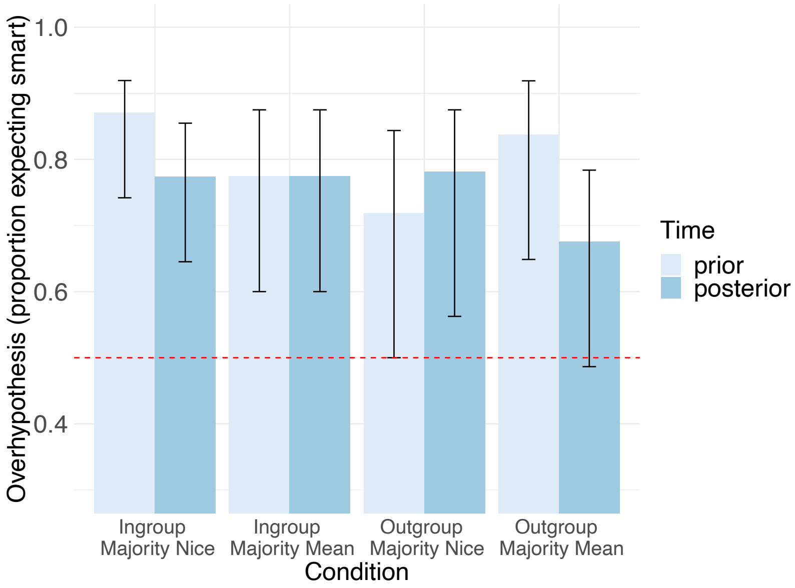

Figure 4. 5: Proportion of children who expected a randomly drawn child from the room to be smart, by condition and time in Experiment 1. The dashed line indicates chance selection (.5), and the error bars indicate bootstrapped 95% CIs.

4.2.3. Discussion

The goal of the present experiment was to examine whether statistically representative counterevidence can change 5- to 6-year-olds' attitudes and beliefs about minimal groups. We assigned children to minimal groups. Then, we showed children the trait distribution of a sample of children randomly drawn from either their ingroup or their outgroup. Lastly, we examined whether the new evidence changed children's attitudes and beliefs about the groups.

After children were assigned to the minimal groups, they showed clear ingroup biases. First, when children were asked to guess from which group the sample was drawn, those who observed the mostly nice sample were more likely to guess the sample was from their ingroup rather than from their outgroup, compared to those who observed the mostly mean sample. Second, overall, children showed more positive attitudes toward their ingroup than toward their outgroup. These results are consistent with the past literature on ingroup bias for minimal groups (Dunham, 2018), and suggest that our minimal group manipulation was successful.

Most importantly, we found that observing the trait distribution of the randomly drawn sample changed children's attitudes in the predicted directions. Consistent with our hypotheses, children's attitudes toward both the ingroup and the outgroup became more positive after observing a mostly nice sample, and more negative after observing a mostly mean sample. Contrary to our hypotheses, children did not process the evidence in a biased way – the effect of the sample was similar for the ingroup and the outgroup. In addition, children showed a negativity bias – the mostly mean sample had a stronger negative effect on their attitudes than the mostly nice sample had a positive effect. This is consistent with past research showing that starting in infancy, negative stimuli have a larger impact on humans' social inferences and decisions than positive stimuli (Hamlin et al., 2010; Vaish et al., 2008).

Observing the sample also changed children's expectations about the distribution of nice vs. mean individuals in the groups. Children were less likely to expect a randomly drawn child from the ingroup and the outgroup to be nice after observing a mostly mean sample from the group. However, observing the mostly nice sample did not have a positive impact on children's expectations. Thus, children showed a negativity bias.

Observing the sample had weaker effects on children's expectations about the distribution of smart vs. not smart individuals in the groups. For the ingroup, given a mostly nice sample, children's expectations changed in the opposite direction as predicted – they were less likely to expect a randomly drawn child to be smart; given a mostly mean sample, their expectations did not change. For the outgroup, children's expectations changed in the predicted directions, but only the mostly mean sample had a statistically significant effect (consistent with a negativity bias). Thus, in general, children did not generalize what they learned about the distribution of trait nice to the distribution of trait smart.

Observing the sample did not change children's essentialist beliefs. An interesting finding was that younger children had stronger essentialist beliefs than older children. To our knowledge, no past studies have investigated the age effect on children's essentialist beliefs about minimal groups. However, a study by Davoodi and colleagues (2020) found that between ages 5 and 10, U.S. children's essentialist beliefs increased for gender, but decreased for categories that are less socially salient, such as sports-team supporters, consistent with our finding.

In conclusion, the present experiment provides initial evidence that 5- to 6-year-olds are sensitive to the trait distribution in statistically representative counterevidence and use that information to change their attitudes and beliefs about minimal groups. Children showed a negativity bias when they were learning from the statistically representative counterevidence – they were affected by negative information more than positive information. Future research should examine whether providing stronger evidence would be more effective in changing children's attitudes and beliefs in the positive direction.

In the next experiment, we will use the same paradigm to examine whether statistically representative counterevidence can change children's attitudes and beliefs about familiar social groups, for which they have stronger prior biases.

4.3. Experiment 2

4.3.1. Methods

Participants

Forty-three White children who were 5 to 6 years of age (20 females; mean age = 6.12; range = 5.03 to 6.96; *SD* = 0.55) participated in the experiment. Our target sample size is 96 White children and 96 Black children. The target sample size would provide us with at least 90% power (at $\alpha = .05$) to detect the effect sizes observed in a similar past study (Baron & Dunham, 2015). Parents filled out a demographic form to indicate the child's and the parents' race and ethnicity, and only White children whose parents are both White participated in the experiment. Participants were tested in a lab room, at children's museums, or over Zoom. Parents of the participants provided written informed consent prior to the experiment session.

Design and Procedure

The design and the procedure are similar to that of Experiment 1, except that the two rooms are filled with children of different racial groups (a Black children room and a White children room). The study employed a 2 (Group condition: Ingroup vs. Outgroup) × 2 (Trait Distribution condition: Majority nice vs. Majority mean) between-subject design. A visual schematic of the procedure is shown in Figure 4.6. We also added a debriefing phase at the end to show participants that children of different races and ethnicities are in the same room, and that all children of different races and ethnicities are nice.

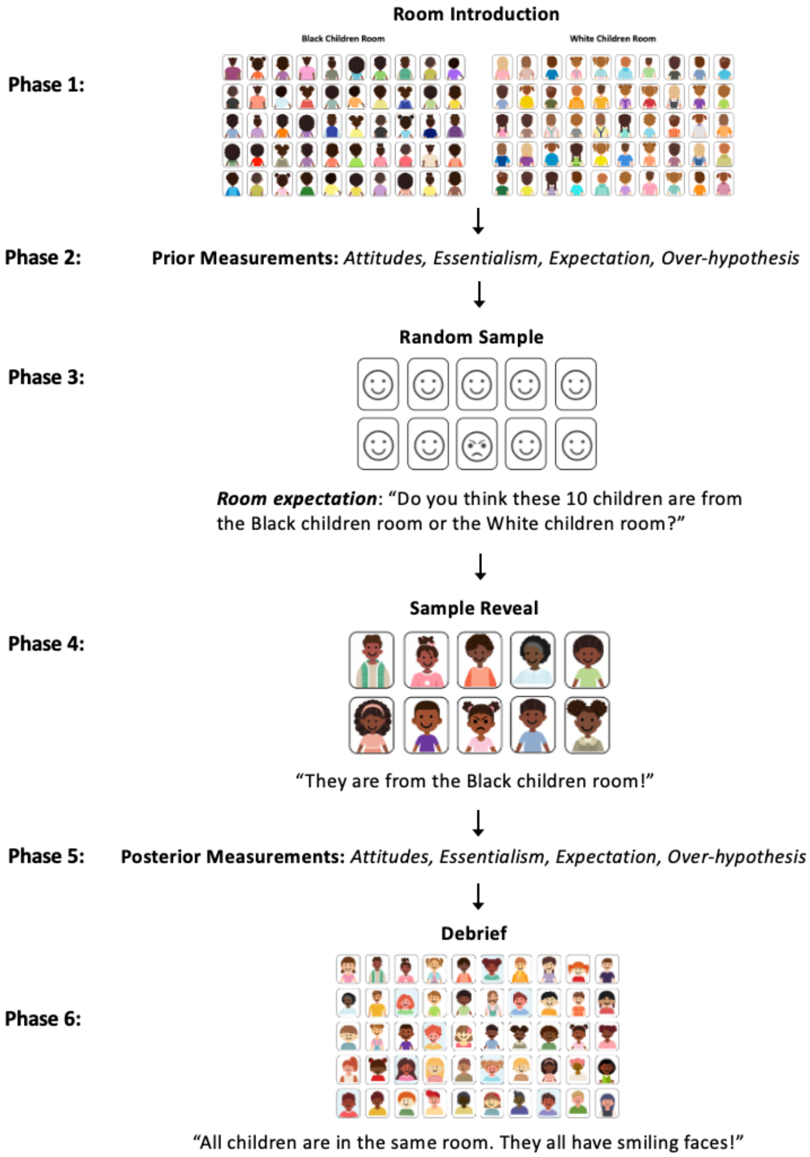
Room Introduction
Black Children Room
White Children Room
Phase 1:
Phase 2:
Prior Measurements: Attitudes, Essentialism, Expectation, Over-hypothesis
Random Sample
Phase 3:
Room expectation: "Do you think these 10 children are from the Black children room or the White children room?"
Sample Reveal
Phase 4:
"They are from the Black children room!"
Phase 5:
Posterior Measurements: Attitudes, Essentialism, Expectation, Over-hypothesis
Debrief
Phase 6:
"All children are in the same room. They all have smiling faces!"

Figure 4. 6: A visual schematic of the procedure in Experiment 2.

4.3.2. Results

Room Expectation

Table 4.2 shows the number of children who expected the Majority Nice sample and the Majority Mean sample to be from the ingroup or the outgroup. We used logistic regression to predict children's room expectations (ingroup = 1, outgroup = 0) from trait distribution condition, age (z-scored), gender, and their interactions. The best-fitting model included the interaction of trait distribution condition and age. With increasing age, children were less likely to expect the majority mean sample to be from the ingroup (β = -1.40, *SE* = 0.66, *p* = .03). With increasing age, children expected that the majority nice sample was more likely than the majority mean sample to be from the ingroup (β = 1.58, *SE* = 0.80, *p* = .049).

Table 4.2: Room expectation results in Experiment 2

Trait distribution condition	Room expectation	
	Ingroup	Outgroup
Majority Nice	11	10
Majority Mean	8	14

Attitudes

The distribution of children's prior and posterior attitude scores is shown in Figure 4.7. We used mixed-effects ANOVAs to predict children's attitude scores from group condition, trait distribution condition, time of measurement, age (z-scored), gender, and their interactions, with random intercepts for participants. We did not find any significant effects. However, after observing a mostly nice sample from either the ingroup or the outgroup, children's attitudes changed in the predicted, positive direction (not statistically significant).

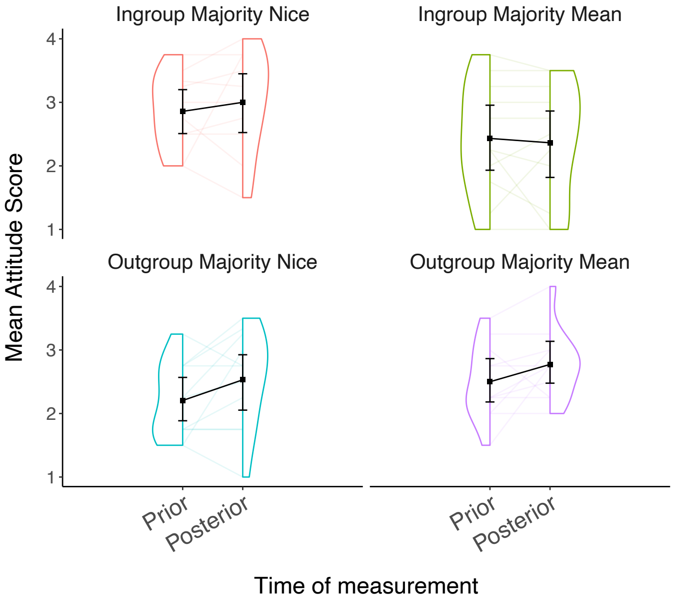

Figure 4. 7: Distribution of children's prior and posterior attitude scores by condition in Experiment 2. The error bars indicate bootstrapped 95% CIs.

Essentialism

The distribution of children's prior and posterior essentialism scores is shown in Figure 4.8. We used mixed-effects ANOVAs to predict children's essentialism scores from group condition, trait distribution condition, time of measurement, age (z-scored), gender, and their interactions, with random intercepts for participants. The best-fitting model included trait distribution condition and age as predictors. Children had stronger essentialist beliefs in the Majority Nice condition than in the Majority Mean condition (β = 0.27, *SE* = 0.10, *p* = .008). Children had weaker essentialist beliefs about racial groups with increasing age (β = -0.36, *SE* = 0.09, *p* < .001).

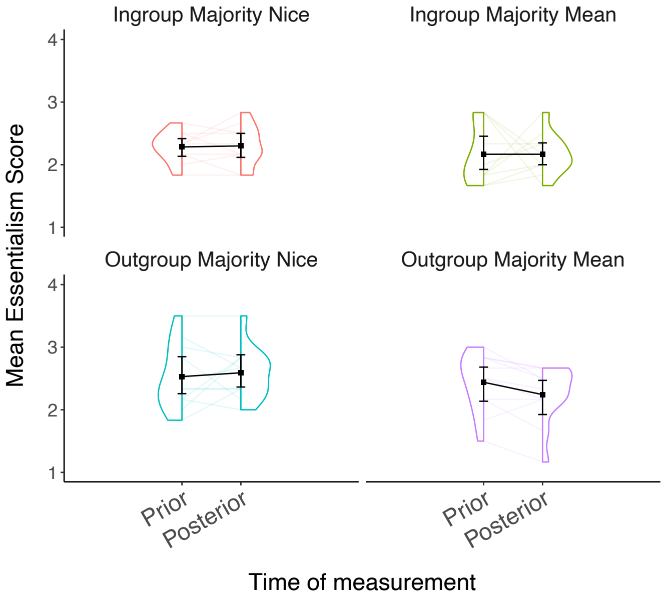

Figure 4. 8: Distribution of children's prior and posterior essentialism scores by condition in Experiment 2. The error bars indicate bootstrapped 95% CIs.

Expectation

The proportion of children who expected a randomly drawn child from the room to be nice is shown in Figure 4.9. We used mixed-effects logistic regression to predict children's expectations (nice = 1, mean = 0) from group condition, trait distribution condition, time of measurement, age (z-scored), gender, and their interactions, with random intercepts for participants. We did not find any significant effects. However, children's expectations changed in the predicted directions: after observing a mostly nice sample about either the ingroup or the outgroup, children's expectations changed in the positive direction (not statistically significant), and after observing a mostly mean sample about either the ingroup or the outgroup, children's expectations changed in the negative direction (not statistically significant).

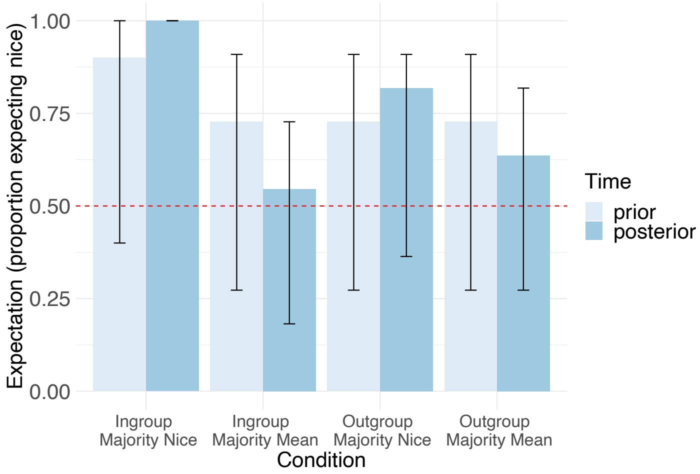

Figure 4. 9: Proportion of children who expected a randomly drawn child from the room to be nice, by condition and time in Experiment 2. The dashed line indicates chance selection (.5), and the error bars indicate bootstrapped 95% CIs.

Over-hypothesis

The proportion of children who expected a randomly drawn child from the room to be smart is shown in Figure 4.10. We used mixed-effects logistic regression to predict children's over-hypothesis (smart = 1, not smart = 0) from group condition, trait distribution condition, time of measurement, age (z-scored), gender, and their interactions, with random intercepts for participants. The best-fitting model included time of measurement as the only predictor. Children were less likely to expect that a random child would be smart after observing a sample (β = -7.14, *SE* = 2.21, *p* = .001).

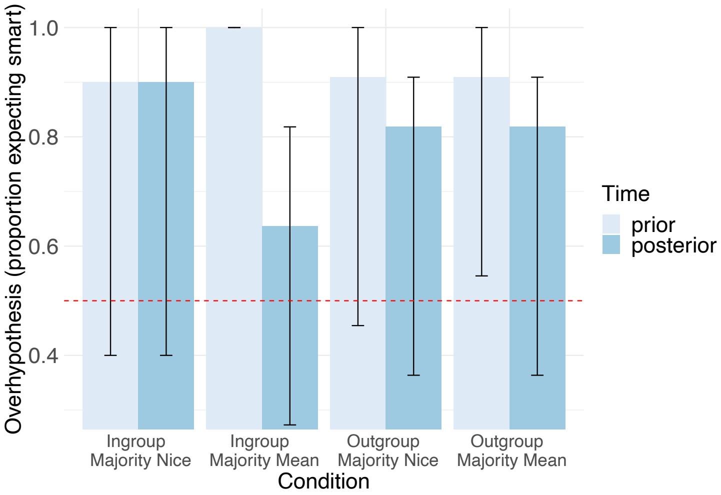

Figure 4. 10: Proportion of children who expected a randomly drawn child from the room to be smart, by condition and time in Experiment 2. The dashed line indicates chance selection (.5), and the error bars indicate bootstrapped 95% CIs.

4.3.3. Discussion

The present experiment examined whether statistically representative counterevidence can change White children's attitudes and beliefs about racial groups. We showed children the trait distribution of a sample of children randomly drawn from either the group of White children or the group of Black children. Then, we measured whether the new evidence changed their attitudes and beliefs about the groups.

In the room expectation measure, we found that children showed greater ingroup biases with increasing age. Older children were more likely to guess that the mostly nice sample was from their ingroup and the mostly mean sample was from the outgroup, compared to younger children. Thus, between the ages of 5 and 6 years, we observed an increase in White children's racial biases.

Most importantly, children's attitudes and expectations about racial groups changed in the hypothesized directions after observing the statistically representative counterevidence. For the attitudes measure, after observing a mostly nice sample from either the ingroup or the outgroup, there was a trend that children's attitudes toward the group became more positive. After observing a mostly mean sample from the ingroup, there was a trend that children's attitudes toward the ingroup became more negative. The only change in the unpredicted direction is that after observing a mostly mean sample from the outgroup, there was a trend that children's attitudes toward the outgroup became more positive. This is probably due to the small sample size of this experiment. For the expectation measure, the trends showed that children were more likely to expect a randomly drawn child from the ingroup or the outgroup to be nice after observing a mostly nice sample from the group, and less likely to expect a randomly drawn child from the ingroup or the outgroup to be nice after observing a mostly mean sample from the group. Thus, these findings provided preliminary evidence that statistically representative counterevidence can change children's attitudes and beliefs about racial groups in the expected directions.

For the over-hypothesis measure, children became less likely to expect a randomly drawn child from the ingroup or the outgroup to be smart after observing a sample, regardless of the trait distribution of the sample. This suggests that children's expectations about the distribution of trait smart in the groups were not affected by the distribution of trait nice in the sample they observed. Thus, children did not generalize what they learned about the distribution of one trait to the distribution of other traits. However, we need a larger sample to assess the robustness of this finding.

The evidence did not have any statistically significant effect on children's essentialist beliefs. However, we found that older children had weaker essentialist beliefs about racial groups than younger children. Past research suggests that the development of essentialist beliefs about social groups depends on environmental and cultural inputs (see Rhodes & Mandalaywala, 2017 for a review). We tested children from counties with diverse racial and ethnic populations[2]. Thus, it is possible that contact with a diverse population allowed children in our experiment to form weaker essentialist beliefs about race with increasing age. In addition, we found an unexpected result that children who were assigned to the Majority Nice condition had stronger essentialist beliefs about racial groups than children assigned to the Majority Mean condition. This unexpected finding might be due to the small sample size of this experiment.

[2] County A: 47.8% White, 33.8% Asian, 22.4% Hispanic or Latino, 10.7% Black, 5.6% Mixed race, 1.0% American Indian and Alaska Native; County B: 77.5% White, 6.3% Hispanic or Latino, 7.9% Asian, 1.9% Black, 11.7% Mixed race (U.S. Census Bureau, 2020).

In conclusion, we found preliminary evidence that White children who are 5 to 6 years of age can use the trait distribution in statistically representative counterevidence to change their attitudes and beliefs about racial groups. A limitation of the present experiment is that the sample size is small. Another limitation is that we have only tested children from one racial group – White children. Past research has shown that while children from the majority and higher status groups manifest the typical ingroup favoritism, children from minority or lower status groups do not: they either prefer the majority/higher status group or show no preference (Newheiser & Olson, 2012; Shutts et al., 2011). In future work, we plan to test 96 White children and 96 Black children in this study to examine how group membership affects children's preexisting biases and how likely they would revise their biases given statistically representative counterevidence.

4.4. General Discussion

In two experiments, we used a novel paradigm of showing statistically representative counterevidence to change 5- to 6-year-olds' attitudes and beliefs about minimal groups and racial groups. We found preliminary evidence that children are sensitive to the trait distribution of a sample randomly drawn from a social group, and that this information was effective in changing their attitudes and beliefs about both minimal groups and racial groups.

Furthermore, children's processing of the statistically representative counterevidence was not filtered by their preexisting biases. Based on past research (Bigler & Liben, 2006; Over et al., 2018; Schug et al., 2013), we hypothesized that the mostly nice sample would have a larger positive effect on the ingroup than the outgroup, and the mostly mean sample would have a larger negative effect on the outgroup than the ingroup. However, our findings showed that children were equally likely to change their attitudes and beliefs about minimal ingroup and minimal outgroup in the respective directions given the sample. We also did not find evidence that children processed the evidence in a biased way for racial groups with the current results. One possibility is that the statistically representative counterevidence is a stronger piece of evidence than evidence used in past studies, therefore children were more likely to take this piece of evidence into account regardless of whether it was consistent with their prior biases. Future studies could test this possibility with other real social groups such as gender and ethnicity.

Children showed a negativity bias when they learned from statistically representative counterevidence about minimal groups, such that the mostly mean sample had a stronger negative effect than the mostly nice sample had a positive effect. So far, we have not found the same negativity bias with racial groups (possibly due to the small sample size). This negativity bias is consistent with past research (Hamlin et al., 2010; Vaish et al., 2008), and suggests that children might need larger amounts of positive evidence in order to effectively change their attitudes and beliefs about social groups in the positive direction.

In order to examine whether children would generalize what they learned about the distribution of trait nice to the distribution of other traits in the social groups, we also measured children's beliefs about the distribution of trait smart before and after they observed the evidence. We found that children did not generalize the distribution of trait nice to the distribution of trait smart. This suggests that children did not form general positive impressions about individuals in the group after observing a mostly nice sample from the group (i.e., they did not show a Halo effect). This finding is also consistent with past research showing that nice (i.e., warmth) and smart (i.e., competence) might be 2 distinct dimensions of social group stereotypes in both adults (Fiske et al., 2002) and children (Baharloo et al., 2022).

Lastly, we found that statistically representative counterevidence did not affect children's essentialist beliefs about minimal groups or racial groups. Overall, children showed weak essentialist beliefs about both minimal groups and racial groups (the mean essentialism scores were 2.36 and 2.34, respectively, on a scale of 1 to 4). Indeed, past research suggests that children do not form essentialist beliefs about minimal groups, unless additional information (e.g., generic language) is provided, and U.S. children do not form strong essentialist beliefs about race until around 10 years of age (Rhodes et al., 2012; Rhodes & Mandalaywala, 2017). Thus, future studies could test older children to examine whether statistically representative counterevidence might also influence children's essentialist beliefs.

These findings make important theoretical contributions to the domains of social and cognitive development. Past studies using the method of exposure to counterstereotypic exemplars showed mixed results in its effectiveness in changing adults' and children's intergroup biases (Block et al., 2022; Gonzalez et al., 2017; Gonzalez et al., 2021; Lai et al., 2014). One possible reason is that adults and children can mentally group the counterstereotypic exemplars into a subtype and view them as exceptions to the group. Leveraging children's understanding that a randomly drawn sample is representative of the group (Denison & Xu, 2019), the current study aimed to prevent subtyping by showing children counterevidence in the form of randomly drawn samples from social groups. The findings suggest that showing statistically representative counterevidence might be a more effective way of presenting counterevidence about social groups to young children. In addition, this study adds to the large body of research showing that children rationally learn from new evidence to update their beliefs (e.g., Kimura & Gopnik, 2019; Kushnir & Gopnik, 2007; Lucas et al., 2014).

These findings also have important real-world implications. In the current study, a minimal intervention of showing children a piece of statistically representative counterevidence for less than a minute was already successful in changing children's attitudes and beliefs in the expected directions. This paradigm can be adapted to design real-world interventions, for example, by showing children larger amounts of statistically representative counterevidence for longer periods of time. In contrast to interventions that reduce bias through intensive cultural exposure (e.g., reducing children's implicit anti-dark-skin bias through a curriculum on African music and musicians; Neto et al., 2015), the present intervention is brief and much easier to implement on a larger scale. In contrast to other types of brief interventions (e.g., exposure to counterstereotypic exemplars), the present intervention might be more effective as it avoids the issue of subtyping.

One limitation of the current study is that we only examined this paradigm with children from one culture (i.e., children in the United States). Children's intergroup attitudes and essentialist beliefs vary across cultures. For example, Israeli children develop essentialist beliefs about ethnicity at a younger age compared to U.S. children's essentialist beliefs about race (Diesendruck et al., 2013). In planned future work, we will examine whether this paradigm would also be effective in changing Jewish and Arab children's attitudes and beliefs about ethnic groups in Israel with our collaborators.

Overall, the current research provides another strong piece of evidence that 5- to 6-year-olds have powerful statistical learning mechanisms, and they can change their deeply ingrained biases about social groups with minimal interventions. The current work paves the way for many exciting future works to investigate the role statistically representative counterevidence can play in changing children's intergroup biases.

Chapter 5: Conclusions

The current dissertation investigates whether there are limits to humans' ability to rationally revise our beliefs. Chapters 2 and 3 examine this question with basic research – Can children and adults revise their most fundamental beliefs about objects and agents given a small number of counterevidence? Chapter 4 examines this question in a domain that goes beyond basic research – Can a minimal intervention of showing children statistically representative counterevidence change their biases about social groups?

5.1. Basic research

5.1.1. Conclusions and implications of the empirical work

In Chapters 2 and 3, we found that the core knowledge principles about objects and agents can be revised in 4- to 6-year-olds and adults given a small amount of counterevidence. Children in our experiments had 4 to 6 years of experience and adults had more than 18 years of experience interacting with objects and agents supporting these principles. However, when they observed just a few events violating these principles, between a third and a half of learners genuinely accepted the counterevidence and revised their beliefs about these principles.

Furthermore, we found some important differences between the object principles and the agent principles. Learners have weaker prior beliefs about the agent principles than the object principles. The agent principles are more easily revisable than the object principles. Learners more readily generalize the new principles about agents to new contexts than new principles about objects. Thus, the core knowledge system of agents is more flexible than the core knowledge system of objects.

5.1.2. Implications for Rational Constructivist and Bayesian frameworks

Chapters 2 and 3 provide strong support for the Rational Constructivist theory and the Bayesian framework of cognitive development (Chater & Oaksford, 2008; Fedyk & Xu, 2018; Gopnik & Wellman, 2012; Tenenbaum et al., 2011; Xu, 2019; Ullman & Tenenbaum, 2020). These frameworks posit that belief revision should always be possible given the right kind of counterevidence, even when we have strong prior beliefs. The principles in the core knowledge systems of objects and agents have been argued to be evolutionarily endowed, innate in humans, and encapsulated (Spelke, 2022). Yet, the current work has shown that they can be revised in human children and adults given a small amount of counterevidence.

Furthermore, the Rational Constructivist and the Bayesian frameworks propose that belief revision should rationally integrate the strength of our prior beliefs and the strength of the evidence. The current work shows that given the same amount of counterevidence (3 to 6 violations of each principle), children and adults are more likely to revise their beliefs about the agent principles, for which they have weaker prior beliefs, compared to the object principles. Thus, when the strength of evidence is equivalent, the strength of prior beliefs determines how likely learners would revise their beliefs.

5.1.3. Implications for the Core Knowledge view

Chapters 2 and 3 also have important implications for the Core Knowledge view (Spelke, 1988, 2000, 2022; Spelke & Kinzler, 2007; Spelke, 2022). First, the current research has shown that the core principles in two of the core knowledge systems (objects and agents) are revisable given little counterevidence. This finding suggests that the core knowledge systems might not be completely encapsulated from conscious reasoning.

Furthermore, the comparison of data from Chapter 2 (objects) and Chapter 3 (agents) suggests that maybe not all core knowledge systems are created equal. The agent system is more flexible than the object system. One possibility for this domain difference is that infants are born with stronger prior beliefs about objects (i.e., the object system is more hard-wired to begin with); another possibility is that children and adults have observed more counterevidence about the psychological principles in everyday life, and therefore have weaker and more flexible beliefs about agents.

More broadly, we speculate that there might be two types of qualitatively different core knowledge systems – one type is more akin to perceptual systems, which are automatic, inflexible, and possibly encapsulated from conscious reasoning, and the other type resembles belief systems, which are more flexible and deliberate. Among the six core knowledge systems discussed in detail in Spelke (2022), we argue that the systems of objects and number (and perhaps space) may be of the first type, whereas the systems of agents and social beings (and perhaps form) are more likely to be of the second type.

For the object system, a large body of research suggests that adults' object representation depends on perceptual mechanisms (Scholl, 2001), and perception of objects is disrupted when objects do not follow the core physical principles such as continuity and cohesion (Scholl & Pylyshyn, 1999; Scholl et al., 2001; vanMarle & Scholl, 2003). Furthermore, object perception seems to be unaffected by the top-down influences of cognition (Firestone & Scholl, 2016).

For the number system, past research has shown clear evidence that the Approximate Number System (ANS) activates automatically and unconsciously in all ages (Izard et al., 2009; Nieder & Dehaene, 2009). The precision of ANS increases during infancy, perhaps due to the improvement of visual acuity (Xu & Arriaga, 2007; Xu & Spelke, 2000). In addition, the neurological signatures of the ANS remain constant from infancy to adulthood, unaffected by years of mathematical education (Hyde & Spelke, 2009, 2011).

On the other hand, the systems of agents and social beings are less automatic and encapsulated, and more likely to be part of our belief systems. Three-month-old infants do not automatically expect agents' actions to be directed to objects; they flexibly learn the goal (objects or locations) of an agent's actions based on the agent's previous behaviors (Woo et al., 2022). While 1-year-old infants and children older than 4-year-olds expect agents to take efficient paths to achieve their goals, 3-year-olds fail to show this expectation, suggesting that the development of the efficiency principle might be discontinuous (Gergely & Csibra, 2003; Gönül & Paulus, 2021).

Similarly, for the system of social beings, while expectations about how individuals interact and affiliate with one another emerge at a young age, these expectations are flexible and can be changed by infants' own social experiences. For instance, infants' social environments modulate their same-race preference – White and Black infants living in monoracial environments prefer faces of their own race, but Black infants living in predominantly White environments do not show same-race preference (Bar-Haim et al., 2006). Infants' linguistic environments also change their expectations about social groups – monolingual infants expect individuals who speak different languages to have different food preferences, but bilingual infants expect them to share food preferences (Liberman et al., 2016).

This distinction between perceptual vs. conceptual core knowledge systems makes interesting predictions that can be tested in future research. For example, children's and adults' revision of the core object principles in Chapter 2 might be less likely to affect the operation of these principles on the perceptual level – participants might revert to principle-consistent predictions about novel events when they are under cognitive load. More generally, learners might be more likely to accept the violations of the agent and social being systems compared to the object and number systems.

5.1.4. Future directions

The findings of Chapters 2 and 3 pave the way for important future research to further probe the limits of humans' ability to revise their beliefs about objects and agents. For instance, in ongoing work, we are investigating whether children and adults can generalize their revised principles about objects to more diverse contexts. Specifically, we show participants new evidence that objects can go through lighter blue walls but not darker blue walls. Then, we ask them to play a maze game to examine whether they would generalize the revised principles from the new evidence to completely different contexts – specifically, whether they would try to go through the lighter blue walls when navigating the maze.

We will also examine whether the core principles of objects and agents can be revised at an even younger age, in infants. Many past studies have shown that infants are surprised and look longer at events that violate their expectations about the object principles and the agent principles, compared to events that are consistent with their expectations (Aguiar & Baillargeon, 1999; Gergely & Csibra, 2003; Leslie & Keeble, 1987; Spelke et al., 1992; Wellman et al., 2016; Woodward, 1998). In a future study, we will show infants a few events violating each principle, and observe whether infants will change their expectations about these principles.

5.2. Beyond basic research

5.2.1. Conclusions and implications of the empirical work

Chapter 4 investigates the limits of humans' ability to rationally revise our beliefs in a domain that goes beyond basic research – changing biases about social groups. We found that 5- to 6-year-olds' biases about minimal groups and racial groups can be changed with a minimal intervention of showing a statistically representative counterevidence about the social groups. These findings provide suggestive evidence that statistically representative counterevidence might be a more effective way of presenting counterevidence about social groups to young children. Furthermore, this method might allow us to design short intervention programs to effectively combat intergroup biases in our society from as early as 5 years of age.

5.2.2. Implications for Rational Constructivist and Bayesian frameworks

Chapter 4 provides more strong support for the Rational Constructivist and the Bayesian frameworks of cognitive development (Chater & Oaksford, 2008; Fedyk & Xu, 2018; Gopnik & Wellman, 2012; Tenenbaum et al., 2011; Xu, 2019; Ullman & Tenenbaum, 2020). Humans are evolutionarily endowed with the tendency to form and attend to coalitions (Cosmides & Tooby, 2010; Pietraszewski et al., 2014), and we are predisposed to prefer ingroup over outgroup even based on minimal membership cues (Dunham, 2018; Mahajan & Wynn, 2012). The current research shows that these evolutionarily ancient and deeply ingrained ingroup biases can also be changed in human children when they are given counterevidence in a minimal intervention.

Furthermore, in past studies, exposure to counterstereotypic exemplars is not always successful in changing children's and adults' intergroup biases (Block et al., 2022; Gonzalez et al., 2017; Gonzalez et al., 2021; Lai et al., 2014). One possible reason is that learners can mentally cluster the exemplars into a subtype, and see the exemplars as exceptions of the social group (Richards & Hewstone, 2001; Hayes et al., 2003). The paradigm in Chapter 4 avoids the issue of subtyping by showing children a piece of evidence that is statistically representative of the entire social group. Subsequently, this stronger piece of counterevidence successfully changed children's intergroup biases. Thus, when the strength of prior biases is equivalent, the strength of the evidence determines how likely learners would change their biases.

5.2.3. Implications for Developmental Intergroup Theory

The Developmental Intergroup Theory (DIT; Bigler & Liben, 2006) proposes that the formation of stereotypes and prejudice depends on both externally driven processes such as the covariation between social categories and attributes, and internally driven processes such as ingroup bias and essentialist beliefs. Chapter 4 shows that externally driven processes, such as new evidence of covariation between social categories and attributes, could directly affect internally driven processes (i.e., ingroup biases), which could in turn have an impact on children's development of prejudice and stereotypes.

In addition, the DIT posits that when children are shown new evidence about social groups, their processing of the new evidence would still be filtered by their preexisting biases (Bigler & Liben, 2006). Chapter 4 provides preliminary evidence that given the right kind of evidence – a sample of individuals who are clearly representative of the entire social group – children might be able to rationally integrate this piece of evidence into their beliefs about the social group, without processing the evidence in a biased way.

5.2.4. Future directions

The findings of Chapter 4 open many possibilities for further investigating the role statistically representative counterevidence can play in changing children's intergroup biases.

First, the current study only examined whether children's attitudes and beliefs can be changed by a sample with the most extreme trait distributions (9 nice vs. 1 mean or 1 nice vs. 9 mean). In future work, we can vary the trait distributions of the sample (7 nice vs. 3 mean, 5 nice vs. 5 mean, or 3 nice vs. 7 mean) to examine if children change their attitudes and beliefs based on the trait distribution of the sample in a graded manner.

Second, future studies can examine the effect of the amount of evidence on children's intergroup biases. For instance, would showing children a second and a third randomly drawn sample from the group, all exhibiting the same distribution of traits, increase children's likelihood of changing their attitudes and beliefs?

Third, future studies can assess the long-term effectiveness of this intervention by measuring children's attitudes and beliefs an hour, a week, or a month after the intervention.

Lastly, the present study only targeted one set of traits (i.e., nice and mean) among a variety of traits that children attribute to different social groups (e.g., Black people are aggressive, girls are bad at math). Future studies can extend this method to target children's beliefs about other types of traits. For example, would a sample of 10 girls, with 9 girls preferring and excelling at math and 1 girl preferring and excelling at reading, change children's gender stereotypes about math and reading?

5.3. Concluding remarks

Taken together, the current dissertation shows that humans have powerful learning mechanisms and suggests that we might have the ability to revise *any* beliefs with new evidence.

Given little counterevidence, children and adults can revise their deeply entrenched beliefs about objects and agents. Given minimal intervention, children can change their deeply ingrained biases about social groups.

However, if these beliefs are evolutionarily endowed, they must have been beneficial in guiding our reasoning and behaviors in the physical, psychological, and social worlds in our evolutionary history. Why should humans revise these beliefs given new evidence? Is the ability to revise the beliefs in the core knowledge systems a *feature* or a *bug* of humans' learning mechanisms? I would argue that this ability might be a *unique feature* of humans' learning mechanisms. Humans occupy the widest range of habitats on Earth among all terrestrial species (Klein, 2009). Inhabiting such diverse environments requires humans to forego our previous knowledge, tools, and social arrangements, and rapidly develop new ones to adapt to new environments (Boyd et al., 2011). We accomplished these feats not only because of our intelligence but also because we have the ability to completely overturn previous beliefs and principles and learn new ones.

In recent decades, humans have also set foot on the Moon, landed rovers on Mars, and photographed other galaxies. One day, humans might migrate to a different planet and inhabit environments that are completely different from the environments on Earth. The current research suggests that even if humans went to a planet where none of the physical, psychological, and social principles that we learned on Earth holds, we would still be able to quickly adapt to the new environments, and learn completely different physical, psychological, and social principles if necessary.

References

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the "light-from-above" prior. *Nature Neuroscience*, *7*(10), 1057–1058.

Aguiar, A., & Baillargeon, R. (1999). 2.5-month-old infants' reasoning about when objects should and should not be occluded. *Cognitive Psychology*, *39*(2), 116–157.

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310.

Amsterlaw, J., & Wellman, H. M. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development*, *7*(2), 139–172.

Augoustinos, M., & Rosewarne, D. L. (2001). Stereotype knowledge and prejudice in children. *British Journal of Developmental Psychology*, *19*(1), 143–156.

Baharloo, R., Fei, X., & Bian, L. (2022, August 1). The development of racial stereotypes about warmth and competence. https://doi.org/10.31234/osf.io/r28yq

Baillargeon, R. (2008). Innate ideas revisited: For a principle of persistence in infants' physical reasoning. *Perspectives on Psychological Science*, *3*(1), 2–13.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Bar-Haim, Y., Ziv, T., Lamy, D., & Hodes, R. M. (2006). Nature and nurture in own-race face processing. *Psychological Science*, *17*(2), 159–163.

Baron, A. S., & Dunham, Y. (2015). Representing 'us' and 'them': Building blocks of intergroup cognition. *Journal of Cognition and Development*, *16*(5), 780–801.

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., Bolz, M., Henrich, J., Setoh, P., Wang, J., & Laurence, S. (2013). Early false-belief understanding in traditional non-Western societies. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1755), 20122654.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bian, L., & Baillargeon, R. (2022). When are similar individuals a group? Early reasoning about similarity and in-group support. *Psychological Science*, *33*(5), 752–764.

Bigler, R. S., Jones, L. C., & Lobliner, D. B. (1997). Social categorization and the formation of intergroup attitudes in children. *Child development*, *68*(3), 530-543.

Bigler, R. S., & Liben, L. S. (2006). A developmental intergroup theory of social stereotypes and prejudice. In *Advances in Child Development and Behavior* (Vol. 34, pp. 39–89). Elsevier.

Block, K., Gonzalez, A. M., Choi, C. J. X., Wong, Z. C., Schmader, T., & Baron, A. S. (2022). Exposure to stereotype-relevant stories shapes children's implicit gender stereotypes. *PLOS ONE*, *17*(8), e0271396.

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*(4), 215–234.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108*(Supplement_2), 10918–10925.

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

Carey, S. (2009). *The origins of concepts.* New York, NY: Oxford University Press.

Chandler, M. J., & Lalonde, C. E. (1994). Surprising, magical and miraculous turns of events: Children's reactions to violations of their early theories of mind and matter. *British Journal of Developmental Psychology*, *12*(1), 83–95.

Chater, N., & Oaksford, M. (Eds.). (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press, USA.

Colomer, M., Bas, J., & Sebastian-Galles, N. (2020). Efficiency as a principle for social preferences in infancy. *Journal of Experimental Child Psychology*, *194*, 104823.

Colomer, M., & Woodward, A. (2023). Should I learn from you? Seeing expectancy violations about action efficiency hinders social learning in infancy. *Cognition*, *230*, 105293.

Cosmides, L., & Tooby, J. (2010). Groups in mind: The coalitional roots of war and morality. *Human morality and sociality: Evolutionary and comparative perspectives*.

Daum, M. M., & Krist, H. (2009). Dynamic action in virtual environments: Constraints on the accessibility of action knowledge in children and adults. *Quarterly Journal of Experimental Psychology*, *62*(2), 335-351.

Davoodi, T., Soley, G., Harris, P. L., & Blake, P. R. (2020). Essentialization of social categories across development in two cultures. *Child Development*, *91*(1), 289–306.

Denison, S., & Xu, F. (2019). Infant statisticians: The origins of reasoning under uncertainty. *Perspectives on Psychological Science*, *14*(4), 499–509.

Diesendruck, G., Goldfein-Elbaz, R., Rhodes, M., Gelman, S., & Neumark, N. (2013). Cross-cultural differences in children's beliefs about the objectivity of social categories. *Child Development*, *84*(6), 1906–1917.

Diesendruck, G., & haLevi, H. (2006). The role of language, appearance, and culture in children's social category-based induction. *Child Development*, *77*(3), 539–553.

Doan, T., Denison, S., Lucas, C., & Gopnik, A. (2015, July). Learning to reason about desires: An infant training study. *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Dunham, Y. (2018). Mere membership. *Trends in Cognitive Sciences*, *22*(9), 780–793.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2006). From American city to Japanese village: A cross-cultural investigation of implicit race attitudes. *Child Development*, *77*(5), 1268–1281.

Dunham, Y., Baron, A. S., & Carey, S. (2011). Consequences of "minimal" group affiliations in children. *Child Development*, *82*(3), 793–811.

Fedyk, M., & Xu, F. (2018). The epistemology of Rational Constructivism. *Review of Philosophy and Psychology*, *9*(2), 343–362.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and brain sciences*, *39*, e229.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, *82*(6), 878.

Fodor, J. (1975). *Language of thought.* Cambridge, MA: MIT Press.

Gelman, S. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, *8*(9), 404–409.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gonzalez, A. M., Steele, J. R., & Baron, A. S. (2017). Reducing children's implicit racial bias through exposure to positive out-group exemplars. *Child Development*, *88*(1), 123–130.

Gonzalez, A. M., Steele, J. R., Chan, E. F., Lim, S. A., & Baron, A. S. (2021). Developmental differences in the malleability of implicit racial bias following exposure to counterstereotypical exemplars. *Developmental Psychology*, *57*(1), 102–113.

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

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108.

Gordon, P. (2004). Numerical cognition without words: evidence from Amazonia. *Science, 306*, 496 – 499.

Gönül, G., & Paulus, M. (2021). Children's reasoning about the efficiency of others' actions: The development of rational action prediction. *Journal of Experimental Child Psychology*, *204*, 105035.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science*, *35*(8), 1407–1455.

Halim, M. L., & Ruble, D. (2010). Gender identity and stereotyping in early and middle childhood. In J. C. Chrisler & D. R. McCreary (Eds.), *Handbook of Gender Research in Psychology* (pp. 495–525). Springer New York.

Hamlin, J. K., Wynn, K., & Bloom, P. (2010). 3-month-olds show a negativity bias in their social evaluations. *Developmental Science*, *13*(6), 923–929.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know?. *Animal behaviour*, *61*(1), 139-151.

Hayes, B. K., Foster, K., & Gadd, N. (2003). Prior knowledge and subtyping effects in children's category learning. *Cognition*, *88*(2), 171–199.

Hershberger, W. (1970). Attached-shadow orientation perceived as depth by chickens reared in an environment illuminated from below. *Journal of Comparative and Physiological Psychology*, *73*(3), 407–411.

Hirschfeld, L. A. (1995). Do children have a theory of race? *Cognition, 54*(2), 209–252.

Huber, S., Krist, H., & Wilkening, F. (2003). Judgment and action knowledge in speed adjustment tasks: Experiments in a virtual environment. *Developmental Science*, *6*(2), 197-210.

Hyde, D. C. & Spelke, E. S. (2009). All numbers are not equal: An electrophysiological investigation of small and large number representations. *Journal of Cognitive Neuroscience, 21*(6), 1039-1053.

Hyde, D. C. & Spelke, E. S. (2011). Neural signatures of number processing in human infants: Evidence for two core systems underlying numerical cognition. *Developmental Science, 14*(2), 360-371.

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences, 106*(25), 10382-10385.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The Naïve Utility Calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Johnson, C. N., & Harris, P. L. (1994). Magic: Special but not excluded. *British Journal of Developmental Psychology*, *12*(1), 35-51.

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Gibson, A., Smith, M., Ge, L., & Pascalis, O. (2005). Three-month-olds, but not newborns, prefer own-race faces. *Developmental Science*, *8*(6), F31–F36.

Kimura, K., & Gopnik, A. (2019). Rational higher-order belief revision in young children. *Child Development*, *90*(1), 91–97.

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, *104*(30), 12577–12580.

Klein, R. G. (2009). *The human career: Human biological and cultural origins*. University of Chicago Press.

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology*, *43*(1), 186–196.

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

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., … Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785.

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

Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children’s causal explanatory reasoning: Causal explanatory reasoning in children. *Child Development*, *81*(3), 929–944.

Legare, C. H., Schult, C. A., Impola, M., & Souza, A. L. (2016). Young children revise explanations in response to new evidence. *Cognitive Development*, *39*, 45–56.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265–288.

Liberman, Z., Woodward, A. L., Sullivan, K. R., & Kinzler, K. D. (2016). Early emerging system for reasoning about the social nature of food. *Proceedings of the National Academy of Sciences*, *113*(34), 9480–9485.

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014a). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, *34*(1), 113–147.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., Markson, L., & Hu, J. (2014b). The child as econometrician: A rational model of preference understanding in children. *PLoS ONE*, *9*(3), e92160.

Ma, L., & Xu, F. (2011). Young children's use of statistical sampling evidence to infer the subjectivity of preferences. *Cognition*, *120*(3), 403–411.

Mahajan, N., & Wynn, K. (2012). Origins of "Us" versus "Them": Prelinguistic infants prefer similar others. *Cognition*, *124*(2), 227–233.

Masnick, A. M., Klahr, D., & Knowles, E. R. (2017). Data-driven belief revision in children and adults. *Journal of Cognition and Development*, *18*(1), 87–109.

McCoy, J., & Ullman, T. (2019). Judgments of effort for magical violations of intuitive physics. *PLOS ONE*, *14*(5), e0217513.

Mullen, B., Brown, R., & Smith, C. (1992). Ingroup bias as a function of salience, relevance, and status: An integration. *European Journal of Social Psychology, 22*(2), 103–122.

Neto, F., da Conceiçao Pinto, M., & Mullet, E. (2016). Can music reduce anti-dark-skin prejudice? A test of a cross-cultural musical education programme. *Psychology of Music*, *44*(3), 388–398.

Newheiser, A.-K., & Olson, K. R. (2012). White and Black American children's implicit intergroup bias. *Journal of Experimental Social Psychology*, *48*(1), 264–270.

Nieder, A. & Dehaene, S. (2009). Representation of number in the brain. *Annual Review of Neuroscience, 32,* 185-208.

Over, H., Eggleston, A., Bell, J., & Dunham, Y. (2018). Young children seek out biased information about social groups. *Developmental Science*, *21*(3), e12580.

Perez, J., & Feigenson, L. (2022). Violations of expectation trigger infants to search for explanations. *Cognition*, *218*, 104942.

Piaget, J. (1954). *The construction of reality in the child.* New York, NY: Routledge.

Pietraszewski, D., Cosmides, L., & Tooby, J. (2014). The content of our cooperation, not the color of our skin: An alliance detection system regulates categorization by coalition and race, but not sex. *PloS one*, *9*(2), e88534.

Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences*, *110*(41), E3965–E3972.

Quinn, P. C., Yahr, J., Kuhn, A., Slater, A. M., & Pascalis, O. (2002). Representation of the gender of human faces by infants: A preference for female. *Perception*, *31*(9), 1109–1121.

Regolin, L., & Vallortigara, G. (1995). Perception of partly occluded objects by young chicks. *Perception & Psychophysics*, *57*(7), 971–976.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*(1), 12–21.

Richards, Z., & Hewstone, M. (2001). Subtyping and subgrouping: Processes for the prevention and promotion of stereotype change. *Personality and Social Psychology Review*, *5*(1), 52–73.

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34), 13526–13531.

Rhodes, M., & Mandalaywala, T. M. (2017). The development and developmental consequences of social essentialism: Social essentialism. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(4), e1437.

Schleihauf, H., Herrmann, E., Fischer, J., & Engelmann, J. M. (2022). How children revise their beliefs in light of reasons. *Child Development*, 93, 1072–1089.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology*, *38*(2), 259–290.

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple object tracking. *Cognition*, *80*(1–2), 159–177.

Schug, M. G., Shusterman, A., Barth, H., & Patalano, A. L. (2013). Minimal-group membership influences children's responses to novel experience with group members. *Developmental Science*, *16*(1), 47–55.

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

Shu, T., Bhandwaldar, A., Gan, C., Smith, K. A., Liu, S., Gutfreund, D., Spelke, E., Tenenbaum, J. B., & Ullman, T. D. (2021). AGENT: A Benchmark for Core Psychological Reasoning. *ArXiv:2102.12321 [Cs]*.

Shutts, K., Kinzler, K. D., Katz, R. C., Tredoux, C., & Spelke, E. S. (2011). Race preferences in children: Insights from South Africa: Social preferences in South Africa. *Developmental Science*, *14*(6), 1283–1291.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking.* New York, NY: Oxford University Press.

Sinclair, S., Dunn, E., & Lowery, B. (2005). The relationship between parental racial attitudes and children's implicit prejudice. *Journal of Experimental Social Psychology*, *41*(3), 283–289.

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

Sodian, B., Licata, M., Kristen-Antonow, S., Paulus, M., Killen, M., & Woodward, A. (2016). Understanding of goals, beliefs, and desires predicts morally relevant theory of mind: A longitudinal investigation. *Child Development*, *87*(4), 1221–1232.

Spelke, E. S. (1988). The origins of physical knowledge. In L. Weiskrantz (Ed.), *Thought without language* (pp. 168–184). Clarendon Press/Oxford University Press.

Spelke, E. S. (2000). Core knowledge. *American Psychologist, 55*(11), 1233–1243.

Spelke, E. S. (2022). *What babies know: Core knowledge and composition volume 1*. Oxford University Press.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological review*, *99*(4), 605.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Stahl, A. E., & Feigenson, L. (2017). Expectancy violations promote learning in young children. *Cognition*, *163*, 1–14.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, *2*, 533-558.

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383–403.

vanMarle, K., & Scholl, B. J. (2003). Attentive tracking of objects versus substances. *Psychological Science*, *14*(5), 498–504.

van Schijndel, T. J. P., Visser, I., van Bers, B. M. C. W., & Raijmakers, M. E. J. (2015). Preschoolers perform more informative experiments after observing theory-violating evidence. *Journal of Experimental Child Psychology*, *131*, 104–119.

Vezzali, L., Giovannini, D., & Capozza, D. (2012). Social antecedents of children's implicit prejudice: Direct contact, extended contact, explicit and implicit teachers' prejudice. *European Journal of Developmental Psychology*, *9*(5), 569–581.

Wang, S., & Baillargeon, R. (2008). Can infants be "taught" to attend to a new physical variable in an event category? The case of height in covering events. *Cognitive Psychology*, *56*(4), 284–326.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual review of psychology*, *43*(1), 337-375.

Wellman, H. M., Kushnir, T., Xu, F., & Brink, K. A. (2016). Infants use statistical sampling to understand the psychological world. *Infancy*, *21*(5), 668-676.

Woo, B. M., Liu, S., & Spelke, E. (2022, June 30). Infants rationally infer the goals of other people's reaches in the absence of first-person experience with reaching actions. PsyArXiv. https://doi.org/10.31234/osf.io/dx2er

Woodward, A. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological Review*, *126*(6), 841–864.

Xu, F. & Arriaga, R. I. (2007). Number discrimination in 10-month-old infants. *British Journal of Developmental Psychology, 25*(1), 103-108.

Xu, F., & Kushnir, T. (Eds.). (2012). *Rational constructivism in cognitive development. Advances in Child Development and Behavior* (Vol. 43). Waltham, MA: Academic Press.

Xu, F., & Kushnir, T. (2013). Infants are rational constructivist learners. *Current Directions in Psychological Science*, *22*(1), 28–32.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1-B11.

Yee, M., & Brown, R. (1994). The development of gender differentiation in young children. *British Journal of Social Psychology*, *33*(2), 183–196.