Children's developing capacity to calibrate the verbal testimony of others with observed evidence when inferring causal relations

Niamh McLoughlin*[1,2], Zoe Finiasz[3], David M. Sobel[4], & Kathleen H. Corriveau[2]

[1]University of Kent, [2]Boston University, [3]Cornell University, [4]Brown University

Manuscript accepted for publication. *Journal of Experimental Child Psychology*

Corresponding author information: Niamh McLoughlin, School of Psychology, University of Kent, Canterbury, UK, CT2 7NP

Email: n.mcloughlin-545@kent.ac.uk

Highlights

- We tested the ability to calibrate verbal testimony with observable causal data.

- Five-year-olds calibrated confident claims with deterministic data.

- Uncertain, accurate, claims about probabilistic data aided children's inferences.

- The capacity to infer causal relations from distinct sources emerges by age 5.

Abstract

Across two studies (N = 120), we investigated the development of children's ability to calibrate the certainty of verbal testimony with observable data that varied in the degree of predictive causal accuracy. In Study 1, four- and 5-year-olds heard a certain or uncertain explanation about deterministic causal relations. The 5-year-olds made more accurate causal inferences when the informant provided a certain, more calibrated explanation. In Study 2, children heard similar explanations about probabilistic relations, making the uncertain informant more calibrated. The 5-year-olds were more likely to infer the correct causal relations when the informant was uncertain, but only when the explanation was attuned to the stochasticity of the individual causal events (or outcomes that *sometimes* occur). These findings imply that the capacity to integrate, and make efficient inferences from, distinct sources of knowledge emerges during the preschool years.

Keywords: Certainty, accuracy, testimony, calibration, causal inference, social learning

Children's developing capacity to calibrate the verbal testimony of others with observed

evidence when inferring causal relations

To acquire everyday knowledge about the world, children must discover the underlying causal structure amongst events. Over the last twenty years, there has been renewed interest in causal inference and discovery (e.g., Gopnik & Wellman, 2012). A fundamental question is how children acquire such knowledge. Much of the research in this field has focused on describing the ways in which children recover causal structure from observation and interaction with the world (e.g., Bonawitz, van Schijndel, Friel, & Schulz, 2012; Griffiths, Sobel, Tenenbaum, & Gopnik, 2011). This research has indicated that the ability to learn about the underlying causes of observed events, and to apply that knowledge to reason about the world, emerges early in development (e.g., Gopnik & Sobel, 2000; Kushnir & Gopnik, 2005).

During the same timeframe, there has been a large literature documenting the role of social learning – the ways in which children acquire knowledge from their interactions with other people. Verbal testimony, defined as the communication of a credible claim (Harris & Koenig, 2006), is an important vehicle for the acquisition of abstract concepts across core domains of knowledge (Callanan & Oakes, 1992; Gelman, 2009; Harris, Pasquini, Duke, Asscher, & Pons, 2006). For example, conversations with more expert others can act as epistemic tools for children's exploration, knowledge construction, and long-term learning outcomes in the domain of science (Rowe, 2012; Kurkul & Corriveau, 2017; Tenenbaum, Snow, Roach & Kurland, 2005). In the current study, we explore the role of others' verbal testimony, as well as observed evidence, on children's inferences about the causal efficacy of objects.

Although causal learning and inference is possible from first-hand observation alone, the acquisition of causal knowledge must be facilitated by collaborative exchanges with others in

order for children to appreciate the multifaceted structures in their physical and cultural

environments (Legare, Sobel & Callanan, 2017). For example, young children actively seek out

causal explanations from their parents (Callanan & Oakes, 1992; Frazier, Gelman, & Wellman,

2009). Moreover, variation in the features of adult-child interactions is related to children's

interpretation of observed data (Fender & Crowley, 2007; Luce, Callanan, & Smilovic, 2013),

their assessment of causal interventions (Kushnir, Wellman, & Gelman, 2008) and the scope of

their subsequent exploration (Bonawitz et al., 2011). This has led some researchers to argue that

children use the same mechanisms to make both inferences about causal events and others'

epistemic competence (Sobel & Kushnir, 2013). An open question is how well children can

integrate the way others generate verbal testimony with causal information that is observed in

real time.

Previous research on testimony suggests that children are not passive recipients of the

claims of other people, but regularly gauge the epistemic competence of informants to determine

their reliability as an information source (Corriveau, Meints, & Harris, 2009; Einav & Robinson,

2011). By the age of 4, children are sensitive to a variety of epistemic characteristics of a

potential learning partner, and are able to track an informant's past accuracy and expertise when

learning about novel words and object labels (see Harris, Koenig, Corriveau & Jaswal, 2018;

Mills, 2013 for reviews). For example, Sabbagh and Baldwin (2001) found that 4-year-olds were

more likely to encode novel word-referents from a speaker who conveyed they were

knowledgeable about an object's label, as compared to a speaker who conveyed cues that they

were ignorant.

Children can also integrate social information with their own observations when

evaluating informants. In a recent study, Birch, Severson, and Baimel (2020) showed that 4 and

5-year-olds prefer to learn from a certain informant who had access to knowledge about the

contents of a box over a certain informant who never had access to the box's contents (and

whose verbal confidence was thus not justified). More generally, others' epistemic competence

and the social cues they use to communicate that competence are not always treated equally

when children infer an informant's reliability. Brosseau-Liard, Cassels, and Birch (2014) found

that 5-year-olds were less likely to trust confident, historically inaccurate speakers compared to

hesitant, previously accurate ones. Epistemic competence similarly trumps social characteristics

across a variety of domains (e.g., Corriveau, Kinzler & Harris, 2013; Jaswal & Neely, 2006;

Vanderbourght & Jaswal, 2011, see Sobel & Finiasz, 2020, for a recent metaanalysis).

　　　Such inferences go beyond evaluating simple claims to how children integrate

information they hear from others that *conflicts* with their own causal inferences. For example,

Young, Alibali and Kalish (2012) showed that 5- to 10-year-old children revised their belief

about ambiguous data most often when a peer disagreed with their hypothesis initially and then

generated neutral or disconfirming evidence (see also Kimura & Gopnik, 2019; Macris & Sobel,

2017). Similarly, Bridgers, Buchsbaum, Seiver, Griffiths, and Gopnik (2015) examined how 4-

and 5-year-olds reasoned about an informant's endorsement that was inconsistent with the causal

evidence (the informant endorsed that one object was more likely to activate a machine, while

the evidence suggested that another object was more likely to do so). Children were more

inclined to use an informant's testimony to guide their causal inference when the informant

communicated that they were knowledgeable, as opposed to naïve, about their initial claim.

Studies have also found that children's false belief capacity predicts whether they understand

that a claim someone makes about a causal relation can be false, as opposed to simply believing

that claim in light of data to the contrary (Sobel, 2015; Sobel et al., 2009). Taken together, these

findings suggest that young children are able to effectively evaluate the epistemic value of informant explanations, yet are sensitive to cases where others generate alternate causal interpretations to their own observations.

In none of these cases, however, is it directly considered when it is appropriate for an informant to be uncertain in their causal claim- in particular, registering that the hesitance of verbal testimony might indicate it is not true. Integrating the epistemic strength of verbal testimony with the truth value of the utterance involves *calibrating* the testimony with observed data. As in previous research (e.g., Tenney, Small, Kondrad, Jaswal, & Spellman, 2011), we define calibration as the relative match between the *social cues that indicate a person's confidence* in their claim and the likelihood that their claim is correct. It is possible that children fail to understand such calibration, especially in the face of an uncertain claim. For example, although Birch et al. (2020) found that young children selectively trusted an informant whose confidence was justified, the older children in their sample (i.e., 8-year-olds) did not favor either of the two *hesitant* informants who differed in their visual access of the contents of a box. Tenney et al. (2011) also found that 5- and 6-year-olds did not treat well-calibrated, hesitant information as more reliable than information generated more confidently. In this study, children tended to trust a witness who provided a confident claim about event details that were both accurate and inaccurate (i.e., an informant who was certain, but their explanation was somewhat incorrect), as compared to a calibrated individual who adjusted her verbal certainty based on the quality of the evidence (i.e., the informant had the same accuracy, but was more hesitant in her claims). This tendency was the same for adults who were under cognitive load.

One central issue that has not been addressed to date is that the calibration of uncertainty can be conceptualized in two different ways. One can calibrate the stochasticity of events in the

aggregate (i.e., if I toss a fair coin repeatedly, and guess the outcome each time, I might be right).

Thus, verbal uncertainty represents the general probabilistic nature of being correct. Another way

is to calibrate the stochasticity of the event itself (i.e., if I toss a fair coin repeatedly, sometimes it

will land on heads). The Birch et al (2020) and Tenney et al. (2011) examples, as well as other

investigations of children's appreciation of the probabilistic accuracy of informants (Pasquini et

al., 2007), only considers the first case and not the second. However, investigations of causal

inference suggest that children only begin to appreciate that the events themselves can be

stochastic around age 5 (e.g., Buchanan & Sobel, 2011; Bullock, Gelman & Baillargeon, 1982).

In other words, when younger children observe that X is a cause, they infer that the causal

efficacious relation should always occur; the ability to appreciate that X might or *sometimes*

makes Y occur develops during the preschool years. This suggests that between the ages of 4 and

5, children begin to appreciate uncertain verbal testimony that highlights, and thus is calibrated

to, individual events, even if they cannot appreciate verbal testimony that is calibrated to the

aggregate of events. The present research was motivated by this question.

The Present Research

We extend previous research on young children's learning of causal relations, and

learning from the testimony of others, to explore the developing ability to attune the confidence

with which informants generate verbal testimony (i.e., the degree of verbal certainty) with the

stochastic nature of the observed data described by that testimony (i.e., the likelihood of a

particular outcome). We investigated children's sensitivity to calibrated certain, and uncertain,

explanations when making inferences about novel causal relations. In Study 1, we manipulated

the certainty with which an informant delivered testimony about deterministic causal outcomes.

Informants generated certain or uncertain testimony describing the efficacy of cues that were

100% and 0% effective. Based on previous results illustrating children's evaluation of confident and knowledgeable informants (Birch et al., 2020; Sabbagh & Baldwin, 2001), we hypothesized that children would make more accurate inferences in the calibrated condition; when asked whether the 100% activation cue was more effective or the 0% activation cue was less effective than other probabilistic cues, children would be more accurate when they heard calibrated testimony.

In Study 2, we replicated our certainty manipulation, but more directly tested the distinction between how verbal testimony might interact with reasoning about cues in the aggregate, as opposed to individual efficacy, by presenting children with certain or uncertain explanations about probabilistic evidence. Critically, two different kinds of uncertain testimony were provided. Some children heard uncertain testimony, similar to the Tenney et al. (2011) and Birch et al. (2020) work (e.g., "*Maybe X causes Y*"); other children heard uncertain testimony about the stochastic nature of the cues (e.g., *"Maybe X sometimes causes Y"*). We predicted that, similar to previous calibration studies, young children would not appreciate the epistemic value of uncertainty in the aggregate calibration condition. Yet, by highlighting the stochastic relations present in the data, children might understand uncertainty as it relates to the individual cues. We also tested the possibility that 5-year-olds would outperform younger children in calibrating verbal testimony to observed data, particularly stochastic data.

Study 1

Method

Participants

Forty-eight 4- and 5-year-olds (24 girls, mean age = 59 months, age range = 48 – 71 months) participated in Study 1. The sample size was determined by power analysis, assuming a

large effect size (Cohen's ω = .5; following the effect sizes documented in Bridgers et al., 2015;

Walker et al., 2017, upon which this method is based) and α = .05, based on an χ^2-test with df =

1. The results of this analysis suggested 26 children per condition. We opted to stop data

collection once the counterbalancing requirements were reached, resulting in n = 24 participants

in each condition.

Children were recruited from a local school and a children's museum in the Northeast

region of the United States from March – December 2018. Children were randomly allocated to

one of the two informant conditions. An equal number of 4- and 5-year-olds participated in each

condition.

Materials

Machine. The machine used in the observation phase was similar to those used in

"blicket detector" paradigms from previous studies on young children's causal reasoning abilities

(e.g., Gopnik & Sobel, 2000). The machine employed in the current research was a 20.32cm x

15.24cm x 7.62 cm black plastic box with a lucite plastic top. The box contained blue LED lights

and a small electronic music player, both of which could be remotely operated. One of the

experimenters used a small remote, hidden out of view of the participant, to make the box light

up and play music for trials in which a block activated the machine. This activation would last

until the child retrieved the block from the top of the machine.

Observation phase stimuli. The block stimuli used in the observation and the test phase

were based on that of Walker et al. (2017). We constructed six blocks (50.80mm wooden cubes)

for the evidence trials. Each block had a plastic rectangular 31.8mm x 10.4mm x 8mm Lego

piece affixed to the top and another to the front of it. The top Lego piece was a different color to

the front Lego piece (see illustration of the stimuli in Figure 1, panel A).

Children viewed three causal trials (i.e., block activated the machine) and three inactive trials (i.e., block did not activate the machine). One of the two Lego pieces on each block always represented a deterministic cue for activation (i.e., 100% or 0%). The other piece represented a probabilistic activation cue (i.e., 66% or 33%). For example, in Figure 1 (panel A), blocks with a black piece always activated the machine (the three blocks have black pieces and all activate the machine). Thus, the black piece was the 100% cue. By contrast, blocks with the yellow piece never activated the machine. This piece was the 0% cue. Blocks with the red piece activated the machine 2 out of 3 times and blocks with the white piece activated the machine 1 out of 3 times. These were the 66% and 33% cues respectively. We refer to the 100% and 0% cues as the *deterministic cues* and the 66% and 33% cues as the *probabilistic cues* (the deterministic cues were the focus of Study 1). The colors of the Lego parts associated with the deterministic and probabilistic properties were counterbalanced across participants; however, for the purpose of describing the procedure below, black, yellow, red, and white indicate the 100%, 0%, 66%, and 33% cues respectively.

Small cards (18.6cm x 7cm) were placed at either side of the machine at the beginning of the observation trials to aid children's categorization of each block as causal or inactive. The causal card depicted an image of the machine lit up with a "thumbs up". The inactive card had an image of an inert machine with a "thumbs down" (see Figure 1, Panel A).

Test phase stimuli. Four additional wooden cube stimuli (same size and color as the blocks used in the observation phase) were used for the test phase. These blocks only had a single Lego piece attached to it. Each block had a different colored piece (i.e., black, yellow, red, white). These blocks were inserted, two at a time, into a box apparatus, which we will call the *hiding* box (see Figure 1, panel B). The hiding box was constructed based on the one used in

Walker et al. (2017). The box was a black 22.2cm x 10.15cm x 5.1cm cardboard box had four

rectangular cut-outs, two at the top and two at the front. The cut-out windows were covered by

dark blue colored felt flaps. For each test trial, the experimenter would reveal the single piece on

each block, either by lifting the top or front flaps. One of the flaps for each block remained

closed during the trials. This was done to ensure that children were only making judgements

between two specific cues at a given time but that they believed they were comparing two of the

blocks (with two Legos) that they viewed in the observation phase. The second Lego piece on

the observed learning block therefore appeared hidden by one of the closed flaps in the test trials.

The orientation of the two test cues in the top and/or front positions was randomized.

Procedure

Observation phase. The experimenter brought children into a quiet room at their school

or off of the museum floor. Children at the museum were tested with their parent/guardian

present. Children tested at their school were tested with only the researchers. The experimenter

sat across from the child at a table. A second experimenter – hitherto referred to as the *informant*

– was seated next to the participant. The experimenter first asked the child a series of questions

about themselves for familiarization and warm-up. Then, she introduced children to the machine

by saying, "I have this machine here, and it lights up and plays music".

The experimenter placed a box filled with the block stimuli on the table and said, "I also

have some toys. Some of these toys will make the machine go and some of them will not". She

then invited the informant to provide testimony about the deterministic cues. In the *Certain*

condition, the informant said, "I know! The Black ones [100% cue] make the machine go and the

Yellow ones [0% cue] do not. I'm really sure." In the *Uncertain* condition, the informant said,

"Um, I don't know. Maybe the Black ones [100% cue] make the machine go and maybe the

Yellow ones [0% cue] do not. I'm not really sure." Following the delivery of this information, the experimenter said, "Alright! Now let's try putting one of them on the machine" and proceeded to administer the evidence trials.

The experimenter placed the six blocks on the machine one at a time and children observed whether each block activated the machine. Children always observed one efficacious block and one inactive block first (order counterbalanced). The first efficacious block was placed with the card that indicated activation; the first inactive block was placed with the card that indicated it failed to activate the machine. After the remaining four blocks were placed on the machine, the experimenter asked the children to help her to categorize the block as either one that "makes the machine go" or one that "does not make the machine go". The child then observed and categorized the other four blocks. The remaining four trials were presented in a random order.

After these six trials, the experimenter again asked the informant to provide her testimony about the blocks. The informant repeated the information she had provided before the beginning of the observation phase. Note that children were able to view all of the evidence while they heard the explanation – the three causal blocks were grouped to one side of the machine and the three inactive blocks were grouped to the other side of the machine. After delivering the second round of testimony, the informant left the room and was not present for the test phase.

Memory checks. After the informant's departure, the experimenter removed the machine, blocks, and category cards from the table. She then administered the three memory check questions. She first asked, "Do you remember which one she [the informant] said would make the machine go?" Following children's response, E asked, "Do you remember which one she said would not make the machine go?" Finally, E asked about the child's perceived certainty of the

testimony (i.e., "Do you remember, was she really sure or not really sure?"). The experimenter did

not provide corrective feedback to the responses. If children failed to provide the correct response

for all three memory check questions, they were excluded from the analyses.

Test phase. After the memory check questions, the experimenter placed the hiding box

on the table. She explained, "This is a hiding box! I can put the blocks in here, and lift up these

flaps and show you a part of the block. I am going to do that now and ask you about the parts

you can see, okay?". When placing the two test blocks in the box for each trial, she would use a

manila folder to obscure the child's view of the process.

Children were presented with six test trials in a random order. For every trial, the

experimenter showed children a pair of cues on different blocks and asked them to indicate

which block was more likely to activate the machine. For example, when shown that one block

has a black cue and the other block has a yellow cue, children were asked "Do you think the

black one or the yellow one makes the machine go?". Children were asked to infer the more

causally predictive cue from a combination of every cue they had observed in the learning phase

(i.e., 100% or 0%, 100% or 66%, 100% or 33%, 66% or 0%, 33% or 0%, 66% or 33%). Five out

of the six test trials involved at least one cue that was specified in the informant's testimony - the

deterministic 100% or 0% cue. Responses to these five trials were the main focus of the

analyses[1].

---

[1] One trial involved the comparison of two cues that were not part of the informant's testimony - the comparison between the 66% and 33% probabilistic cues. We realized that, based on our current predictions, it would be difficult to draw any meaningful inferences regarding performance on this trial across conditions and thus excluded it from the following analyses.
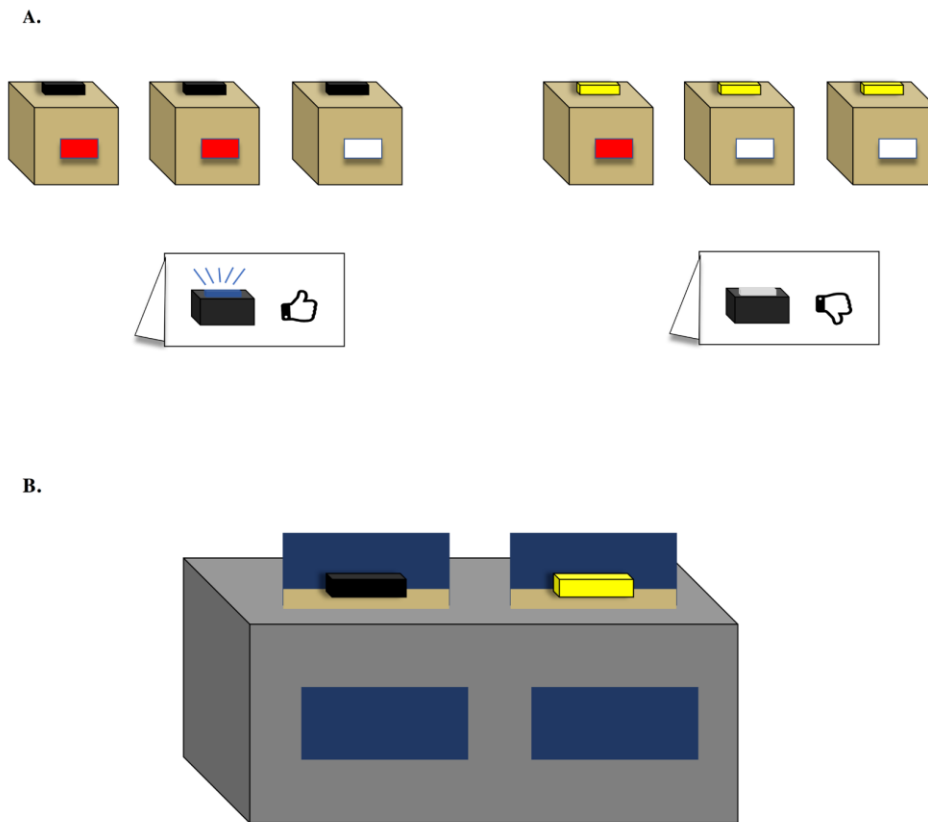
Figure 1. The block stimuli and memory cards used for the observation phase (panel A) and the hiding box apparatus used in the test phase (panel B).

Results and Discussion

Preliminary Analyses

The number of children who correctly recalled the color of the deterministic cues identified by the informant (approximately 89% of responses) and the certainty of the informant (73% of responses) during the memory check phase is reported in Appendix A (see Table S1 for the breakdown by condition and question type). One child did not provide correct responses for any of the three memory check questions and was excluded for the remainder of the analyses.

For each test trial, children were given a score of 1 if they chose the block with the cue that indicated a higher likelihood of activation, otherwise they were given a score of 0. Using

mixed-effects logistic regression models, the preliminary analyses revealed that there was no

effect of the counterbalancing variables (order of first learning block, color associated with the

predictive cue; all p's > .10), participant variables (gender, testing site; both p's > .87) nor trial

type (all p's > .23) on children's causal judgements in the test phase. We did not consider these

variables further.

Main Analyses

We conducted a mixed-effects binomial logistic regression model using the *glmer*

function of the *lme4* package in R statistical software (version 3.4.2) to explore the effect of

Informant Condition (categorical predictor: Certain vs. Uncertain) and Age (categorical

predictor: 4-year-old vs. 5-year-old) on whether children chose the more predictive cue on each

test phase question[2]. The models included Informant Condition and Age as fixed effects and

participant ID as a random effect to account for variability of individual responses to the test

trials. We entered Informant Condition in a first step, Age in a second step, and, in a third model,

we added the interaction between Informant and Age. All of the data files are openly available at

https://osf.io/raqh3/?view_only=6b2c171ed8924d828beae3da67724732

The final model revealed a significant main effect of Age, $\beta = 2.21$, $SE = 0.83$, $z = 2.04$, p

= .008, and significant Informant Condition x Age interaction, $\beta = -2.86$, $SE = 1.09$, $z = -2.62$, p

= .009. To investigate the interaction further, we ran two separate mixed-effects logistic

regression models within each age group. The results showed a significant main effect of

Informant Condition among the 5-year-olds, $\beta = -2.83$, $SE = 0.99$, z = -2.42, p =.02, OR = 0.09,

95% CI = [0.01, 0.64], indicating that children were more likely to identify the correct causal

---

[2] All of the reported interaction effects hold when age is entered as a continuous predictor in the main models (see Figure S1 and S2 in Appendix B for visualization of the significant interactions).

cues in the *Certain* condition compared to the *Uncertain* condition (see Figure 2). There was no

effect of Informant Condition on 4-year-olds' responses, $\beta = 0.60$, *SE* = 0.62, z = 0.98, *p* =.33.
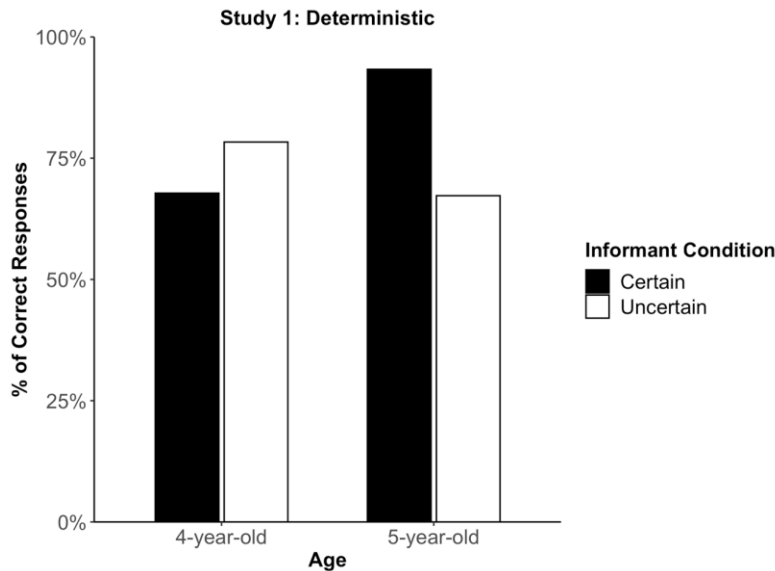

Figure 2. The proportion of test trials that children chose the more causally predictive cue as a function of informant condition and age.

The ability to appropriately integrate the verbal testimony of others and observable

evidence when making causal inferences about deterministic data was evident among the older

children. Despite some research suggesting that 4-year-olds show preferences for confident and

knowledgeable informants (Birch et al., 2020; Sabbagh & Baldwin, 2001), the younger children

in Study 1 did not show sensitivity to the relative match between the verbal certainty cues and

the accuracy of the data. The findings of Study 1 extends previous work by showing that hearing

others confidently state information that is consistent with observed data fostered judgements

among 5-year-olds about the likelihood that causal features are efficacious.

An important open question is whether children in this age range are also sensitive to a

speaker's confidence when they generate verbal information that is not about deterministic

causes but rather are probabilistic, or are indicative of outcomes that do not always occur. Here,

uncertain verbal testimony could support children's causal inferences because it is calibrated to the unreliability of the associated causal data (e.g., *"Maybe X causes Y"*). By contrast, generating certain testimony about a probabilistic cue is inaccurate because there is an alternative, deterministic causal cue. If children are capable of calibrating the strength of the testimony to probabilistic data, they should make more accurate causal inferences when such testimony is hesitant as opposed to certain. However, based on previous findings suggesting that young children might not recognize the truth value of uncertain claims (Birch et al., 2020; Tenney et al., 2011), we also tested whether testimony calibrated to individual stochastic events (e.g., *"Maybe X sometimes causes Y")* would lead to more accurate causal inferences.

In Study 2, children were presented with the same causal inference paradigm outlined in Study 1. In this study, we manipulated the certainty with which the informant delivered an explanation about the two probabilistic data cues - children either heard a certain or uncertain explanation about the 66% and 33% cues. We included two uncertain conditions. In one of these conditions, the uncertain informant (like the informant in the certain condition) provided testimony about the data in the aggregate (e.g., "Maybe the Red one [66%] makes the machine go"). In the other condition, the uncertain informant acknowledged the stochastic nature of the data (e.g., "Maybe the Red one *sometimes* make the machine go"). We anticipated that children might be particularly sensitive to the calibrated testimony when the uncertain informant provided an accurate description of the probabilistic outcomes, or in the *Uncertain + Sometimes* condition.

Study 2

Method

Participants

Seventy-two 4- and 5-year-olds (31 girls, mean age = 60 months, age range = 48 – 72 months) were recruited to participate in Study 2. Children in the *Certain* and *Uncertain* conditions were recruited from the same Study 1 locations between March - December 2018. The children in the *Uncertain + Sometimes* condition were recruited at a later time point from the same two locations between June – August 2019. There was an equal number of 4- and 5-year-olds in each condition. Four children were excluded in the final analyses because they did not pass the memory check phase (see below).

Materials and Procedure

The materials used were identical to that of Study 1. The observation and test phases were also similar to Study 1. In the observation phase, the key difference was the content of the informant's testimony that children heard before and after observing the experimenter place the six learning blocks on the machine. In the *Certain* condition, the informant said: "I know! The Red ones [66% cue] make the machine go and the White ones [33% cue] do not. I'm really sure." In the *Uncertain* condition, the informant said: "Um, I don't know. Maybe the Red ones [66% cue] make the machine go and maybe the White ones [33% cue] do not. I'm not really sure". In the *Uncertain + Sometimes* condition, the informant said: "Um, I don't know. Maybe the Red ones [66% cue] sometimes make the machine go and maybe the White ones [33% cue] sometimes do not. I'm not really sure".

In all conditions, children were then asked to infer the more causally predictive cue from a combination of every cue they had observed during the learning trials (i.e., 66% or 33%, 100%

or 66%, 100% or 33%, 66% or 0%, 33% or 0%, 100% or 0%)[3]. Note that, unlike Study 1, there

is a conflict for children between the correct answer according to the observed data and the cue

that is mentioned in the testimony in two of these trials (conflict trials: 100% or 66%, 33% or

0%). For example, in order to choose the more predictive cue in the 100% or 66% test trial,

children would need to discount the testimony that the 66% cue is efficacious and favor the

deterministic 100% cue in their response.

Results and Discussion

Preliminary Analyses

The number of children who correctly recalled the color of the probabilistic cues

(approximately 78% of responses) and the certainty of the informant (64% of responses) during

the memory check phase are reported in Appendix A (see Table S2 for the breakdown by

condition and question type). Four children did not provide correct responses for any of the three

memory check questions and were excluded for the remainder of the analyses.

Responses to the test questions were scored in the same manner as Study 1. Mixed-

effects logistic regression models revealed that there was no effect of the counterbalancing

variables (order of first learning block, color associated with the predictive cue; all *p*-values>

.19), nor participant variables (gender, testing site; both *p*'s > .74) on correct responses to the test

trials in Study 2. There was a main effect of Trial Type: overall, children were more likely to

choose the more accurate causal cue on the trials that were generally consistent with the

testimony provided by the informant (66% or 33%, 100% or 33%, 66% or 0%) compared to the

two trials that pose a potential conflict between the testimony and observed evidence (100% or

---

[3] As in Study 1, one of the test trials involved children's inference of two cues that were not identified in the verbal testimony (i.e., the comparison between the 100% and 0% deterministic cues). We decided to drop this trial and only focus on the five trials that related to the testimony, and thus to our research hypothesis, in the main analyses.

66% and 33% or 0%; all *p*-values < .003). We retained this variable in the following models to

control for this significant main effect, see below.

Main Analyses

We conducted a mixed-effects binomial logistic regression model using the *glmer*

function of the *lme4* package in R statistical software (version 3.4.2) to examine the effect of

Informant Condition (categorical predictor: Certain vs. Uncertain vs. Uncertain + Sometimes)

and Age (categorical predictor: 4-year-old vs. 5-year-old) on whether children inferred the more

predictive cue in the test trials. The models included Informant Condition and Age, and their

interaction, as fixed effects and participant ID as a random effect to account for variability of

individual responses in the test phase.

The results yielded (with the Certain condition as the reference level) a significant

Informant Condition x Age interaction, $\beta = 1.68$, $SE = 0.67$, $z = 2.52$, $p = .012$, for the

comparison between the *Certain* and *Uncertain + Sometimes* condition. To check the comparison

between the two Uncertain conditions, we then defined the *Uncertain + Sometimes* condition as

the reference level in the model. The results showed a significant main effect of Age, $\beta = 1.50$,

$SE = 0.49$, $z = 3.10$, $p = .002$, and a significant Informant Condition x Age interaction, $\beta = -1.51$,

$SE = 0.66$, $z = -2.28$, $p = .022$, for the comparison between the *Uncertain* and *Uncertain +*

Sometimes condition. Thus, the effect of Age differed in the *Uncertain + Sometimes* condition

compared to both the *Certain* and *Uncertain* conditions. There was no such difference between

the *Certain* and *Uncertain* conditions.

To examine this interaction further, we conducted separate mixed-effects logistic

regression models within each age group. There was a significant effect of Informant Condition

among the 5-year-olds: children in the *Uncertain + Sometimes* condition were more likely to

choose the more accurate causal cue in comparison to the 5-year-olds in the *Certain* condition, β = -1.06, *SE* = 0.49, z = -2.18, *p* =.029, OR = .35, 95% CI = [.13, .90], and *Uncertain* condition, β = -1.07, *SE* = 0.48, z = -2.24, *p* =.025, OR = .34, 95% CI = [.14, .84] (see Figure 3).

Recall that children were invited to make decisions about some trials that were consistent with the informant's testimony (66% or 33%, 100% or 33%, 66% or 0%) and two trials where the more accurate causal cue could potentially be in conflict with the testimony (100% or 66% and 33% or 0%). The results of this model, that included the variable of Trial Type, suggested that 5-year-olds were generally less likely to choose the causally correct cue on the two conflict trials in comparison to each of the three other trials (all *p*'s < .004).

There was no effect of Informant Condition among the younger children (see Figure 3). There was a significant effect of trial type in this age group however: 4-year-olds were also less likely to infer the causally correct cue in the conflict trials in comparison to the three consistent trials (all *p*'s < .03[4]).

---

[4] The comparison between the 66% or 0% trial (consistent trial) and the 100% or 66% trial (conflict trial) was not significant among this age group (*p* = .21).
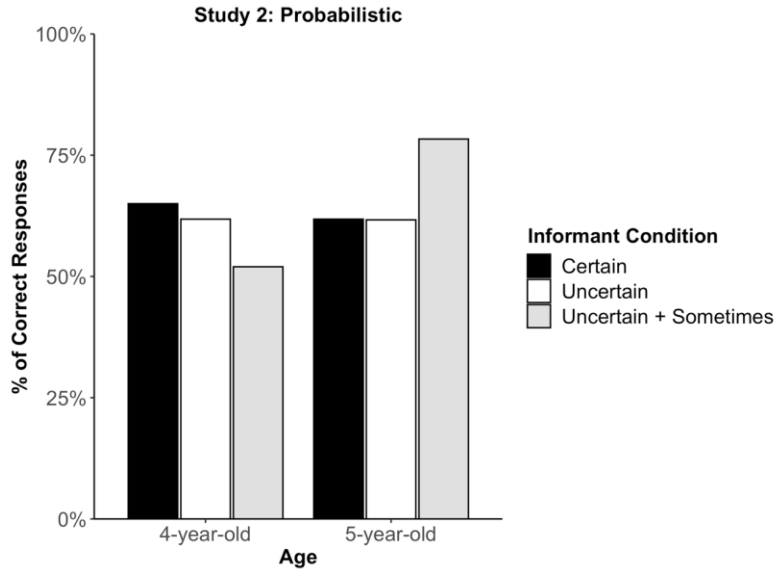
Figure 3. The proportion of test trials that children chose the more causally predictive cue as a function of informant condition and age.

Consistent vs conflict trials. The results suggest that 4- and 5-year-old children tended to

use the testimony to guide their judgements over their own observations of the causal evidence

when both sources of knowledge were in conflict. Yet, the older children showed more accurate

causal inferences in the *Uncertain + Sometimes* condition in comparison to the two other

conditions (whereas the younger children exhibited similar levels of learning across the three

informant conditions). To further understand the observed boost in older children's causal

inferences in this condition, we ran separate analyses on the consistent and conflict trials among

this age group. There were no significant comparisons between the *Uncertain + Sometimes* and

the two other conditions (both p's > .30) for 5-year-olds' performance on the consistent trials

(66% or 33%, 100% or 33%, 66% or 0%). In contrast, 5-year-olds were significantly more likely

to choose the causally correct cue in the conflict trials in the *Uncertain + Sometimes* compared to

the *Uncertain* condition, $\beta = -1.41$, *SE* = 0.51, z = -2.79, p =.005, OR = 0.25, 95% CI = [0.09,

0.66], and *Certain* condition, β = -1.14, *SE* = 0.51, z = -2.23, *p* =.03, OR = 0.32, 95% CI = [0.12,

0.87].

The ability to integrate, and to appropriately discount, informant testimony about

probabilistic cues in favor of more causally predictive evidence was evident among the 5-year-

olds in this sample. Importantly, the boost in children's causal inferences emerged when the

informant's claim most accurately represented the probabilistic causal information (i.e., when the

informant was uncertain, but accurate about the stochastic nature of the data and said "Maybe it

sometimes makes the machine go") as opposed to information about the cues in the aggregate.

General Discussion

The results of the two studies suggest that children's capacity to calibrate verbal

testimony with first-hand observations when learning about novel causal relations emerges

during the preschool years. Five-year-olds, but not 4-year-olds, demonstrated the ability to attune

the verbal certainty of an explanation to the predictive accuracy of causal data. . In Study 1, the

5-year-olds showed greater accuracy when they heard a certain explanation about deterministic

outcomes, as compared to an uncertain explanation. In Study 2, we observed a similar

developmental trend; five-year-old children learned more effectively about the nature of the

probabilistic cues after hearing an uncertain explanation. Importantly, the older children's causal

inferences relied on the calibrated informant not only conveying verbal cues to uncertainty about

the outcomes, but also providing an accurate explanation about probabilistic events (or outcomes

that *sometimes* occur). Further analysis of 5-year-olds' causal inferences in the test phase

suggested that children who heard an explanation about the stochastic nature of the probabilistic

cues were more likely to appropriately *discount* those cues in favor of more causally predictive

evidence. Taken with the age-related change described above, the present studies offer novel

insights into the effects of informant calibration on children's causal understanding in a potential

learning environment.

The finding that 5-year-olds successfully drew from both the testimonial and observed

deterministic evidence in Study 1 complements previous findings on children's epistemic

evaluations of confident informants (e.g., Birch et al., 2020; Brosseau-Liard et al., 2014), albeit

with some minor differences. For example, Birch et al. (2020) found that 4- and 5-year-olds

selectively trust an informant whose verbal confidence positively correlated with their

knowledge access. Here, we extend these findings to show that 5-year-olds in our study made

more accurate inferences about causal relations from an informant whose verbal certainty was

justified. Furthermore, the results of Study 1 demonstrate the potential benefits of explanation for

children's causal learning outcomes (Walker et al., 2017). When an adult provided a causal

explanation that was consistent with children's first-hand observations, both based on accuracy

and level of confidence, 5-year-olds were likely to endorse the efficacious causal relations.

The pattern of results in Study 2 makes a novel contribution to studies exploring

children's understanding of uncertain informants. The older children in our studies were more

likely to use uncertain testimony calibrated to individual probabilistic events to facilitate their

causal inferences. An interesting question is what motivates the development of this capacity for

calibration. One possibility is that the results can be explained by children's developing capacity

to understand stochastic causal relations in their environment (Buchanan & Sobel, 2011; Bullock

et al., 1982), and thus verbal testimony that highlights such relations scaffolds their

understanding of the causal system.

Another plausible explanation for the observed age-related changes depends on children's

recognition of the motivation behind others' testimony – their mental states. For instance, Sobel

et al. (2009) found that 3- to 5-year-olds made correct inferences about probabilistic data when

an informant provided social cues that she expected the data to be deterministic. When given an

object that made the machine go 2/3 times, she expressed surprise on the trial that it failed to

activate the machine and given an object that made the machine go 1/3 times, she expressed

surprised on the trial that it activated the machine. The study showed that, without the

informant's surprise cues, children could not infer which object was more likely to activate the

machine, whereas when the cues were provided, only children who passed a standard false belief

measure could do so, controlling for age. This result suggests that children's understanding of

others' mental states could be related to their ability to integrate information generated by others

with interpretations of observed data.

There was some indication that children were more sensitive to cues to certainty than

cues to uncertainty. The older children in our sample were not at ceiling in the accurate,

uncertain condition in Study 2; five-year-olds still sometimes interpreted the uncertain testimony

in this condition at face value and chose the probabilistic cues over the deterministic evidence on

the relevant trials. Further, in the memory check phase, children were generally more likely to

recognize verbal certainty (or when the informant was "really sure") than verbal uncertainty (or

when the informant was "not really sure"), and correctly recall all of the testimonial pieces of

information when the confidence of the informant was consistent with, and endorsed,

deterministic outcomes (in the Certain/Deterministic condition). Although previous work

suggests young children show systematic differences in their behavior on the basis of verbal

epistemic cues (e.g., Bridgers et al., 2015; Sabbagh & Baldwin, 2001), one possibility is that the

younger children in our sample might not have the metacognitive skills to explicitly reflect on

the certainty of the informant (Ruffman, Rustin, Garnham & Parkin, 2001). Another possibility

is that the current phrasing of the memory check questions did not directly tap into children's understanding of a person's level of confidence. Further research is necessary to discern between these two possibilities.

Children in the present study heard an explanation about the causal system both prior to and after viewing the evidence. Because the informant provided an explanation before the observation phase, this may have primed children to pay attention to the relevant cues, and potentially lead them to weigh the testimony more heavily than the first-hand evidence. This might be particularly true in Study 2 where the informant's explanation was not consistent with the general causal structure of the objects (she ignored the more obvious, deterministic evidence). Future research should explore whether children's causal judgements would differ if they had the opportunity to observe the evidence before hearing a claim about that evidence. For example, Decker et al. (2015) asked 6- to 12-year-olds to learn probabilistic relations between a stimulus and a reward (either positive or negative), and then introduced them to informants who generated false information about those relations. The children were more likely to weigh their own observations over erroneous verbal instruction when making judgements about the probability of an outcome. While Decker et al.'s (2015) study used a different age range and paradigm from the present study, incorporating these findings does suggest that hearing verbal testimony before observing data could promote the calibration pattern that we posit here. Further research should investigate this issue. It would also be important to test whether children's causal inferences predict how they independently interact with the objects, and their generalizations to novel causal systems.

To conclude, the present research set out to address a gap in the literatures on the acquisition of children's causal knowledge from their own observations of the world (Gopnik, &

Wellman, 2012) and from their interactions with other people (Harris et al., 2018; Mills, 2013;

Sobel & Kushnir, 2013). Our results show the nuanced interaction between the verbal and

observed information that children are privy to when inferring causal relations. By the age of 5,

children were able to attune subtle differences in the social cues that people use to convey their

epistemic competence to the accuracy of those claims when making causal discoveries. More

generally, these findings suggest that the capacity to integrate disparate sources of evidence

emerges relatively early in development.

References

Birch, S., Severson, R. L., & Baimel, A. (2020). Children's understanding of when a person's confidence and hesitancy is a cue to their credibility. *PloS One, 15*, e0227026. doi: 10.1371/journal.pone.0227026

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition, 120*, 322–330. doi: 10.1016/j.cognition.2010.10.001

Bonawitz, E.B., van Schijndel, T., Friel, D., & Schulz, L. (2012). Balancing theories and evidence in children's exploration, explanations, and learning. *Cognitive Psychology, 64*, 215-234. doi: 10.1016/j.cogpsych.2011.12.002

Bridgers, S., Buchsbaum, D., Seiver, E., Griffiths, T. L., & Gopnik, A. (2016). Children's causal inferences from conflicting testimony and observations. *Developmental Psychology, 52*, 9–18. doi: 10.1037/a0039830

Brosseau-Liard, P., Cassels, T., & Birch, S. (2014). You seem certain but you were wrong before: Developmental change in preschoolers' relative trust in accurate versus confident speakers. *PloS One, 9*, e108308. doi: 10.1371/journal.pone.0108308

Buchanan, D. W., & Sobel, D. M. (2011). Mechanism-based causal reasoning in young children. *Child Development, 82*, 2053-2066. doi: 10.1111/j.1467-8624.2011.01646.x

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

Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development, 7*, 213–233. doi: 10.1016/0885-2014(92)9001

Corriveau, K. H., Kinzler, K. D., & Harris, P. L. (2013). Accuracy trumps accent in children's endorsement of object labels. *Developmental Psychology, 49*, 470-479. doi: 10.1037/a0030604

Corriveau, K. H., Meints, K., & Harris, P. L. (2009). Early tracking of informant accuracy and inaccuracy. *The British Journal of Developmental Psychology, 27*, 331–342. doi: 10.1348/026151008x310229

Decker, J. H., Lourenco, F. S., Doll, B. B., & Hartley, C. A. (2015). Experiential reward learning outweighs instruction prior to adulthood. *Cognitive, Affective & Behavioral Neuroscience, 15*, 310–320. doi: 10.3758/s13415-014-0332-5

Einav, S., & Robinson, E. J. (2011). When being right is not enough: Four-year-olds distinguish knowledgeable informants from merely accurate informants. *Psychological Science, 22*, 1250–1253. doi: 10.1177/0956797611416998

Fender, J. G., & Crowley, K. (2007). How parent explanation changes what children learn from everyday scientific thinking. *Journal of Applied Developmental Psychology, 28*, 189–210. doi: 10.1016/j.appdev.2007.02.007

Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult-child conversation. *Child Development, 80*, 1592–1611. doi: 10.1111/j.1467-8624.2009.01356.x

Gelman, S. A. (2009). Learning from others: Children's construction of concepts. *Annual Review of Psychology, 60*, 115–140, doi: 10.1146/annurev.psych.59.103006.093659

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

Gopnik A, & Wellman H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*, 1085-1108. doi: 10.1037/a0028044

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*, 1407-1455. doi: 10.1111/j.1551-6709.2011.01203.x

Harris P. L., & Koenig M. A. (2006). Trust in testimony: How children learn about science and religion. *Child Development, 77*, 505–524 doi: 10.1111/j.1467-8624.2006.00886.x

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology, 69*, 251-273. doi: 10.1146/annurev-psych-122216-011710

Harris P. L., Pasquini, E. S., Duke S., Asscher J. J., & Pons F. (2006). Germs and angels: The role of testimony in young children's ontology. *Developmental Science, 9*, 76–96. doi: 10.1111/j.1467-7687.2005.00465.x

Jaswal, V. K., & Neely, L. A. (2006). Adults don't always know best: Preschoolers use past reliability over age when learning new words. *Psychological Science, 17*, 757-758. doi: 10.1111/j.1467-9280.2006.01778.x

Kimura, K., & Gopnik, A. (2019). Rational higher-order belief revision in young children. *Child Development, 90*, 91-97. doi: 10.1111/cdev.13143

Kurkul K., & Corriveau K. H. (2017). Question, explanation, follow-up: A mechanism for
    learning from others? *Child Development, 89*, 280-294. doi:10.1111/cdev.12726

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

Kushnir, T., Wellman, H. M., & Gelman, S. A. (2008). The role of preschoolers' social
    understanding in evaluating the informativeness of causal interventions. *Cognition, 107*,
    1084–1092. doi: 10.1016/j.cognition.2007.10.004

Legare, C. H., Sobel, D. M., & Callanan, M. (2017). Causal learning is collaborative: Examining
    explanation and exploration in social contexts. *Psychonomic Bulletin & Review, 24*,
    1548–1554. doi: 10.3758/s13423-017-1351-3

Luce, M. R., Callanan, M. A., & Smilovic, S. (2013). Links between parents' epistemological
    stance and children's evidence talk. *Developmental Psychology, 49*, 454–461. doi:
    10.1037/a0031249

Macris, D. M., & Sobel, D. M. (2017). The role of evidence diversity and explanation in 4-and 5-
    year-olds' resolution of counterevidence. *Journal of Cognition and Development, 18*,
    358-374. doi: 10.1080/15248372.2017.1323755

Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from
    others. Developmental Psychology, 49, 404–418. doi: 10.1037/a0029500

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the
    relative accuracy of informants. *Developmental Psychology, 43*, 1216-1226. doi:
    10.1037/0012-1649.43.5.1216

Row, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-
    directed speech in vocabulary development. *Child Development, 83*, 17–74, doi:
    10.1111/j.1467-8624.2012.01805.x

Ruffman, T., Rustin, C., Garnham, W., & Parkin, A. J. (2001). Source monitoring and false
    memories in children: relation to certainty and executive functioning. *Journal of
    Experimental Child Psychology, 80*, 95–111. doi: 10.1006/jecp.2001.2632

Sabbagh, M. A., & Baldwin, D. A. (2001). Learning words from knowledgeable versus ignorant
    speakers: links between preschoolers' theory of mind and semantic development. *Child
    Development, 72*, 1054–1070. doi: 10.1111/1467-8624.00334

Sobel, D. M. (2015). Can you do it? How preschoolers judge whether others have learned.
    Journal of Cognition and Development, 16, 492-508. doi:
    10.1080/15248372.2013.815621

Sobel, D. M., & Finiasz, Z. (2020). How children learn from others: An analysis of selective
    word learning. *Child Development*. doi: 10.1111/cdev.13415

Sobel D. M., & Legare C. H. (2014). Causal learning in children. *Wiley Interdisciplinary
    Reviews: Cognitive Science, 5*, 413-427. doi: 10.1002/wcs.1291

Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: how children evaluate the reliability of
    testimony as a process of rational inference. *Psychological Review, 120*, 779–797. doi:
    10.1037/a0034191

Sobel, D. M., Sommerville, J. A., Travers L. V., Blumenthal E. J., & Stoddard E. (2009). The
    role of probability and intentionality in preschoolers' causal generalizations. Journal of
    Cognition and Development, 10, 262-284. doi: 10.1080/15248370903389416

Tenenbaum, H.R., Snow, C.E., Roach, K.A., & Kurland, B. (2005). Talking and reading science: Longitudinal data on sex differences in mother-child conversations in low-income families. *Applied Developmental Psychology, 26*, 1–19. doi: 10.1016/j.appdev.2004.10.004

Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., & Spellman, B. A. (2011). Accuracy, confidence, and calibration: How young children and adults assess credibility. *Developmental Psychology, 47*, 1065–1077. doi: 10.1037/a0023273

VanderBorght, M., & Jaswal, V. K. (2009). Who knows best? Preschoolers sometimes prefer child informants over adult informants. Infant and Child Development: An International *Journal of Research and Practice, 18*, 61-71. doi: 10.1002/icd.591

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child Development, 88*, 229–246. doi: 10.1111/cdev.12590

Young, A. G., Alibali, M. W., & Kalish, C. W. (2012). Disagreement and causal learning: Others' hypotheses affect children's evaluations of evidence. *Developmental Psychology, 48*, 1242–1253. doi: 10.1037/a0027540

Supplementary Information

Appendix A. Children's Performance on the Memory Check Questions

Table S1. *The number (and percentage) of children who correctly recalled the deterministic cues and the certainty of the informant by condition in Study 1.*

	Informant Condition			
	Certain		Uncertain	
	n	%	*n*	%
Recall of 100% cue	23	95.83	18	75.00
Recall of 0% cue	22	91.67	22	91.67
Recall of certainty	22	91.67	13	54.17

Table S2. *The number (and percentage) of children who correctly recalled the probabilistic cues and the certainty of the informant by condition in Study 2.*

	Informant Condition					
	Certain		Uncertain		Uncertain 'Sometimes'	
	n	%	*n*	%	*n*	%
Recall of 66% cue	21	87.50	18	75.00	14	58.33
Recall of 33% cue	22	91.67	21	87.50	16	66.67
Recall of certainty	15	62.50	15	62.50	16	66.67

Appendix B. The Proportion of Correct Test Responses as Function of Informant

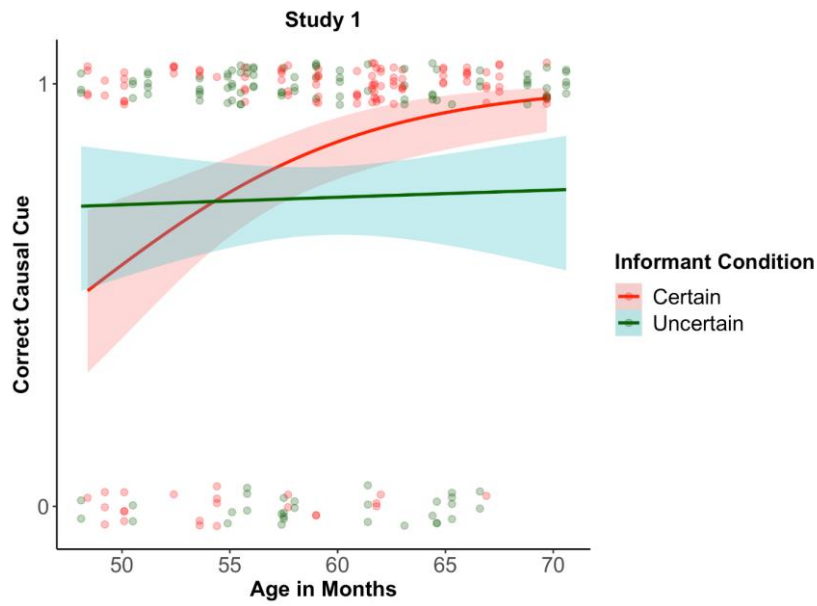**Condition and Age (Continuous Predictor) in Studies 1 and 2**

Figure S1.

Figure S2.