Developmental shifts in children's and adults' use of informant accuracy and calibration in testimonial reasoning

Rebekah A. Gelpí[1,2], Sophie Bridgers[3,*], Amy Whalen[4,*], and Daphna Buchsbaum[5, 1]

[1]Department of Psychology, University of Toronto

[2]Schwartz Reisman Institute for Technology and Society, University of Toronto

[3]Department of Psychology, Stanford University

[4]School of Biology, University of St. Andrews

[5]Department of Cognitive and Psychological Sciences, Brown University

Author Note

Rebekah A. Gelpí https://orcid.org/0000-0001-6899-0520

We have no known conflicts of interest to declare.

* All work was completed prior to SB joining DeepMind and AW joining Amazon.

Correspondence concerning this article should be addressed to Rebekah A. Gelpí, Department of Computer Science, Whiting School of Engineering, Malone Hall, 340 N. Charles St., Baltimore, MD 21218. Email: rebekah@jhu.edu

Abstract

This study investigates how adults as well as 3- and 4-year-old children integrate testimony and physical evidence when evaluating informant reliability in a causal inference task. Previous research indicates that young children are sensitive to informants past accuracy and expressed certainty but are less consistent in evaluating whether an informants certainty is calibrated appropriately to their actual knowledge, also known as *self-knowledge*. Across five studies, we examine the development of children's use of others' self-knowledge on a causal inference task where informants expressing differing levels of certainty endorsed one of two blocks as effective for making a machine play music, but physical evidence contradicted this endorsement, favouring the other block, comparing children's responses to adults and to a computational model of causal testimonial learning. We find that while adults' inferences are most compatible with a model that captures both the previous accuracy of informants as well as the calibration of informants' confidence, 4-year-olds are better captured by a model incorporating informants' past accuracy but not their self-knowledge, while 3-year-olds show only limited evidence of tracking accuracy. These findings suggest a developmental shift in the ability to reason not only about whether informants are correct, but about whether their confidence appropriately reflects their underlying knowledge and epistemic access.

Keywords: causal reasoning, social learning, testimony, self-knowledge

Developmental shifts in children's and adults' use of informant accuracy and calibration in testimonial reasoning

Introduction

Imagine that you are trying to solve a set of unfamiliar puzzles, and you ask two friends about how to solve the first one. Your friends differ in their confidence – one friend is sure of his solution, while the other admits she isn't sure, and guesses the solution. If both of your friends are wrong, then your friends have been shown to be equally inaccurate. But if both friends indicate they are sure of the solution to the second puzzle, you might be more likely to try the solution that your friend who was previously guessing proposes, because she was self-aware of her own lack of knowledge about the first puzzle, while the other friend was not only incorrect, but inappropriately confident in his own knowledge.

This example highlights that when people learn from social testimony, they often draw inferences not only about the topic of the testimony (e.g., how to solve the puzzle) but also about the reliability of the informants themselves (Buchsbaum, Bridgers, et al., 2012; Tenney et al., 2011). In the same way that others' behaviour can be attributed to personal traits or to the situations that individuals are in (e.g., Cao et al., 2024; Seiver et al., 2013), there can be multiple reasons why an informant's testimony might be accurate or inaccurate. Some informants might have more knowledge than others, or individuals might sometimes be in a situation where they know something that another individual happens not to know (Juteau et al., 2025; Schmid et al., 2024). When informants are inaccurate, it might be because they are making reasonable assumptions given their knowledge, are overconfident given their lack of knowledge, or even that they are intentionally trying to deceive a listener (Palmquist & Kondrad, 2024; Palmquist et al., 2022; Pozzi & Mazzarella, 2024; Ronfard & Lane, 2019).

In this paper, we explore how children and adults reason about a somewhat less explored property that informants can exhibit, namely their *calibration*—whether an informants expressed confidence reliably tracks their knowledge (and the accuracy of their

testimony). A calibrated informant knows when they do or do not know the answer to a question, while an uncalibrated informant might be underconfident (always indicating a lack of certainty, even when they are correct) or overconfident (always indicating that they know the answer, even when their answer turns out to be incorrect). Even when informants have been equally accurate (or inaccurate) in the past, a learner would prefer to obtain information from a well-calibrated informant, as it is more likely that the informant will be correct when the informant indicates confidence in their answer.

Here, we extend prior work on the topic of calibration, investigating the developmental trajectory of children's and adults' use of information about not only an informant's accuracy and expressed confidence, but also the calibration of an informant's statements about their degree of confidence. By replicating and extending a causal testimonial paradigm introduced by Buchsbaum, Bridgers, et al. (2012) and Bridgers et al. (2016), we provide converging evidence that adults are willing to extend the benefit of the doubt to informants who have acknowledged their own uncertainty, moreso than to informants who have been confidently wrong. We also include new experiments with 3- and 4-year-old children, age groups that differ in their sensitivity to epistemic cues, including a novel two-informant task for 4-year-olds, and formally compare children's and adults' behaviour to a computational model of testimonial causal learning, evaluating the extent to which each age group's choices are best fit by models that include tracking of informants' accuracy, confidence, and calibration. Together, these findings clarify an emerging developmental shift in children's use of cues to informants' information quality and self-knowledge when they must integrate social and physical evidence to jointly reason about people and causal systems.

The Development of Selective Trust in Testimony

Although children are generally highly trusting of others from an early age (Jaswal et al., 2010), when children learn from the testimony of those around them, they must learn to use this information *selectively*, understanding that some kinds of informants are

better than others, that some cues are more reliable than others, and that more testimony is not always better (Gelpí & Buchsbaum, 2024; Gweon, 2021; Harris et al., 2018; Kendal et al., 2018). For example, children are more likely to endorse the testimony of a familiar source than an unfamiliar one (Corriveau, Harris, et al., 2009), an authority over a subordinate (Bernard et al., 2016), an in-group member over an out-group member (Kinzler et al., 2011), domain experts over novices (Koenig & Jaswal, 2011; Kushnir et al., 2013), someone with direct access to information over someone reporting hearsay (Aboody et al., 2022; Butler et al., 2018; Gelpí, Whalen, et al., 2025), and majorities over dissenters (Corriveau, Fusaro, & Harris, 2009; Haun & Tomasello, 2011; Walker & Andrade, 1996), although how children weigh these different factors varies considerably across development (Tong et al., 2020) and culture (Enesco et al., 2016; Sebastián-Enesco et al., 2020).

Children can also use their own observations to reason about causal relationships (for a review, see Goddu & Gopnik, 2024). By the time children are 3 years old—and in some cases even in infancy—children can evaluate information about the causal efficacy of multiple different objects and infer properties of the objects such as their category or causal strength (Kushnir & Gopnik, 2005, 2007; Waismeyer et al., 2015).

In addition to balancing information about multiple causes, however, children also must account for the fact that they may encounter testimony from others that conflicts with their own observations. This simultaneous social and causal inference is challenging, because it is consistent with two possibilities: on one hand, the child could be receiving inaccurate and thus unreliable testimony, while on the other hand the child's own observations might be incomplete. Three-year-olds often defer to informant testimony even when it conflicts with their own direct observations, while 4- and 5-year-olds tend to rely on their own observations (Jaswal et al., 2014; Ma & Ganea, 2010), even when this sometimes means disregarding the testimony of a majority of individuals (Gelpí, Otsubo, et al., 2025).

As children begin to not only integrate social testimony with their own observations to reason about causal events, but also use information derived from their causal inference

to reason about the properties of individuals (see e.g., Buchsbaum, Seiver, et al., 2012), they can use their own observations and prior knowledge to reason about informant qualities such as reliability. Thus, when an informant has been shown to be previously inaccurate—either by providing labels that conflict with a child's existing knowledge, or by being shown to be incorrect—children are less likely to endorse the claims of this informant, and reduce their estimates of the informant's knowledge (Birch et al., 2008; Hermansen et al., 2021; Koenig & Harris, 2005; Koenig et al., 2004; Pasquini et al., 2007; Ronfard & Lane, 2019). Recent work shows that children also adapt their evidentiary standards based on prior informational reliability. Orticio et al. (2024) found that 4–7-year-olds exposed to detectable inaccuracies subsequently engaged in more extensive fact-checking before accepting novel claims. Nevertheless, young children may still endorse the testimony of inaccurate informants when other sources of information are not available (Vanderbilt et al., 2014) or costly to acquire (Brosseau-Liard, 2014), suggesting that children do not rely on informants' history of accuracy alone when deciding whether to extend trust to an informant.

Children's Understanding of Confidence and Calibration

In addition to cues such as accuracy and expertise, an important cue that can predict whether potential informants possess relevant knowledge is an informant's confidence. As a cue to the reliability of testimony, even very young children recognize nonverbal cues and behaviours (such as shrugging or head tilting for an uncertain informant) that correspond to others' confidence; for example, 2-year-old children use this information to selectively learn about object labels (Birch et al., 2010) and imitate novel actions (BrosseauLiard & PoulinDubois, 2014). By 3 to 4 years old, children additionally use informants' professed confidence in the form of verbal statements, such as "I know which one is the blicket", to determine whether to attend to a speaker's label for a toy (Sabbagh & Baldwin, 2001).

Despite this, children's preference for confident informants exhibits substantial

between-individual variability, and is also only moderately stable within individual children over time, suggesting that situational factors may also play a role in when children preferentially learn from confident informants (Juteau et al., 2019). Likewise, just as children scrutinize the accuracy of informants and selectively learn from those shown to be reliable, children exhibit an increasing tendency with age to disregard the endorsement of a confident informant when it conflicts with the predictions of a previously accurate but hesitant informant (Brosseau-Liard et al., 2014; Fobert et al., 2024; Jaswal & Malone, 2007). Related work (McLoughlin et al., 2021) has shown that preschoolers can integrate informants expressed certainty with causal evidence in blicket-style tasks, particularly when uncertainty is appropriate given probabilistic causal structure, although this paradigm did not require children to evaluate informants calibration across a history of accuracy.

Nevertheless, children's evaluation of when confidence is misplaced does not always align with adults' intuitions. For example, while children aged 7 years and older are more likely to think that individuals who exhibit intellectual humility—i.e., those who do not exhibit excessive confidence in their beliefs when it is not warranted—are nicer, more knowledgeable, and better sources to learn from (Bowes et al., 2025; Hagá & Olson, 2017), younger children do not consistently prefer intellectually humble to intellectually arrogant informants. Further, unlike older children, 5- and 6-year-old children defer to informants even when they express confidence about "unknowable" facts (such as how many leaves are on all of the trees in the world), while only older children reserve such deference for informants who express confidence about "knowable" facts (such as how many bones are in a rabbit's body; Kominsky et al., 2016).

One potential explanation for this development in children's evaluation of confidence is that it reflects children's emerging awareness of others' metacognitive processes. Although children can leverage some information about their own knowledge, in particular their own degree of uncertainty, to guide behaviours such as question-asking and exploration by ages 2-3 (Baer & Kidd, 2022), evidence for the relationship between

4-year-old children's own metacognitive abilities and the selectivity of their social learning is less clear (Baer et al., 2021; Dutemple et al., 2023; Resendes et al., 2021).

Similarly, while there is converging evidence that adults use information about the calibration of informants to determine their credibility (Stanciu & Fiser, 2022; Tenney et al., 2008), the evidence that young children do so is somewhat more mixed, with some studies finding that 3- to 6-year-old children do not use it (Bridgers et al., 2016; Fobert et al., 2024; Tenney et al., 2011) and other findings showing that children as young as 4 attend to informants' calibration (Birch et al., 2020). However, there is also evidence that children's testimonial learning is selective in some key ways that might facilitate their reasoning about calibration under some circumstances. For example, in a set of experiments by Kushnir and Koenig (2017), 4-year-old children were willing to endorse the testimony of an informant who previously professed ignorance, but not one who was previously confidently wrong. This suggests that when tracking a history of accuracy, children may distinguish between informants based on differing degrees of awareness of their own lack of knowledge.

Bridgers et al. (2016) provided a particularly relevant test of these ideas in a blicket detector paradigm that required children to integrate an informant's endorsement of a potential cause with their expressed confidence, and with conflicting directly observed evidence (the block the informant endorsed was always observed to in fact be less causally efficacious). Four-year-old children showed sensitivity to both confidence and the strength of the directly observed evidence, but they did not reliably penalize information from informants who had previously been confidently incorrect more than informants who were previously incorrect but had explicitly expressed their uncertainty. Taken together, it remains unclear whether preschool-age children are able to evaluate the calibration of informants' confidence relative to their knowledge, or whether they are primarily relying on confidence as a more general clue to reliability, with an understanding of calibration emerging only later in the school-age years.

If preschool-age children are aware not only that certain behaviours and words are associated with more knowledge, but have a deeper representation of the relationship between informants' professed confidence and their accuracy—i.e., that informants with a better metacognitive awareness of their own knowledge states are more likely to be accurate when they express confidence than informants with poor metacognition—then they should treat the testimony of informants who were confidently incorrect differently from that of informants who guessed incorrectly.

Nevertheless, even if they do not necessarily do so in the same ways as adults, shifts in children's choices across development may provide insight into how children's representation of others' calibration changes over time. By comparing children's and adults' behaviour to the predictions of a computational model of reasoning from social testimony, we can provide a clearer picture of the extent to which children's choices are consistent with a model that evaluates others' testimony in light of their likely metacognitive awareness, in comparison to a simpler model that treats confidence or hesitance from an informant similarly, irrespective of the calibration of the informant's past testimony. Thus, across five studies, we investigate the development of 3- and 4-year-old children's use of others' self-knowledge as a cue to reliability, comparing their choices to those of adults, and propose a computational model that captures the developmental trajectory of children's and adults' use of informants' calibration in their reasoning about social testimony.

Model of Testimonial Causal Learning

Many models of causal learning (e.g., Griffiths et al., 2011; Perfors et al., 2011) have argued that both children and adults often update their beliefs about causal relationships in ways consistent with Bayes' rule: $P(h|d) \propto P(d|h)P(h)$, such that people update their prior degree of belief in a hypothesis h proportional to the likelihood of observing data d conditioned on that hypothesis being true. These models have also been extended to capture not only how people reason about physical evidence, but also about social testimony (e.g., Goodman & Frank, 2016; Shafto et al., 2014). These models propose that

people interpret the utterances of others through the assumption that others are attempting to communicate rationally and efficiently, providing learners with an inductive framework to draw richer conclusions than simple observation of data would provide. Here, we adopt a Bayesian framework as a normative benchmark to characterize the inferences implied by different representational assumptions about how informants generate their testimony.

For example, if a learner is faced with a machine with three buttons (A, B, and C) that might activate it, being provided with physical evidence that A activates the machine does not intrinsically change the probability that B activates the machine. However, if an informant states that they are confident that B activates the machine, and their testimony is incorrect, the informant's testimony about C may be less reliable. Similarly, when people combine physical evidence and social testimony, they have the opportunity to evaluate the quality of the testimony itself. An informant whose testimony aligns poorly with the physical evidence an observer has encountered might be considered to have less reliable testimony in the future. However, in the context of calibration, a previously inaccurate informant might still provide good testimony in the future if they acknowledged that their prior incorrect testimony was a guess, while a previously inaccurate informant who professed confidence in their answer might reflect the informant having a poor sense of their own knowledge, reducing the reliability of their future testimony.

Here, we extend an existing hierarchical Bayesian model of learning from a combination of physical evidence and social testimony (Figure 1; see also Buchsbaum, Bridgers, et al., 2012). Learners in this model obtain information about a causal system C both from informants who differ in past accuracy and expressed confidence, and from their own observations. Our model is defined in terms of observed variables representing causal outcomes, statements by an informant about the causal strengths of potential causes, and about her level of certainty about her causal knowledge. The model also has hidden variables representing the actual causal strengths of the potential causes, the informants general level of knowledgeability, her specific knowledge of the individual causes, and her

level of self-knowledge–how well she knows what she knows. We capture the complex relationships among these variables in a graphical model (see Figure 1), and describe them in more detail below.

Each cause $c \in C$ has a corresponding causal power W_c that corresponds to an effect $\omega_c = (\rho \cdot W_c) + (1 - \rho)(1 - W_c)$, where ρ is the relatively high probability that an object with high causal efficacy will almost always generate an effect. The global prior over the proportion of objects with high or low causal efficacy is governed by the parameter γ, such that $P(\omega_c = \rho) = \gamma$ and $P(\omega_c = 1 - \rho) = 1 - \gamma$, where γ. As a result, when an object is tested for N trials and succeeds at eliciting a causal effect x times, the success rate can be determined by:

$$P(E_c|\omega_c) = \omega_c^x (1 - \omega_c)^{N-x} \tag{1}$$

This model assumes that people have a prior κ over the proportion of individuals within a population that are generally knowledge $g \in \{0, 1\}$; individuals in the population are generally knowledge $P(g = 1)$ with probability κ or, generally unknowledgeable $P(g = 0)$ with probability $1 - \kappa$. Individuals have specific knowledge about a particular cause K_c depending on their general knowledge and the free parameter τ, such that:

$$P(K_c = 1|G) = \begin{cases} \tau & G = 1 \\ 1 - \tau & G = 0 \end{cases} \tag{2}$$

Where the parameter τ captures the intuition that someone who is generally knowledgeable is likely to also know about a specific cause (but occasionally may be ignorant about a specific example), while someone who is generally unknowledgable is less likely to know about any specific case, but may still occasionally have knowledge. When an informant has specific knowledge about a given cause, the informant will provide a correct report about the causal strength of that cause $P(R_c = W_c|K_c = 1) = 1 - \varepsilon$, with a small probability ε of misreporting. If the informant does not have local knowledge, the report

will be a guess $P(R_c = W_c|K_c = 0) = 0.5$.

Finally, an informant's testimony also includes a statement about the informant's level of certainty Q_c (e.g., "I'm certain" or "I'm just guessing"). Here, we extend the model in Buchsbaum, Bridgers, et al. (2012) by modelling learners' treatment of certainty as indicative of knowledge conditional on calibration, rather than as an unconditional cue. An informant's reported level of certainty depends on their specific knowledge of this cause K_c and on their self-knowledge (calibration) S. If an informant possesses accurate self-knowledge, meaning they know what they do and don't know, then their expressed confidence will align with their specific knowledge such that $P(Q_c = K_c|s = 1) = 1 - \delta$, with a small probability δ of misreporting their confidence. If an individual does not possess accurate self-knowledge, their expressed confidence will be poorly calibrated to their true knowledge, and thus they will randomly express high or low confidence in their report $P(Q_c|s = 0) = 0.5$. The global prior η determines the proportion of individuals who possess self-knowledge and thus express appropriately calibrated confidence.

After observing physical evidence and informant testimony, learners compute the joint probability of each cause c and potential data $\mathcal{D} = \{E, R, Q\}$ given the global variables G, S:

$$L_c(\cdot|G, S) = \sum_{K_c, W_c} P(W_c)P(K_c|G)P(R_c|K_c, W_c)P(Q_c|K_c, S)P(E_c|W_c) \tag{3}$$

and the posterior over global variables G and S given $\mathcal{D}$:

$$P(G, S|\mathcal{D}) \propto P(G)P(S) \prod_{c \in C} L_c(\cdot|G, S) \tag{4}$$

and marginalize over the posterior of G and S to infer the posterior probability for each cause:

$$P(W_c = 1|\mathcal{D}) = \sum_{G,S} P(G, S|\mathcal{D})P(W_c = 1 \mid G, S, \mathcal{D}) \tag{5}$$

Lastly, we assume that learners endorse a cause proportional to the posterior predictive expectation:

$$\begin{aligned} P(\text{choose } c) &\propto \mathbb{E}[\omega_c|\mathcal{D}] \\ &\propto P(W_c = 1|\mathcal{D}) \cdot \rho + P(W_c = 0|\mathcal{D}) \cdot (1 - \rho) \end{aligned}$$

Lesioned Model

Since children might struggle to represent informants' calibration, we also consider a series of lesioned models which represent how a learner who does not track past accuracy or general knowledgabilty, or one who does not track calibration (or both), should normatively reason in this same scenario. To do this, we remove the terms G and S as well as the free parameters κ and η, representing informants' general knowledge and self-knowledge respectively. A learner lacking the general knowledge term would assume that knowing about one cause is not correlated to knowing about other causes, while a learner lacking the self-knowledge term believes that informants only express confidence when they really know an answer.

When the model lacks the general knowledge term, specific knowledge depends only on the τ parameter, i.e. $P(K_c) = \tau \cdot (K_c) + (1 - \tau) \cdot (1 - K_c)$. Similarly, when the model lacks the self-knowledge term, the informant's confidence in their report Q_c depends only on the informant's local knowledge:

$$P(Q_c) = \begin{cases} Q_c = K_c & 1 - \delta \\ Q_c \neq K_c & \delta \end{cases} \tag{6}$$

Model Evaluation

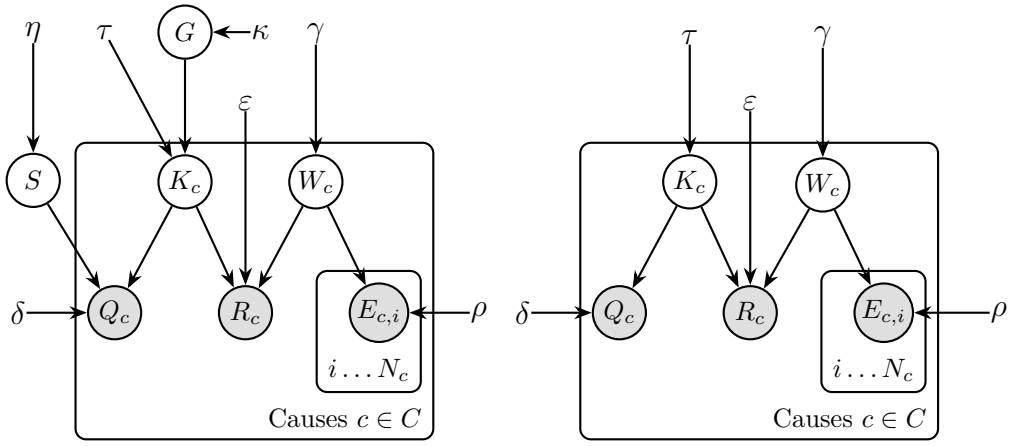
We present five experiments that test the extent to which which adults and children attend to the calibration of informants' confidence when integrating physical evidence and testimony. In order to evaluate the extent to which the model predictions are able to

generalize to new experimental settings, we first fit the models to adults' and children's choices on one experiment, and use the fitted parameters to estimate the fit on other experiments. As the lesioned models are nested within the full model, we compare the model fits using the Bayesian information criterion (BIC), penalizing the additional complexity of the full model according to the number of free parameters.

The BIC values can then be used to compute an approximate Bayes factor between models using the formula $\log(\mathrm{BF}_{01}) = 0.5 \cdot (\mathrm{BIC}_0 - \mathrm{BIC}_1)$. Using the criteria of Kass and Raftery (1995), a natural log Bayes factor of $\log(\mathrm{BF}_{01}) \geq 1.15$ is needed to provide meaningful positive evidence of M_1, $\log(\mathrm{BF}_{01}) \gtrapprox 2.5$ for strong evidence of M_1, and $\log(\mathrm{BF}_{01}) \gtrapprox 5$ for decisive evidence in favour of model M_1.

Across all models, we fixed the parameters $\rho = 0.95$ (corresponding to a prior that most causes are highly efficacious) and $\varepsilon = 0.01$ (corresponding to a prior that knowledgeable people rarely misspeak), fitting the others to the data. To compute the best-fitting parameter values for each model, we use an implementation of the limited-memory Broyden-Fletcher-Goldfarb-Shannon algorithm (L-BFGS), bounding δ (how often calibrated people misstate their confidence) between 0.0001 and 0.5, γ (how common efficacious causes are) between 0.0001 and 0.9999, and bounding general knowledge, specific knowledge and self-knowledge terms (τ, κ, and η) between 0.5 and 0.9999 (as values below 0.5 can have unintuitive implications, e.g. when $\tau < 0.5$, the model assumes that people are less likely to have specific knowledge when they have general knowledge than when they do not).

In the following experiments, we present scenarios motivated by this model to adults and children in order to evaluate to what degree their choices reflect the use of information about the accuracy of informants' testimony, as well as the calibration of informants' expressed confidence in their testimony. By formalizing the predictions of models of varying complexity, we can evaluate which models most closely reflect the choices of adults and children and use this information to draw conclusions about adults' and



(a) *Full model* (b) *Lesioned model*

Figure 1

Causal graphical model for the full model (a, left) and lesioned model (b, right) of physical evidence and testimony in Experiments 1–5. In the lesioned model, the nodes for G and S and hyperparameters κ and η are removed, as it is assumed that all agents are fully calibrated and that having specific knowledge about one cause is not informative about general knowledge.

children's testimonial reasoning capacities.

Experiment 1

To investigate how reasoning about the calibration of informants' expressed confidence changes across development, our first experiment adapts the task of Bridgers et al. (2016) to adults. In this task, participants were introduced to a causal system (a machine) and two potential causes (blocks). Participants first received testimony from an informant who expressed either high confidence ('I know...') or uncertainty ('I'm guessing...') regarding which block was effective at making the machine light up and play music. Subsequently, participants observed statistical evidence that conflicted with the informant's claim. We measured participants' causal judgments at three time points: (1) after the informant's testimony but before observing the physical evidence (Prior), (2) after observing the physical evidence (Causal), and (3) after the informant returned to provide

high confidence testimony about the causal efficacy of a novel pair of blocks (Generalization). Critically, in the causal phase, the previously uncertain informant is well calibrated (they said they were guessing and turned out to be incorrect), while the previously confident informant is poorly calibrated (they said they knew, but the physical evidence contradicts their testimony). The generalization phase therefore provides a test of whether participants treat informants' prior calibration as evidence of their self-knowledge, and therefore informative for future trust in their confident testimony.

Across conditions, the conflicting physical evidence was either deterministic (one block always activated the machine and the other never did) or probabilistic (one block activated the machine more often than the other), allowing us to vary whether incorrect testimony could plausibly be attributed to stochastic causal structure. For example, when a confident informant has provided testimony that conflicts with probabilistic evidence, it is plausible that the pattern of evidence is simply "unlucky", and thus our model predicts that a learner would continue to assign some probability that the informant is still well-calibrated. However, when their claims are met with conflicting deterministic evidence, this is much less likely, and our model predicts that a learner would treat that informant as poorly calibrated. Given that prior work with adults has shown that they are sensitive to the calibration of informants, this experiment will allow us to test the predictions of the model. By fitting the parameters to this experiment, we can additionally evaluate the extent to which the model can generalize to subsequent experiments, providing converging predictive validity to the model's predictions.

Methods

Participants and Design

296 U.S. participants were recruited from Amazon's Mechanical Turk service and were compensated $0.50 for completing the task. Participants were required to have 50 successful HITs with an approval rating of over 95% in order to be eligible for the task. Participants were assigned to one of four between-subjects conditions: confident

deterministic condition ($N = 74$), confident probabilistic condition ($N = 74$), naive deterministic condition ($N = 71$), or naive probabilistic condition ($N = 77$). This yielded a 2 x 2 factorial design, with the informant's reported knowledge (confident or naive) and the evidence that people encountered (conflicting deterministic data or conflicting probabilistic data) as factors.

Materials and Procedure

Participants completed an online survey task where two cartoon characters, an informant and an assistant, presented information about a machine. The informant introduced a green box with a black top as a machine that was capable of lighting up and playing music when certain objects were placed on it. She introduced two different blocks and explained that one block almost always activated the machine (the *endorsed* block) and another block almost never did (the *unendorsed* block). In the Confident conditions, the informant claimed she knew which block was better at activating the machine, while in the naive conditions, the informant stated that she was just guessing as to which block was better. After introducing the blocks, the informant stated that she needed to leave and that her assistant would continue the experiment. The informant then left, and was not present during the demonstration of the physical evidence.

Before observing any evidence, the assistant asked participants which one of the two blocks was more likely to make the machine activate (the *prior* choice). Then, participants saw the assistant demonstrate the blocks, which exhibited different patterns of causal efficacy, depending on whether they were in one of the Deterministic or Probabilistic conditions.

In the Deterministic conditions, the endorsed block activated the machine 0/6 times, while the unendorsed block activated the machine 6/6 times. In the Probabilistic conditions, the endorsed block activated the machine 2/6 times, while the unendorsed block activated the machine 2/3 times. The two blocks thus had an identical number of activations, but the unendorsed block had a higher observed probability of activating the

machine than the endorsed block (as in Kushnir & Gopnik, 2007). Participants were not given any explanation for the variability in the machines behavior or the observed outcomes of the blocks. After observing this evidence, participants were again asked to choose which block was more likely to activate the machine (the *causal* choice).

Finally, in the generalization phase, the original informant returned and presented two new unobserved blocks. In both conditions, she stated that she *knew* that one of the two blocks almost always activated the machine and that the other block never did. After the informant left again, the assistant asked participants which block was more likely to activate the machine (the *generalization* choice).

The task used four distinct blocks (a green rectangle, blue cylinder, pink disk, and orange square), as in Bridgers et al. (2016), which were presented in fixed pairs, one pair in the conflict condition and the second pair in the generalization condition. Across participants, which pair was shown first, which block within the pair was endorsed by the informant, and the leftright position of the blocks was counterbalanced.

Results and Discussion

The results from Experiment 1 are presented in Figure 2. In order to investigate the effects of confidence, causal evidence, and informants' self-knowledge (calibration) on adults' endorsements of the informants' claims, we first conducted a Bayesian mixed-effects logistic regression, with the reported informant knowledge (confident or naive), conflicting evidence (probabilistic or deterministic), and within-subject choice type (prior, causal, or generalization) as predictors of participants' endorsement of the informant. Because of the large number of potential interactions between variables, we instantiate the model with a horseshoe prior with $df = 1$ to penalize large numbers of non-zero estimates, and follow Kruschke (2018) in setting the boundaries of the regions of practical equivalence (ROPE) to [-0.18, 0.18] for model coefficients.

Overall, participants' choices reflected a sensitivity to informants' initial confidence (Figure 2). Before observing evidence (in the prior rating), participants were more likely to

select a block endorsed by a confident informant(97.3%) rather than a naive informant (78.6%) ($OR = 9.91$, 0.00% in ROPE). They were also much less likely to select the informant's endorsed block after observing conflicting evidence ($OR = 0.0047$, 0.00% in ROPE). Participants were more likely to select a block endorsed by a confident informant (20.3%) than a naive informant (1.4%) even after observing conflicting evidence; this effect was driven by a willingness to continue to endorse the confident informant's testimony in the Probabilistic condition ($OR = 17.71$, 0.00% in ROPE), while participants in the Deterministic condition dismissed the testimony irrespective of condition ($OR = 7.40$, 0.35% in ROPE).

Lastly, participants' choices in the generalization phase reflected a sensitivity to the calibration of informants. In the Probabilistic condition, there was no conclusive difference between participants' chosen block in the generalization phase when a confident or naive informant had endorsed the less effective block ($OR = 1.32$, 28.10% in ROPE). Conversely, in the Deterministic condition, participants were less likely to choose the initially confident informant's endorsed block than the initially naive informant's endorsed block ($OR = 0.36$, 0.00% in ROPE), suggesting that the evidence participants observed in the Deterministic condition was sufficiently strong that the best explanation for the expressed confidence of the informant is that she had poor self-knowledge, leading her to state that she was confident about an answer even when she lacked knowledge about the causal system. This contrasts with the Probabilistic condition, where the data is sufficiently ambiguous that there is still a possibility that the confident informant's claim was appropriate or at least understandable, such that the informant's expressed confidence could be well-calibrated; the fact that some proportion of participants continued to endorse a confident informant in the Probabilistic condition is consistent with participants extending her the benefit of the doubt.

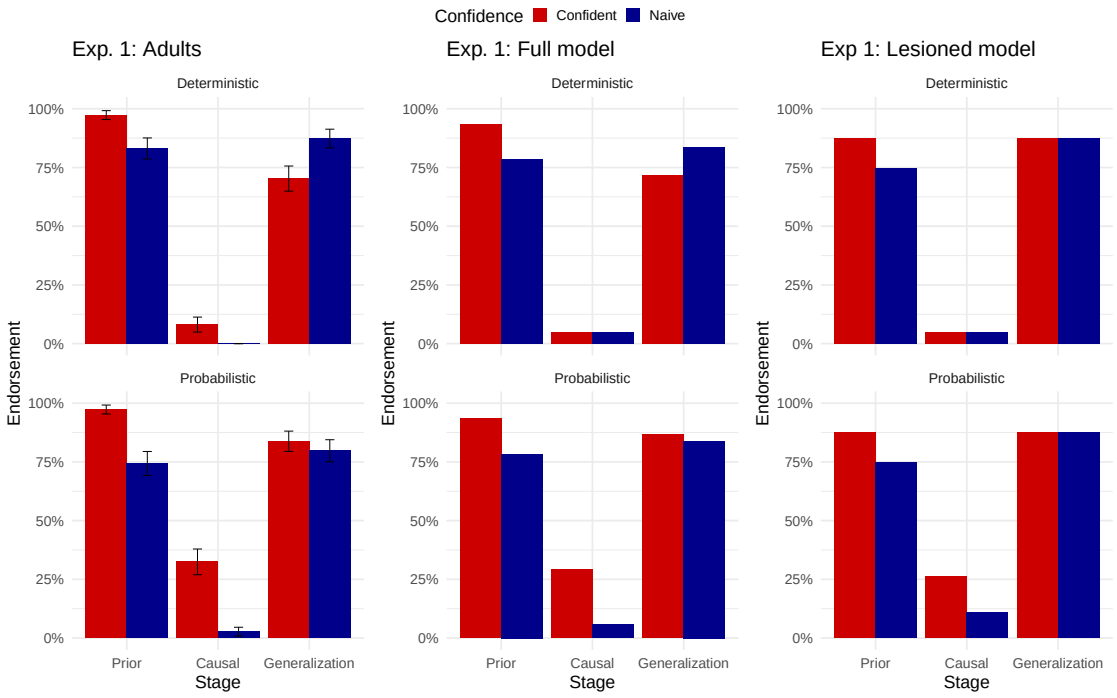


Figure 2

Empirical human data (left) and model predictions for the full model (middle) and lesioned model without global knowledge or self-knowledge (right) for Experiment 1. In the Deterministic condition (top row), adults are more likely to disregard the testimony from the previously confident informant, consistent with the predictions of the full model, but not the lesioned model. In the Probabilistic condition (bottom), some adults maintain the possibility that the confident informant is correct after observing the conflicting physical evidence, as the confident informant's claim is unlikely but still possible.

Model Comparison

Qualitatively, these findings align with our model's predictions that adults use informants' testimony to infer not only the accuracy of an informant's claim but also update their beliefs about the informant's level of self-knowledge, treating the testimony of poorly calibrated informants as less informative than the testimony of informants who had guessed incorrectly. The strength of the evidence also determined the willingness that people had to continue to endorse the claims of a confident informant, such that when

participants observed probabilistic evidence that contradicted a confident informant's claim, they were more willing to continue to rely on the informant's testimony in a new environment.

To provide a quantitative evaluation of the model's ability to capture the behaviour of adults on this task, we compute the best-fitting parameters for each model (full model, global knowledge only, self-knowledge only, and lesioned) and compare the performance of all four models. The full model (BIC = 676.11) was the best performing model, outperforming the model including only global knowledge ($\log \mathrm{BF}_{01} = 18.55$), the model including only self-knowledge ($\log \mathrm{BF}_{01} = 14.71$), and the lesioned model ($\log \mathrm{BF}_{01} = 25.41$), providing decisive evidence in favour of the full model relative to the other models, suggesting that despite the greater complexity of the full model, its greater explanatory power allows it to capture a broader element of how participants reasoned about the testimony in Experiment 1.

The best-fitting free parameter values for the full model corresponded to $\gamma = 0.61$, $\delta = 0.01$, $\tau = 0.96$, $\kappa = 0.98$, and $\eta = 0.80$, but a range of parameter values provide comparably good fit to the data. This corresponds to a world in which most people are globally knowledgeable, and most people who are globally knowledgeable are locally knowledgeable, but that a substantial minority of people are poorly calibrated and thus liable to appear overly confident. In addition to their poorer fit, the lesioned models make less implausible assumptions about certain parameter values. For example, the model lacking global knowledge assumes that poor calibration is relatively widespread ($\eta = 0.64$), while the model lacking self-knowledge and the model lacking both self-knowledge and global knowledge assume that calibrated people misspeak very frequently ($\delta = 0.27$ and 0.29, respectively).

Experiment 2

In Experiment 1, we found that adults appropriately adjusted their evaluation of the causal strength of objects according to both testimony from informants of varying

certainty and probabilistic or deterministic causal evidence that conflicted with the claims of the informants. In Experiment 2, we expand the scope of our experiment to evaluate settings where the informants' testimony is correct, allowing us to examine whether people's endorsements in the generalization condition, which were still generally high, are reduced relative to a case where the informants' testimony was consistent with the observed evidence. By using the best-fitting parameters from Experiment 1, this further allows us to test the replicability of the model predictions and its ability to capture and generalize the predictions of participants on whose data the model is not directly fitted.

Methods

Participants and Design

300 U.S. participants were recruited from Amazon's Mechanical Turk service and were compensated $0.50 for completing the task. Participants were required to have 50 successful HITs with an approval rating of over 95% in order to be eligible for the task. 23 participants were excluded due to failing an attention check question, yielding 277 participants for the final sample. Participants were assigned to one of four between-subjects conditions: confident consistent condition ($N = 64$), confident conflicting condition ($N = 70$), naive consistent condition ($N = 74$), or naive conflicting condition ($N = 69$). This yielded a 2 x 2 factorial design, with the informant's reported knowledge (confident or naive) and the evidence that people encountered (consistent or conflicting with the informant's testimony) as factors.

Materials and Procedure

The procedure for Experiment 2 was largely the same as Experiment 1, with the following changes. First, in the *consistent* evidence conditions, the evidence that participants observed in the causal phase was consistent with the informant's testimony (i.e., the block endorsed by the informant activated the machine more than the unendorsed block). In the *conflicting* evidence conditions, the unendorsed block was more effective, as in Experiment 1. The pattern of evidence shown to participants during the causal phase

was always deterministic, so the confident and naive conditions with conflicting evidence corresponded to the confident deterministic and confident naive conditions in Experiment 1.

Secondly, rather than providing a forced choice response to which block was more likely to make the machine activate, participants provided a rating from 0 to 10 of how likely each block was to activate the machine. Although we solicited adults' judgments using forced choice ratings in Experiment 1 to allow us to directly compare to children's performance (in Bridgers et al., 2016, and in Experiment 3, below), the greater range of potential responses with a rating task might allow adults to show more nuanced, finer-grained differences between their responses in different conditions of the experiment.

Results and Discussion

To assess the effects of the reported knowledge state of the informants and whether the evidence conflicted with the testimony, we used a mixed-effects linear regression with informants' reported knowledge (confident or naive), evidence condition (consistent or conflicting), and rating type (prior, causal, generalization) as predictors of participants' block ratings (measured as the difference in ratings between the informant's endorsed block and the unendorsed block).

First, we conducted an omnibus ANOVA, finding significant main effects of informants' reported knowledge ($\chi^2(1) = 86.55$, $p < .0001$) and the rating type ($\chi^2(2) = 706.90$, $p < .0001$). Further, the effects of reported knowledge and pattern of evidence interacted with the rating type, and a three-way interaction between reported knowledge, pattern of evidence, and rating type also emerged (all $\chi^2(2) > 47.54$, $p < .0001$).

To investigate these interactions, we conducted a series of follow-up comparisons. Unsurprisingly, given participants had not yet observed any evidence, participants in both evidence conditions had higher ratings for the informant's endorsed block when the informant indicated that they were confident than when the informant indicated they were naive (both $t(802) > 7.58$, $p < .0001$). However, when participants observed the causal evidence, they provided slightly higher ratings to the block endorsed the confident

informant than that of the naive informant ($t(802) = 4.20$, $p < .0001$) when the evidence conflicted with the informant's initial claim, although the ratings were very low overall for both informants. When the evidence was consistent with the informant's initial claim, there was no difference between the block endorsed by the confident vs. naive informants ($t(802) = -0.38$, $p = .70$), strongly preferring the endorsed block in both cases.

Critically, while participants continued to show no difference in their generalization rating between the initially confident and initially naive informants when the evidence was consistent with the informant's initial claim ($t(802) = 0.70$, $p = .48$), participants gave *lower* ratings to the block endorsed by the initially confident informant than to the initially naive informant when the data conflicted with the informant's initial claim ($t(802) = 7.22$, $p < .0001$).

Furthermore, when the informant was initially confident, participants were less likely to endorse the block chosen by the informant in the generalization trial when the causal evidence was inconsistent was the informant's initial claim than when the evidence was consistent ($t(802) = –9.593$, $p <.0001$), while this effect was only marginal for the initially naive informant ($t(802) = –1.877$, $p = .061$), suggesting that seeing conflicting evidence reduced participants' perceptions of the informant's self-knowledge and global knowledge, particularly when the informant had been confidently wrong about a prior claim.

Model Comparison

To evaluate participants' rating choices in light of the model, which reflected choice probabilities, we assume that participants maximized, always choosing the block with the higher rating and choosing randomly when blocks were equally rated, rather than probability matching (i.e., choosing blocks in proportion with their rating) as all models were fit much better by assuming that participants maximized (all log $BF_{01} \geq 118.07$).

Even though the parameters were not fit to the data from this experiment, participants on this task were very well fit by the full model's predictions (Figure 3). We once again found that the full model's predictions (BIC = 486.89) were a better fit to

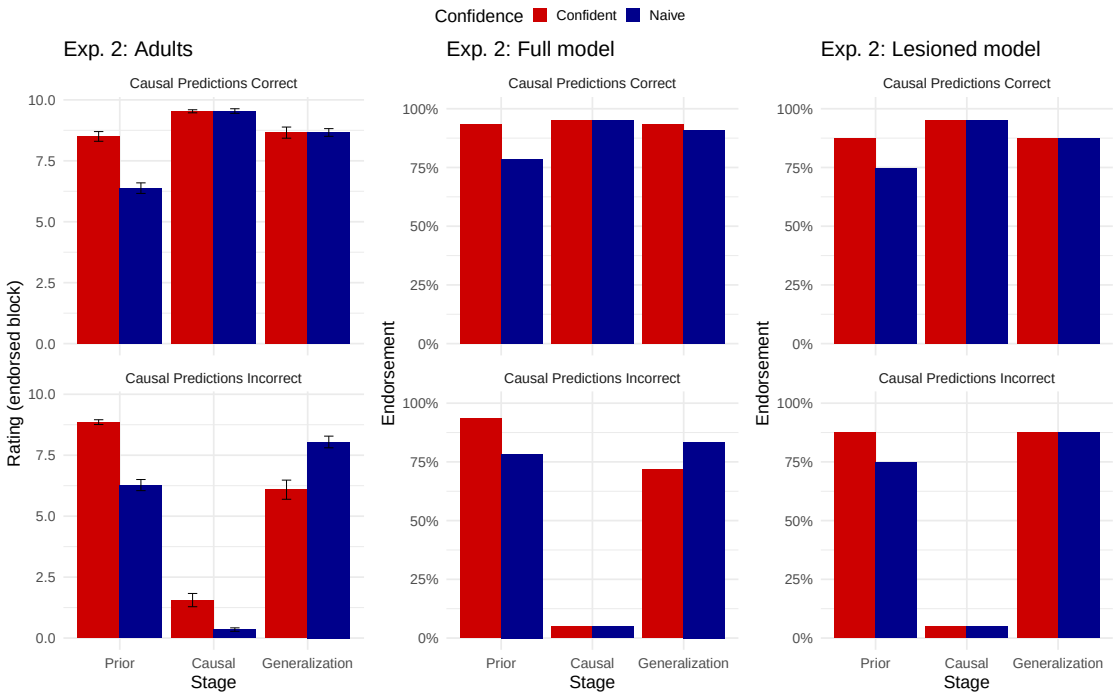


Figure 3

Empirical human data (left) and model predictions for the full model (middle) and lesioned model without global knowledge or self-knowledge (right) for Experiment 2. Adults showed the same pattern in their ratings in the generalization phase as in Experiment 1 and the full model predictions, dismissing the testimony from the confident informant who made an incorrect prediction (bottom). When the informant's prediction was correct (top), adults, as with the other models, were willing to endorse the testimony from both the confident and the naive informant.

human data on the task than any of the lesioned models, outperforming the model lacking global knowledge (log $BF_{01} = 21.94$), the model lacking self-knowledge (log $BF_{01} = 23.42$), and the model lacking both global knowledge and self-knowledge (log $BF_{01} = 28.36$), providing decisive evidence in favour of the full model.

Experiment 3

In Experiments 1 and 2, we showed that adults' choices on a reasoning task were captured well by a model of testimonial causal learning that assumes that learners track

both the global knowledge environment as well as the calibration of individual informants when deciding how much credence to lend to the endorsements of an informant. The results in Experiment 2 also provided further empirical validation of the model predictions, as the parameters fit to Experiment 1 predicted adults' performance in Experiment 2 better than any of the lesioned models.

The findings from Bridgers et al. (2016) suggest that 4-year-olds have some emerging understanding that confident informants can nevertheless be wrong, but that under some circumstances—e.g., when the data is more ambiguous—it may be reasonable to continue to lend credence to a confident informant. On the other hand, some studies have suggested that 3-year-olds distinguish accurate from inaccurate informants (Birch et al., 2008; Hermansen et al., 2021, but see Clément et al., 2004; Koenig and Harris, 2005), but there is limited evidence that 3-year-old children distinguish between informants whose accuracy is more uncertain or probabilistic in nature (e.g. Pasquini et al., 2007), despite the fact that 3-year-old children make sophisticated probabilistic inferences in other domains (e.g. Denison et al., 2006; Gopnik et al., 2004).

In this study, we thus adapt a similar experimental paradigm to Experiments 1 and 2 and to Bridgers et al. (2016) to 3-year-olds to investigate how 3-year-olds reason about conflicting evidence of varying levels of strength. Due to the potential demands on young children's attention and memory, we divide the within-subjects task completed by adults into a set of baseline conditions, representing the prior choice made by adults, and the conflict conditions, corresponding to the causal and generalization choices made by adults.

Methods

Participants and Design

118 U.S. 3-year-olds (M_{age} = 43.4 months, SD = 3.3, 57 girls, 61 boys) were recruited from local museums and preschools. Children were randomly assigned to one of six between-subjects conditions: the confident baseline condition (N = 20), the naive baseline condition (N = 20), the confident deterministic conflict condition (N = 19), the

naive deterministic conflict condition ($N = 19$), the confident probabilistic conflict condition ($N = 20$), or the naive probabilistic conflict condition ($N = 20$). This yielded a 2 x 3 factorial design with the informant's reported knowledge (confident or naive) and the evidence that children encountered (no data/baseline, conflicting deterministic data, and conflicting probabilistic data) as factors.

Materials and Procedure

Stimuli. Children were introduced to a machine consisting of a wooden box with a Lucite top that could light up and played music. The machine's activation was controlled by a hidden switch visible to the experimenter; the machine could be activated when children placed certain objects on top of the machine. The objects were four wooden blocks that differed in their shape and colour (a green circle, yellow square, pink triangle, and blue arch).

Study Procedure. Children sat at a table with two experimenters, the *informant* and the *assistant*. The informant introduced children to the machine and brought out two of the four wooden blocks. The informant explained that one block almost always activated the machine (the *endorsed* block) and one block almost never did (the *unendorsed* block). The informant's reported knowledge differed depending on the *certainty condition* (confident/naive). In the Confident condition, the informant expressed knowledge about the blocks, stating "I have played with these blocks a lot and so I really *know* which block is better at making the machine go". In the naive condition, the informant expressed ignorance about the blocks:

> I have never ever played with these blocks before, and I have no idea which block is better at making the machine go. Hmm... even though I dont know anything about them, Im just going to guess that the [endorsed block] almost always makes the machine go. And Im going to guess that the [unendorsed block] almost never makes it go. So, Im guessing that the [endorsed block] is better at making the machine go.

In addition to the differing statements made by the informant in the naive and confident conditions, the informant's nonverbal cues (e.g., furrowed brow, hesitation; see also Jaswal & Malone, 2007) also differed between the two conditions.

After providing the initial testimony, the informant left the room, leaving only the assistant, who up until this point had been uninvolved in the experiment. The assistant behaved differently depending on the *evidence condition* (baseline/deterministic conflict/probabilistic conflict).

Baseline condition. In the baseline condition, the assistant asked the child to choose just one of the blocks introduced to them by the experimenter to intervene on the machine to make the machine activate. After choosing a block, the child was also asked to recall which block the informant had endorsed as being the better block to activate the machine.

Deterministic conflict condition. In this condition, there were two phases, the *causal phase* and the *generalization phase*. In the causal phase, rather than immediately asking children to endorse a block, the assistant demonstrated each block on the machine. The endorsed block activated the machine 0/6 times, while the unendorsed block activated the machine 6/6 times. Thus, the physical evidence for the machine activation contradicted the informant's testimony about which was the better block.

After the assistant demonstrated the physical evidence, just as in the baseline condition, the assistant asked the child to intervene on the machine to make it activate, and subsequently to recall which block the informant had previously endorsed.

The causal phase was followed by the generalization phase. In this phase, the informant returned with two new blocks (the two blocks not initially presented to the child) and endorsed one block while unendorsing the other block. Regardless of whether the informant initially demonstrated knowledge or uncertainty about the first set of blocks, the informant claimed to *know* which was the better block in the generalization trial.

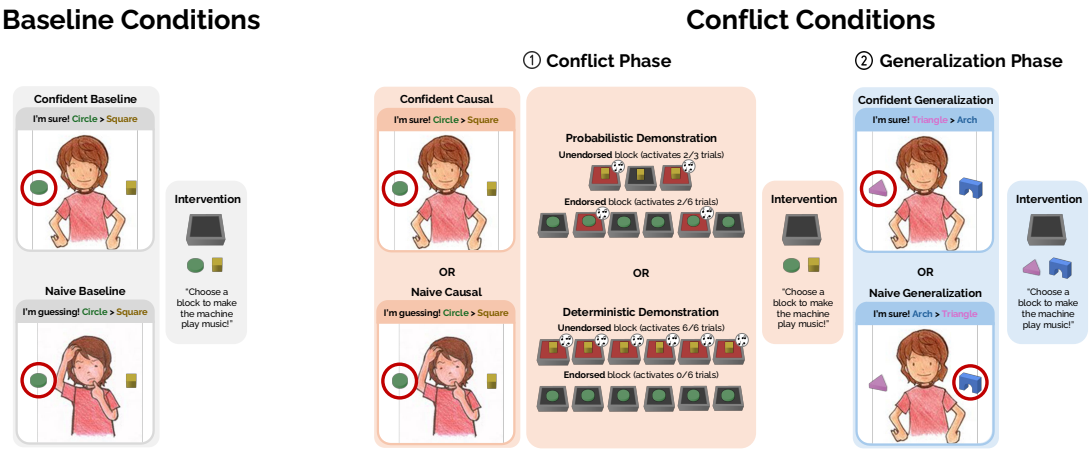


Figure 4

Description of testimony and interventions presented to children in the baseline conditions (left) and the conflict conditions (right). Children encountered either a confident (top) or naive (bottom) informant provide testimony about two blocks. In the baseline conditions, they chose a block to intervene on without encountering evidence. In the conflict conditions, children first saw probabilistic or deterministic evidence that conflicted with the informant's claim. Then, after choosing the block from the first two they believed would activate the machine, children saw the same informant provide testimony about two new blocks. Both the initially confident and initially naive informant were confident in the generalization phase, but endorsed different blocks.

Probabilistic conflict condition. This condition was identical to the deterministic conflict condition, except that the initial demonstration of the physical evidence presented by the assistant did not deterministically contradict the endorsement made by the informant, but instead did so probabilistically. The endorsed block activated the machine 2/6 times (on the second and fifth trials), while the unendorsed block activated the machine 2/3 times (on the first and third trials).

Results and Discussion

Results for Experiment 3 are included in Figure 5. Similarly to Experiment 1, we conducted a Bayesian generalized logistic regression, evaluating the effects of reported informant knowledge (confident or naive), conflicting evidence (probabilistic or deterministic), and choice type (baseline, causal, or generalization) as predictors of children's endorsement of the informant. Because children in Experiment 4 participated only in the baseline condition (corresponding to the prior choice for adults in Experiment 1) or in the conflict conditions (corresponding to the causal and generalization choices for adults in Experiment 1), we cannot directly test the strength of the changes between the baseline ratings and the causal and generalization ratings within subjects; instead, we compare the effects of reported informant knowledge across the five response types (baseline, probabilistic causal, probabilistic generalization, deterministic causal, deterministic generalization).

Overall, 3-year-old children were sensitive to the conflicting evidence they encountered, choosing the block endorsed by the informant significantly more often in the Baseline condition (85.0%) relative to the causal choice in either the Deterministic (15.4%), $OR = 61.5$, 95% CI = [12.5, 354], 0.00% in ROPE, or Probabilistic evidence conditions (22.5%), $OR = 29.7$, 95% CI = [8.21, 110], 0.00% in ROPE, irrespective of the informant's initially reported certainty. Similarly, they chose the endorsed block more often in the generalization trials (79.7%) than in the causal trials (19.0%), 0.00% in ROPE for all contrasts. There were no conclusive differences between the Baseline condition and the generalization trials, suggesting that 3-year-old children reduced their confidence in an informant's testimony when physical evidence contradicted the claims of the informants, but were generally willing to extend a similar degree of credence to the informant's testimony when the informant made claims about a previously unobserved block.

Interestingly, however, they did not conclusively show an effect of informant certainty across any condition, although there was anecdotal evidence that children were

more likely to choose the endorsed block in the Probabilistic causal rating when the informant was confident (35.0%) than when the informant was naive (10.0%), $OR = 5.70$, 95% CI = [1.06, 41.8], 0.95% in ROPE, a pattern that was similar to the choices of 4-year-olds in Bridgers et al. (2016). This may provide some limited evidence that 3-year-old children are able to differentially track how evidence strength influences willingness to accept testimony depending on the confidence of the informant, although possibly to a lesser degree than older children.

Model Comparison

As with adults in Experiment 1, we computed the likelihood of children's data under the four potential models (full, lacking global knowledge, lacking self-knowledge, lacking both) and fit the parameters to children's choices.

3-year-old children were best fit by the model lacking self-knowledge (BIC = 210.10), although there was only incidental evidence supporting it over the model lacking both self-knowledge and global knowledge (log $BF_{01} = 0.31$). There was positive evidence supporting the model lacking self-knowledge over the full model (log $BF_{01} = 3.26$) and the model lacking global knowledge but not self-knowledge (log $BF_{01} = 2.95$). Across the best-fitting parameters of all models, the δ parameter (reflecting the frequency of misstating one's calibration when providing testimony) was high (≥ 0.24), providing further evidence that 3-year-olds treat confidence signals that conflict with the outcome of the physical evidence as mostly uninformative.

Additionally, to examine how children's inferences change across development, we also compared the model's predictions to the data from Bridgers et al. (2016), where 4-year-old children completed a similar task. The best-fitting model for 4-year-olds was once again the model lacking self-knowledge (BIC = 188.12), but unlike 3-year-olds, there was positive evidence supporting this model over the model lacking both self-knowledge and global knowledge (log $BF_{01} = 2.42$), the model lacking global knowledge only (log $BF_{01} = 1.82$) and the full model (log $BF_{01} = 1.18$).

Together, these findings suggest that children between 3 and 4 continue to develop substantially in their ability to track informants' accuracy, with 4-year-olds and perhaps some 3-year-olds reducing their endorsement of informants' future testimony after encountering conflicting evidence. However, neither 3-year-olds nor 4-year-olds readily use information about the calibration of informants' statements to reason about the likely reliability of their future testimony.

Experiment 4

Comparing the findings for 3-year-olds (in Experiment 3) and 4-year-olds (from Bridgers et al., 2016) with the predictions of our model, we show that children in this age range respond in a manner consistent with tracking global knowledge, but not self-knowledge. When confronted with an informant who makes an inaccurate claim, they become less likely to endorse that informant's claims in the future (and at least in the case of 4-year-olds, adjust their willingness to continue to endorse the informant depending on the strength of the countervailing evidence). However, they do not treat the mistakes of a confidently wrong informant as uniquely undermining the reliability of that informant's confidence (i.e., relative to a circumstance where someone makes a mistake about a domain they are aware they are ignorant of, but are confident about their knowledge of a different domain).

One explanation for this pattern may be that children's understanding of what confidence in one's own knowledge represents is still developing, and 4-year-old children do not yet represent a the difference between an inappropriately confident informant who makes an incorrect claim, from an informant who is incorrect but is appropriately uncertain about guesses, as some findings suggest (Fobert et al., 2024; Kominsky et al., 2016; Tenney et al., 2011, but see Kushnir and Koenig, 2017).

Another possibility is that 4-year-olds do represent this, but may struggle to exhibit this due to task demands. For example, if children are explicitly presented with two informants whose calibration differs (as other studies have used to assess e.g. 3- and

4-year-olds children's perceptions of accurate vs. inaccurate informants; Birch et al., 2008; Koenig & Harris, 2005; Koenig et al., 2004; Pasquini et al., 2007), this may make the difference between a calibrated informant and a miscalibrated informant clearer, leading children to behave in a manner better captured by the full model rather than one of the lesioned models.

To provide a better comparison with children's choices, in Experiment 4 we first run a version of a task with two informants with adults, allowing us to evaluate developmental differences on our task between adults and 4-year-olds.

Methods

Participants and Design

101 U.S. participants were recruited from Amazon's Mechanical Turk service and were compensated $0.50 for completing the task. Participants were required to have 50 successful HITs with an approval rating of over 95% in order to be eligible for the task. Participants were assigned to one of two between-subjects conditions: initially confident condition ($N = 53$) and initially naive condition ($N = 48$).

Materials and Procedure

Participants completed an online web task similar to Experiments 1 and 2. They were introduced by an assistant character to a machine that that plays music when certain blocks are placed on it. The assistant further introduced participants to two informants, who provided testimony about what would happen when different blocks were placed on the machine.

The testimony from the informants varied based on the two conditions. In the *initially confident* condition, *both* informants indicated that they knew which block would activate the machine, and endorsed the same block (e.g., blue cylinder over orange square). In the *initially naive* condition, one informant indicated that she knew which block would activate the machine, while the other informant stated that she did not know and would guess. Both informants also endorsed the same block. Participants were then asked to rate

the likelihood that each block would activate the machine on a scale from 0 to 10 (the prior rating).

The remainder of the experiment was the same in both conditions. After the informants provided their testimony, the assistant returned and showed a pattern of deterministic evidence that conflicted with the testimony provided by both informants (e.g., the orange square activated the machine 6/6 times, while the blue cylinder activated it 0/6 times). Then, as before, the assistant asked the participant to rate the likelihood that each block would activate the machine on a scale from 0 to 10 (the causal rating).

Finally, the informants returned and provided testimony about two unobserved blocks (green rectangle and pink disk; identity counterbalanced with the blocks used in the first two phases). This time, both informants stated that they knew which block would activate the machine (irrespective of the initial confidence of the second informant), but endorsed opposite blocks. After this testimony, the assistant asked the participant to rate the likelihood that each block would activate the machine on a scale from 0 to 10 (the generalization rating).

After providing the generalization ratings, participants also answered a forced-choice question where they were asked to select the block from the generalization trial that they thought was more likely to activate the machine.

Results and Discussion

To assess the effects of the reported knowledge state of the informants and whether the evidence conflicted with the testimony, we used a Bayesian mixed-effects linear regression with the second informant's reported knowledge (confident or naive) and rating type (prior, causal, generalization) as predictors of participants' block ratings (measured as the difference in ratings between the informant's endorsed block and the unendorsed block).

In the prior rating, participants rated the endorsed block more highly than the unendorsed block ($\beta = 7.12$, 95% CI = [6.41, 7.72], 0.00% in ROPE), but there was no conclusive evidence that participants were more likely to rate the endorsed block more

highly when the second informant was confident compared to when the second informant was naive (β = -0.508, 95% CI = [-1.80, 0.77], 59.14% in ROPE). Conversely, in the causal rating, participants were more likely to rate the unendorsed block more highly (β = -7.13, 95% CI = [-7.79, -6.51], 0.00% in ROPE), which also did not conclusively show evidence of differing based on the second informant's initially reported knowledge level (β = 0.29, 95% CI = [-1.07, 1.51], 67.90% in ROPE).

However, when the second informant was initially naive, participants gave higher ratings to the block she endorsed than the block endorsed by the first informant (β = -1.97, 95% CI = [-2.88, -1.11], 0.00% in ROPE), while the ratings when the second informant also professed confidence did not conclusively differ by block (β = -0.14, 95% CI = [-1.08, 0.75], 87.19% in ROPE).

In the final forced-choice question, participants were more likely to choose the block endorsed by the second informant when she was initially naive than the first informant (OR = 2.35, 95% CI = [1.35, 4.44], 0.00% in ROPE), while no such difference emerged when the second informant also professed confidence (OR = 0.92, 95% CI = [0.51, 1.57], 46.70% in ROPE).

Model Comparison

Similar to our analysis for Experiment 2, we first converted participants' ratings to choices by assuming that they would always choose the block they gave a higher rating, or choose randomly when the blocks had equal ratings, and used the fitted parameters from Experiment 1 to test the full model and the three lesioned models. The best fitting model was the full model (BIC = 228.52), which performed inconclusively better than the model lacking self-knowledge (log BF_{01} = 0.93), and better than the model lacking global knowledge (log BF_{01} = 6.49) or lacking both (log BF_{01} = 5.07).

When including the maximization question, there was strong evidence supporting it over the other models, including the model lacking self-knowledge (log BF_{01} = 3.46), the model lacking global knowledge (log BF_{01} = 10.01) and the model lacking both

self-knowledge and global knowledge (log $BF_{01} = 4.71$).

Experiment 5

In Experiment 4, we found that adults once more made choices consistent with the idea that they track the calibration of informants. Although the model comparisons were somewhat inconclusive, part of the reason for this may be that the choices in Experiment 4 made quite similar predictions across most of the conditions (Figure 6), such that the lesioned model could adequately capture much of the variability in people's responses. However, in the key condition where the models make distinct predictions, participants chose responses consistent with the full model, favouring the second informant more than the first informant when the second informant had been initially naive, and thus more likely to be calibrated regarding their own knowledge than the informant who had been confidently wrong.

In Experiment 5, we adapt the task from Experiment 4 for 4-year-olds, similarly dividing the experiment into baseline conditions and a contrast condition to reduce the total number of judgments required to be elicited from children. If children are better able to track the difference between the calibration of two informants, we predict that children will respond similarly to adults in Experiment 4, endorsing an initially naive informant over an initially confident informant when both informants previously responded incorrectly but now indicate confidence about their endorsement of conflicting novel objects.

Methods

Participants and Design

180 Canadian 4-year-olds ($M_{age} = 53.9$ months, $SD = 3.5$, 65 girls, 80 boys) were recruited from local museums and preschools. To be eligible to participate, children had to be typically developing and understand sufficient English to understand the experimenter instructions. 35 participants were excluded due to experimenter error ($N = 12$), speaking insufficient English ($N = 9$), falling outside the age range ($N = 4$), participation in a previous version of the task ($N = 4$), child inattentiveness ($N = 3$), technical difficulties

with the experiment ($N = 2$), and not being typically developing ($N = 1$), leaving a final sample of 145 children included in the final sample for analysis.

170 parents completed an optional demographic questionnaire for their children. Participating children had the following demographic breakdown (percentages do not sum to 100% as parents could select multiple options): Caucasian/White (50.6%), East Asian (31.8%), South Asian (15.9%), Black (6.5%), Southeast Asian (5.9%), Latin American (5.3%), Middle Eastern (2.9%), Aboriginal (0.6%), and Other (4.7%).

Children were assigned to one of four between-subjects conditions. The first three were the *baseline* conditions ($N = 36$ each), while the fourth was the *contrast* condition ($N = 49$).

Materials and Procedure

Stimuli. Children were introduced to a machine consisting of a grey box with a green plastic top that could play music. The machine's activation was controlled by a hidden switch visible to the experimenter; the machine could be activated when children placed certain objects on top of the machine. The objects were four wooden blocks that differed in their shape and colour (a red circle, yellow square, blue triangle, purple star).

Informants' testimony about the machine was provided to children by means of a video displayed on a laptop, and children used headphones to listen to the statements made by the informants.

Study Procedure. In all conditions, children were first familiarized with the machine and its functions by an experimenter. They were told that two informants would provide them with information about the blocks before the child would be able to intervene on the machine. The information provided to the children and the opportunities for intervention varied depending on the conditions.

Baseline Conditions. Children first watched two informants provide information about the blocks. In the *Baseline Same* condition, both informants endorsed the same block (e.g., the red circle); one informant expressed confidence about her

knowledge and the other stated that she was guessing. In the *Baseline Confidence* condition, the informants endorsed opposing blocks, while one informant expressed confidence and the other stated that she was guessing. In the *Baseline Opposite* condition, both informants expressed confidence while endorsing opposing blocks, but unlike for the other conditions, the order that the informants provided testimony was not counterbalanced, to investigate for a potential primacy effect.

After watching the informants provide information about the blocks, children were invited to intervene on the machine, selecting one of the blocks to make the machine play music; if a child did not initially make a selection, they were reminded of the names of the objects and prompted once more to choose a block.

After the child made a selection, children were asked a memory check question about the block that each informant had endorsed at making the machine play music, and the informant's level of confidence in her endorsement (sure or just guessing).

Contrast Condition. There were two phases to this condition. In the *causal* phase, children were similarly introduced to two informants via video, who provided information about the blocks. The initial demonstration was identical to the Baseline Same condition: both informants endorsed the same block, with one informant indicating certainty and the other indicating that she was just guessing. Then, the experimenter demonstrated the physical evidence to the child. The block that had been endorsed by the informants was shown to activate the machine and play music 0/6 times, while the unendorsed block activated the machine and played music 6/6 times. The order of demonstrations, as well as the identity of the endorsed block, was counterbalanced between participants. After observing all of the physical evidence, children were asked by the experimenter which block they thought was better at making the machine play music.

The experimenter then took out photos of both informants, and asked the child a memory check question about which block each informant had endorsed at making the machine play music, and the informant's level of confidence in her observations.

Next, the experiment proceeded to the *generalization* phase. In this phase, the same two informants provided testimony about two new, unobserved blocks over video, with the order of presentation of the informants counterbalanced. This time, both the previously confident informant and the previously naive informant indicated that they were certain of the answer, but endorsed opposing blocks. After listening to the testimony from both informants, they were asked a memory check question by the experimenter about the informants' endorsements and level of confidence in their statements. If children answered incorrectly, the experimenter would place the correct block with the picture of the informant who endorsed the block, and state the informants' confidence in their claim (both confident).

Finally, children were asked by the experimenter to choose a block to place on the machine, and were asked to explain their choice.

Results and Discussion

To analyze children's choices across the conditions, we used a series of Bayesian generalized logistic regression models. In particular, we were interested in how children's choices changed in the contrast conditions relative to matched baseline conditions.

First, we compared the Baseline Same condition and the causal phase of the Contrast condition. In these two phases, informants made the same endorsements, only differing in whether children saw the conflicting causal evidence before making their testimony. Unsurprisingly, children were more likely to choose the block endorsed by both informants before observing any evidence ($\mu = 0.82$, 95% CI = [0.68, 0.94], 0.00% in ROPE), while no children chose the block endorsed by the informants after observing the causal evidence, where the endorsed block activated the machine 0/6 times and the unendorsed block activated it 6/6 times ($\mu = 0.001$, 95% CI = [0.00, 0.02], 0.00% in ROPE). These conditions also differed substantially from each other ($OR = 5288$, 95% CI = [46, $1.6 \cdot 10^9$], 0.00% in ROPE).

Next, we compared the Baseline Opposite and Baseline Confidence conditions to the

generalization phase of the Contrast condition. The Baseline Opposite condition corresponded to the generalization phase in that both informants in the generalization phase expressed confidence in the opposite blocks, while the Baseline Confidence condition corresponded in that one informant had previously expressed that she was just guessing, while the other informant had expressed certainty. In the Baseline Confidence condition, children were more likely to select the block endorsed by the confident rather than naive informant ($\mu = 0.72$, 95% CI = [0.57, 0.89], 0.00% in ROPE). In the Baseline Opposite condition, children were more likely to endorse the first informant who had provided testimony ($\mu = 0.79$, 95% CI = [0.64, 0.92], 0.00% in ROPE). Children's choices in the generalization phase of the Contrast condition did not differ meaningfully from chance ($\mu = 0.55$, 95% CI = 0.41, 0.68, 40.3% in ROPE), but also did not differ conclusively from the choice in either baseline condition (Baseline Confidence: 9.55% in ROPE; Baseline Opposite: 0.71% in ROPE).

Model Comparison

To examine the fit of the full model and the three lesioned models on Experiment 5, we used the parameter values derived from fitting the model to the 4-year-olds' data from Bridgers et al. (2016). Unexpectedly, we found that the best-fitting model was the model without global knowledge or self-knowledge (BIC = 204.51), outperforming the other models ($\log \text{BF}_{01} \geq 3.71$). Although this finding is somewhat unexpected given our findings that 4-year-old children and possibly 3-year-old children, this may be because all models made relatively similar predictions across trials, meaning that the full and partially lesioned models did not sufficiently improve fit to justify their complexity with respect to the fully lesioned model lacking both the self-knowledge and the global knowledge parameter. One reason for this may be that the models with 4-year-olds were best fit with a fairly high γ parameter (the prior over the frequency of causally efficacious objects), which pushes the model predictions towards 50% when it encounters conflicting testimony, or when the testimony is considered weak (i.e., testimony from a naive informant, or an

informant inferred to have poor general knowledge).

General Discussion

Across five experiments, we investigated how both preschool-aged children and adults reason about causal systems when integrating physical evidence and testimony from informants who vary in their accuracy, their confidence, and the calibration of their confidence (self-knowledge). Adults' choices were consistently best captured by a model that represents both informants' global knowledge and the extent to which their confidence is calibrated to their accuracy, treating confident but inaccurate informants as less reliable than informants who were also incorrect, but appropriately acknowledged when they were uncertain. In contrast, although 4-year-olds tracked informants' record of accuracy and adjusted their evaluations of informants' current and future claims according to the strength of the physical evidence they observed, their choices were better captured by a model that represented tracking of accuracy (general knowledgeability) but not informants' self-knowledge. 3-year-olds updated their beliefs about the efficacy of causal systems when informants were inaccurate, but the effects of informants' confidence on 3-year-olds' inferences were fairly weak, and they did not integrate the calibration of informants' expressed confidence into their evaluations.

Together, these findings provide evidence of a developmental shift in testimonial reasoning within early childhood as well as between childhood and adulthood. Three-year-olds on our tasks may reconcile conflicting testimony by relying on informant testimony when it is available (even if an informant has been inaccurate before), dismissing testimony when newer information conflicts with it, but not explicitly modelling the knowledge state of the informant. The 4-year-old children in our tasks reliably adjusted the degree of belief they were willing to extend a confident informant when the conflicting evidence was weaker (probabilistic) rather than deterministic, but did not necessarily consider the impact of an informant's inaccurate testimony on the reliability of their confidence.

These findings—and our computational model of testimonial reasoning—shed additional light on the somewhat mixed picture of testimonial reasoning in very young children. Although there is some evidence that children as young as 3 years old track the accuracy of informants (Birch et al., 2008; Hermansen et al., 2021), this effect appears weaker or more fragile than the effects observed with older children (Clément et al., 2004; Koenig & Harris, 2005; Pasquini et al., 2007), and may appear more robustly only when children are judging the accuracy of knowledge about already very familiar information, e.g. the labels of familiar objects. In our task, children were learning about the functions of a novel machine, so 3-year-olds may not have considered the informants' previous inaccuracy to imply a lack of general knowledge. Alternatively, children under 4 years old may treat testimony which is possible but probabilistically more unlikely as equivalent to erroneous testimony (e.g., Pasquini et al., 2007); however, this effect did not appear to influence children's subsequent testimony on the generalization trials, suggesting that this impression was not necessarily long-lived.

Interestingly, both 3- and 4-year-olds' willingness to grant informants who were confidently wrong the same benefit of the doubt as those who were aware that they did not know the answer is consistent with recent work arguing that young children perceive confidence primarily as a situational cue rather than a stable trait-like characteristic that has implications for an informant's general knowledge and metacognition (Juteau et al., 2024, 2025)—even if, at least in the case of 4-year-olds, children were less likely to endorse new testimony overall after encountering conflicting evidence, suggesting that they were tracking the ongoing accuracy of informants. If children consider confidence as a characteristic that reflects the situation one is in rather than an awareness of one's own knowledge state, this would explain the children on our tasks do not perceive the testimony of an informant who has previously been wrong despite being confident to be undermined by poor calibration. Given that the tendency to make situational compared to person-specific attributions has been found to vary across cultures (Cao et al., 2024; Choi

et al., 1999), this may also influence both how young children and adults perceive testimony from confident informants.

Another similar perspective has drawn from the findings that young children often exhibit robust trust in testimony, even after informants' claims have been contradicted by evidence (Heyman et al., 2013; Jaswal et al., 2010, 2014). Although we found that children quickly updated their beliefs to align with the physical evidence of causal efficacy that they observed, and, at least in the case of 4-year-olds, became less willing to endorse testimony from previously inaccurate informants irrespective of their expressed confidence, children's inferences about the confident and naive informant may not have been the same as adults'. For example, even though young children correctly reject the testimony of informants who consistently provide inaccurate responses on a sticker-finding task, they often consider this behaviour to be due to the fact that the informant is less smart, while adults perceive it as a strategy of deliberate deception (Ronfard & Lane, 2019).

In the same way that children's and adults' inferences when faced with an informant giving advice that is "suspiciously bad" changes from a belief that the informant is incompetent to a belief that the informant is competent at deliberately deceiving the listener, tracking informants' metacognitive knowledge states likely involves some degree of mentalizing about the informant. Although children typically succeed on classic false-belief tasks of theory of mind by the age of 4 years old (Wellman et al., 2001), understanding the metacognitive elements of others' self-knowledge may require more complex representations that are still in the process of developing. Just as inhibitory control and second-order theory of mind abilities have been found to correlate with understanding how listeners' epistemic states may differ from reality when they are being deceived (Tay et al., 2024), metacognitive representation may be necessary to represent an informant who *thinks that they know* but that the child knows often has undue confidence. Recent work has found that theory of mind performance is positively associated with children's use of accuracy as a cue to selective social learning (Dutemple et al., 2023; Resendes et al., 2021); if reasoning

about calibration involves similar cognitive processes, this would predict that children with higher scores on theory of mind tasks (e.g., Korkman et al., 2007; Wellman & Liu, 2004) would be more likely to make choices consistent with the full model on our tasks.

Three-year-olds in particular may also find it difficult to inhibit the tendency to endorse testimony provided deliberately and intentionally to them by an informant, possibly because they perceive the testimony to hold not only epistemic value, but *interpersonal* value in participating in a shared obligation between the speaker and listener (Koenig et al., 2022) and establishing good relations with social partners (Jaswal & Kondrad, 2016). Particularly when confidence can be associated with positive personality traits such as intelligence by young children (Birch et al., 2020), children might treat some testimonial reasoning tasks as partially affiliative in nature, particularly when interacting with experimenters who are visible and physically present (as in Experiment 3; see also Jaswal et al., 2010).

In addition to their implications on our understanding of the cognitive processes underlying how children evaluate others' confidence in early development, these findings have broader implications for our understanding of testimony in an epistemic environment in which children and adults increasingly encounter information from non-human sources. An emerging line of work has tested trust in emerging technologies such as voice assistants (Girouard-Hallam & Danovitch, 2022) and generative AI (Ding et al., 2025), which may provide confident answers that are nevertheless incorrect. Indeed, recent work on reliance in AI systems has found that while adults are sensitive to some elements of the accuracy and calibration of tools such as large language models, certain characteristics of responses such as providing explanations can increase reliance irrespective of accuracy (Kim et al., 2025). Broader insight into how reasoning about the accuracy and calibration changes across childhood and into adulthood will be critical in order to understand how people's willingness to extend selective trust not just to other people but also to these increasingly prevalent technologies.

Together, these findings contribute to a growing literature on the developmental changes in testimonial reasoning about confidence in early childhood (Brosseau-Liard et al., 2014; Fobert et al., 2024; Juteau et al., 2024, 2025), highlighting the progression from first tracking informants' record of accuracy and subsequently tracking the calibration of informants' testimony, and establishing a computational model that captures this trajectory between early childhood and adulthood. Our work also informs existing research into the extent to which evidential reasoning influences children's and adults' belief changes; computational and empirical work has established that, particularly when evidence is ambiguous, individuals can diverge in their beliefs after encountering the same data (Fryer et al., 2019; Gelpí, León-Villagrá, et al., 2025; Jern et al., 2014). A similar phenomenon may drive an "illusion of consensus" when the source of an individual's knowledgeor the dependency between multiple informantsis not clear (Aboody et al., 2022; Alister et al., 2022; Desai et al., 2022; Gelpí, Whalen, et al., 2025), which may be more difficult for children to identify.

Acknowledgements

This work was supported by funding from the Social Sciences and Humanities Research Council of Canada and the Canada Foundation for Innovation to Daphna Buchsbaum, and a SURF/Rose Hills Fellowship to Amy Whalen. We are grateful for all the families for their participation, and to the Computational Cognitive Development lab for their help with data collection. We would especially like to thank Nafisa Bhuiyan and Jing Yi Wang for their help with data collection and analysis. Thank you also to Alison Gopnik and Tom Griffiths for their assistance with conceptualization of earlier versions of the task, and for help facilitating data collection.

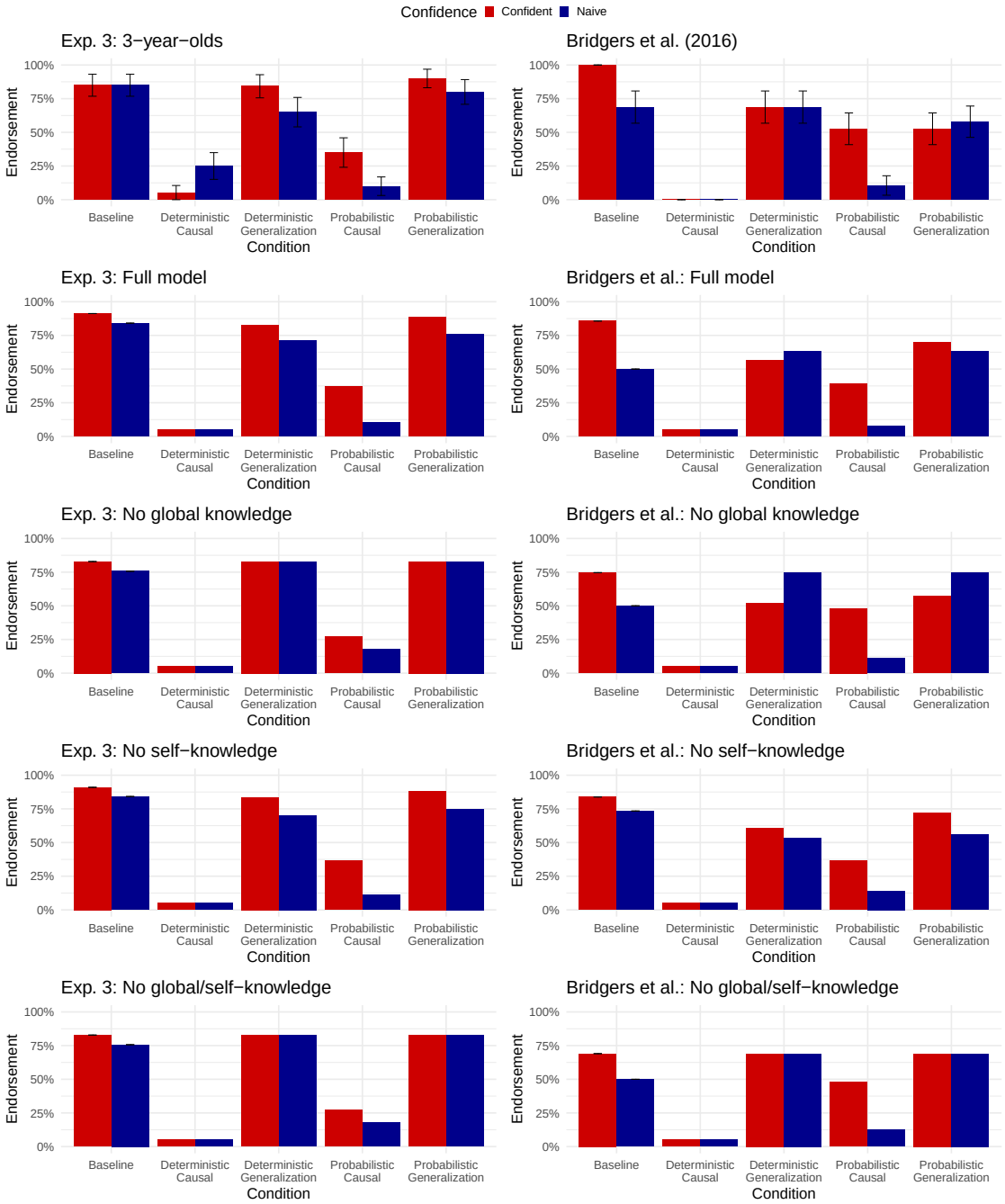


Figure 5

Results from Experiment 3 (left) and from Bridgers et al. (2016) (right), showing children's choices across all five conditions (top row), the full model predictions (row 2), model with no global knowledge (row 3), model with no self-knowledge (row 4), and fully lesioned model (row 5). Both 3-year-olds and 4-year-olds were best fit by the model with no self-knowledge (row 4), although 3-year-olds were only marginally better fit by the model with no self-knowledge than the full model.

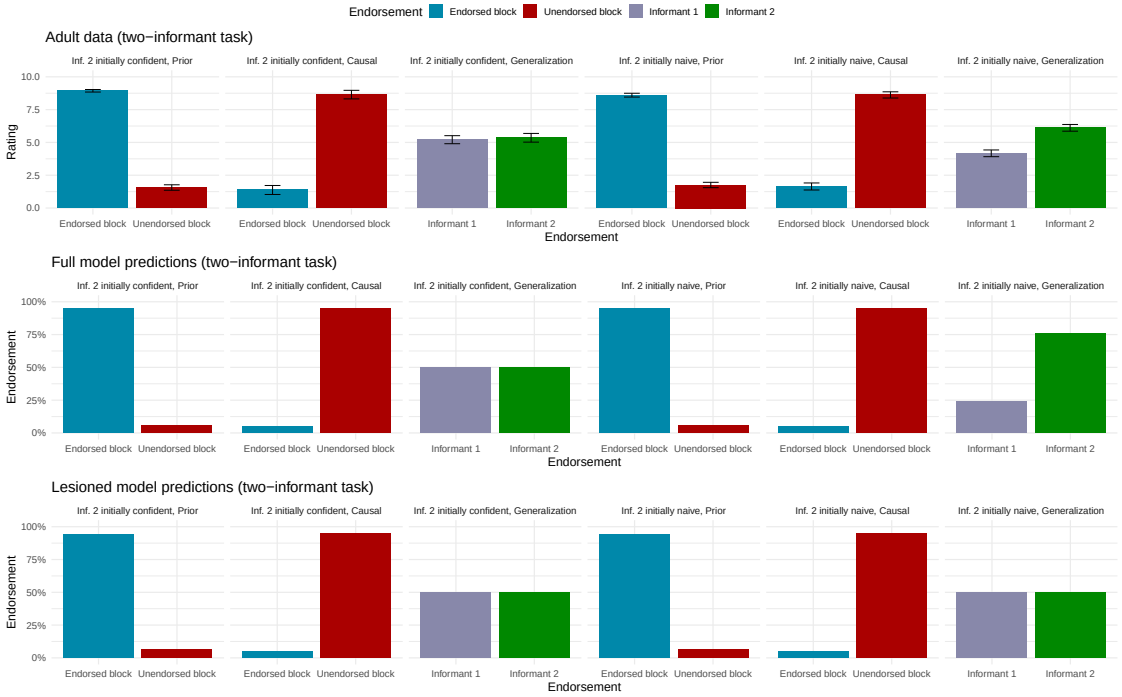


Figure 6

Human data (top) and model predictions for full (middle) and fully lesioned (bottom) model for Experiment 4. Like the full model, participants were more likely to endorse the second informant when the informant had initially been naive.

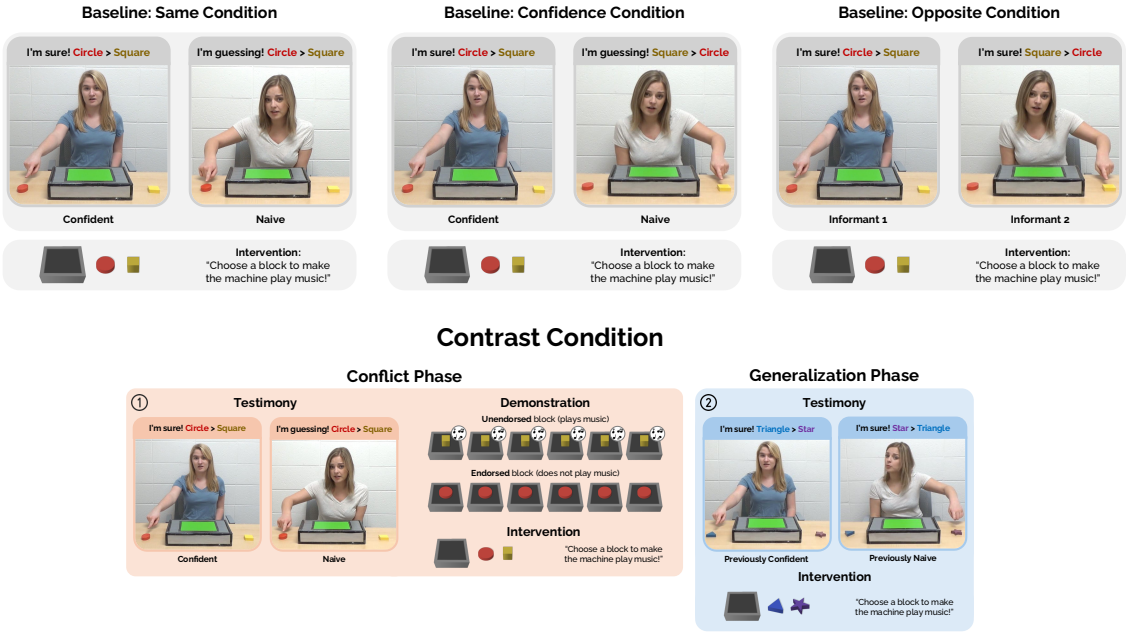


Figure 7

Description of testimony and interventions presented to children in the baseline conditions (top) and contrast condition (bottom). In all conditions, children saw two informants provide initial testimony, and then subsequently choose an object to intervene on the machine. In the contrast condition, children additionally saw a demonstration of physical evidence before making the first intervention, and then saw the same informants endorse two new objects.

References

Aboody, R., Yousif, S. R., Sheskin, M., & Keil, F. (2022). Says who? Children consider informants sources when deciding whom to believe. *Journal of Experimental Psychology: General*. https://doi.org/10.31234/osf.io/z6sbr

Alister, M., Perfors, A., & Ransom, K. (2022). *Source independence affects argument persuasiveness when the relevance is clear* (tech. rep.). PsyArXiv. https://doi.org/10.31234/osf.io/5yx36

Baer, C., & Kidd, C. (2022). Learning with certainty in childhood. *Trends in Cognitive Sciences*, *26*(10), 887–896. https://doi.org/10.1016/j.tics.2022.07.010

Baer, C., Malik, P., & Odic, D. (2021). Are childrens judgments of anothers accuracy linked to their metacognitive confidence judgments? *Metacognition and Learning*, *16*(2), 485–516. https://doi.org/10.1007/s11409-021-09263-x

Bernard, S., Castelain, T., Mercier, H., Kaufmann, L., Van der Henst, J.-B., & Clément, F. (2016). The boss is always right: Preschoolers endorse the testimony of a dominant over that of a subordinate. *Journal of Experimental Child Psychology*, *152*, 307–317. https://doi.org/10.1016/j.jecp.2016.08.007

Birch, S. A. J., Akmal, N., & Frampton, K. L. (2010). Twoyearolds are vigilant of others nonverbal cues to credibility. *Developmental Science*, *13*(2), 363–369. https://doi.org/10.1111/j.1467-7687.2009.00906.x

Birch, S. A. J., Severson, R. L., & Baimel, A. (2020). Children's understanding of when a person's confidence and hesitancy is a cue to their credibility (V. Capraro, Ed.). *PLOS ONE*, *15*(1), e0227026. https://doi.org/10.1371/journal.pone.0227026

Birch, S. A. J., Vauthier, S., & Bloom, P. (2008). Three- and four-year-olds spontaneously use others past performance to guide their learning. *Cognition*, *107*, 1018–1034. https://doi.org/10.1016/j.cognition.2007.12.008

Bowes, S. M., Novick, K., Lourenco, S. F., & Tasimi, A. (2025). Do children value intellectual humility over intellectual arrogance? *Developmental Psychology*. https://doi.org/10.1037/dev0001991

Bridgers, S., Buchsbaum, D., Seiver, E., Griffiths, T. L., & Gopnik, A. (2016). Childrens causal inferences from conflicting testimony and observations. *Developmental Psychology*, *52*(1), 9–18. https://doi.org/10.1037/a0039830

Brosseau-Liard, P. E. (2014). Selective, but Only if It Is Free: Children Trust Inaccurate Individuals More when Alternative Sources Are Costly: Cost and Selectivity. *Infant and Child Development*, *23*(2), 194–209. https://doi.org/10.1002/icd.1828

Brosseau-Liard, P. E., Cassels, T., & Birch, S. A. J. (2014). You Seem Certain but You Were Wrong Before: Developmental Change in Preschoolers Relative Trust in Accurate versus Confident Speakers (A. Senju, Ed.). *PLoS ONE*, *9*(9), e108308. https://doi.org/10.1371/journal.pone.0108308

BrosseauLiard, P. E., & PoulinDubois, D. (2014). Sensitivity to Confidence Cues Increases during the Second Year of Life. *Infancy*, *19*(5), 461–475. https://doi.org/10.1111/infa.12056

Buchsbaum, D., Bridgers, S., Whalen, A., Seiver, E., Grifths, T. L., & Gopnik, A. (2012). Do I know that you know what you know? Modeling testimony in causal inference. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*, 156–151.

Buchsbaum, D., Seiver, E., Bridgers, S., & Gopnik, A. (2012). Learning about Causes from People and about People as Causes. In *Advances in Child Development and Behavior* (pp. 125–160, Vol. 43). Elsevier.

Butler, L. P., Schmidt, M. F. H., Tavassolie, N. S., & Gibbs, H. M. (2018). Childrens evaluation of verified and unverified claims. *Journal of Experimental Child Psychology*, *176*, 73–83. https://doi.org/10.1016/j.jecp.2018.07.007

Cao, A., Carstensen, A., Gao, S., & Frank, M. C. (2024). United StatesChina differences in cognition and perception across 12 tasks: Replicability, robustness, and

within-culture variation. *Journal of Experimental Psychology: General*, *153*(11), 2657–2685. https://doi.org/10.1037/xge0001559

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin*, *125*(1), 47–63. https://doi.org/10.1037/0033-2909.125.1.47

Clément, F., Koenig, M., & Harris, P. (2004). The Ontogenesis of Trust. *Mind & Language*, *19*(4), 360–379. https://doi.org/10.1111/j.0268-1064.2004.00263.x

Corriveau, K. H., Fusaro, M., & Harris, P. L. (2009). Going With the Flow: Preschoolers Prefer Nondissenters as Informants. *Psychological Science*, *20*(3), 372–377. https://doi.org/10.1111/j.1467-9280.2009.02291.x

Corriveau, K. H., Harris, P. L., Meins, E., Fernyhough, C., Arnott, B., Elliott, L., Liddle, B., Hearn, A., Vittorini, L., & de Rosnay, M. (2009). Young Childrens Trust in Their Mothers Claims: Longitudinal Links With Attachment Security in Infancy. *Child Development*, *80*(3), 750–761. https://doi.org/10.1111/j.1467-8624.2009.01295.x

Denison, S., Konopczynski, K., Garcia, V., & Xu, F. (2006). Probabilistic Reasoning in Preschoolers: Random Sampling and Base Rate. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *28*, 1216–1221.

Desai, S. C., Xie, B., & Hayes, B. K. (2022). Getting to the source of the illusion of consensus. *Cognition*, *223*, 105023. https://doi.org/10.1016/j.cognition.2022.105023

Ding, Y., Facciani, M., Joyce, E., Poudel, A., Bhattacharya, S., Veeramani, B., Aguinaga, S., & Weninger, T. (2025). Citations and Trust in LLM Generated Responses. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(22), 23787–23795. https://doi.org/10.1609/aaai.v39i22.34550

Dutemple, E., Hakimi, H., & Poulin-Dubois, D. (2023). Do I know what they know? Linking metacognition, theory of mind, and selective social learning. *Journal of*

Experimental Child Psychology, *227*, 105572. https://doi.org/10.1016/j.jecp.2022.105572

Enesco, I., Sebastián-Enesco, C., Guerrero, S., Quan, S., & Garijo, S. (2016). What Makes Children Defy Majorities? The Role of Dissenters in Chinese and Spanish Preschoolers Social Judgments. *Frontiers in Psychology*, *7*.

Fobert, S., Varin, R., Cossette, I., Fournier, K. R. C., & BrosseauLiard, P. E. (2024). Children presume confident informants will be accurate (until proven otherwise). *Infant and Child Development*, *33*(6), e2551. https://doi.org/10.1002/icd.2551

Fryer, R. G., Harms, P., & Jackson, M. O. (2019). Updating Beliefs when Evidence is Open to Interpretation: Implications for Bias and Polarization. *Journal of the European Economic Association*, *17*(5), 1470–1501. https://doi.org/10.1093/jeea/jvy025

Gelpí, R. A., & Buchsbaum, D. (2024). Children as Cultural Explorers: How Imitation, Pedagogy, and Selective Trust Prepare Children for Learning in the Cultural Niche. In J. J. Tehrani, J. Kendal, & R. L. Kendal (Eds.), *The Oxford Handbook of Cultural Evolution* (1st ed.). Oxford University Press.

Gelpí, R. A., León-Villagrá, P., Cunningham, W. A., Lucas, C. G., & Buchsbaum, D. (2025). Resource-rational belief revision can mitigate as well as amplify polarization. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Gelpí, R. A., Otsubo, K., Whalen, A., & Buchsbaum, D. (2025). Investigating Sensitivity to Shared Information and Personal Experience in Childrens Use of Majority Information. *Open Mind*, *9*, 240–265. https://doi.org/10.1162/opmi_a_00182

Gelpí, R. A., Whalen, A., Griffiths, T. L., Xu, F., & Buchsbaum, D. (2025). Can children and adults balance majority size with information quality in learning from preferences? *Journal of Experimental Psychology: General*, *154*(5), 1388–1406. https://doi.org/10.1037/xge0001724

Girouard-Hallam, L. N., & Danovitch, J. H. (2022). Childrens trust in and learning from voice assistants. *Developmental Psychology*, *58*(4), 646–661. https://doi.org/10.1037/dev0001318

Goddu, M. K., & Gopnik, A. (2024). The development of human causal learning and reasoning. *Nature Reviews Psychology*, *3*(5), 319–339. https://doi.org/10.1038/s44159-024-00300-5

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. *Psychological Review*, *111*(1), 3–32. https://doi.org/10.1037/0033-295X.111.1.3

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. https://doi.org/10.1111/j.1551-6709.2011.01203.x

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910. https://doi.org/10.1016/j.tics.2021.07.008

Hagá, S., & Olson, K. R. (2017). If I only had a little humility, I would be perfect: Childrens and adults perceptions of intellectually arrogant, humble, and diffident people. *The Journal of Positive Psychology*, *12*(1), 87–98. https://doi.org/10.1080/17439760.2016.1167943

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive Foundations of Learning from Testimony. *Annual Review of Psychology*, *69*(1), 251–273. https://doi.org/10.1146/annurev-psych-122216-011710

Haun, D. B. M., & Tomasello, M. (2011). Conformity to Peer Pressure in Preschool Children. *Child Development*, *82*(6), 1759–1767. https://doi.org/10.1111/j.1467-8624.2011.01666.x

Hermansen, T. K., Ronfard, S., Harris, P. L., Pons, F., & Zambrana, I. M. (2021). Young children update their trust in an informants claim when experience tells them otherwise. *Journal of Experimental Child Psychology*, *205*, 105063. https://doi.org/10.1016/j.jecp.2020.105063

Heyman, G. D., Sritanyaratana, L., & Vanderbilt, K. E. (2013). Young Children's Trust in Overtly Misleading Advice. *Cognitive Science*, *37*(4), 646–667. https://doi.org/10.1111/cogs.12020

Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young Children Have a Specific, Highly Robust Bias to Trust Testimony. *Psychological Science*, *21*(10), 1541–1547. https://doi.org/10.1177/0956797610383438

Jaswal, V. K., & Kondrad, R. L. (2016). Why Children Are Not Always Epistemically Vigilant: Cognitive Limits and Social Considerations. *Child Development Perspectives*, *10*(4), 240–244. https://doi.org/10.1111/cdep.12187

Jaswal, V. K., & Malone, L. S. (2007). Turning Believers into Skeptics: 3-Year-Olds' Sensitivity to Cues to Speaker Credibility. *Journal of Cognition and Development*, *8*(3), 263–283. https://doi.org/10.1080/15248370701446392

Jaswal, V. K., PérezEdgar, K., Kondrad, R. L., Palmquist, C. M., Cole, C. A., & Cole, C. E. (2014). Can't stop believing: Inhibitory control and resistance to misleading testimony. *Developmental Science*, *17*(6), 965–976. https://doi.org/10.1111/desc.12187

Jern, A., Chang, K.-m. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224. https://doi.org/10.1037/a0035941

Juteau, A.-L., Cossette, I., Millette, M.-P., & Brosseau-Liard, P. (2019). Individual Differences in Childrens Preference to Learn From a Confident Informant. *Frontiers in Psychology*, *10*, 2006. https://doi.org/10.3389/fpsyg.2019.02006

Juteau, A.-L., Holmes, C. J., & Brosseau-Liard, P. (2025). Confidence cues: Epistemic or social? *Cognitive Development*, *74*, 101574. https://doi.org/10.1016/j.cogdev.2025.101574

Juteau, A.-L., Ibrahim, Y. A., McIntee, S.-E., Varin, R., & Brosseau-Liard, P. E. (2024). Do children interpret informants confidence as person-specific or situational? (S. Triberti, Ed.). *PLOS ONE*, *19*(5), e0298183. https://doi.org/10.1371/journal.pone.0298183

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social Learning Strategies: Bridge-Building between Fields. *Trends in Cognitive Sciences*, *22*(7), 651–665. https://doi.org/10.1016/j.tics.2018.04.003

Kim, S. S. Y., Vaughan, J. W., Liao, Q. V., Lombrozo, T., & Russakovsky, O. (2025). Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3706598.3714020

Kinzler, K. D., Corriveau, K. H., & Harris, P. L. (2011). Childrens selective trust in native-accented speakers. *Developmental Science*, *14*(1), 106–111. https://doi.org/10.1111/j.1467-7687.2010.00965.x

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in Testimony: Children's Use of True and False Statements. *Psychological Science*, *15*(10), 694–698. https://doi.org/10.1111/j.0956-7976.2004.00742.x

Koenig, M. A., & Harris, P. L. (2005). Preschoolers Mistrust Ignorant and Inaccurate Speakers. *Child Development*, *76*(6), 1261–1277. https://doi.org/10.1111/j.1467-8624.2005.00849.x

Koenig, M. A., & Jaswal, V. K. (2011). Characterizing Childrens Expectations About Expertise and Incompetence: Halo or Pitchfork Effects? *Child Development*, *82*(5), 1634–1647. https://doi.org/10.1111/j.1467-8624.2011.01618.x

Koenig, M. A., Li, P. H., & McMyler, B. (2022). Interpersonal trust in children's testimonial learning. *Mind & Language*, *37*(5), 955–974. https://doi.org/10.1111/mila.12361

Kominsky, J. F., Langthorne, P., & Keil, F. C. (2016). The better part of not knowing: Virtuous ignorance. *Developmental Psychology*, *52*(1), 31–45. https://doi.org/10.1037/dev0000065

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY-II: Clinical and interpretive manual*. NCS Pearson.

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Kushnir, T., Vredenburgh, C., & Schneider, L. (2013). Who can help me fix this toy? The distinction between causal knowledge and word knowledge guides preschoolers selective requests for information. *Developmental Psychology*, *49*(3), 446–453. https://doi.org/10.1037/a0031649

Kushnir, T., & Gopnik, A. (2005). Young Children Infer Causal Strength From Probabilities and Interventions. *Psychological Science*, *16*(9), 678–683. https://doi.org/10.1111/j.1467-9280.2005.01595.x

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior

spatial assumptions. *Developmental Psychology*, *43*(1), 186–196. https://doi.org/10.1037/0012-1649.43.1.186

Kushnir, T., & Koenig, M. A. (2017). What I dont know wont hurt you: The relation between professed ignorance and later knowledge claims. *Developmental Psychology*, *53*(5), 826–835. https://doi.org/10.1037/dev0000294

Ma, L., & Ganea, P. A. (2010). Dealing with conflicting information: Young childrens reliance on what they see versus what they are told. *Developmental Science*, *13*(1), 151–160. https://doi.org/10.1111/j.1467-7687.2009.00878.x

McLoughlin, N., Finiasz, Z., Sobel, D. M., & Corriveau, K. H. (2021). Childrens developing capacity to calibrate the verbal testimony of others with observed evidence when inferring causal relations. *Journal of Experimental Child Psychology*, *210*, 105183. https://doi.org/10.1016/j.jecp.2021.105183

Orticio, E., Meyer, M., & Kidd, C. (2024). Exposure to detectable inaccuracies makes children more diligent fact-checkers of novel claims. *Nature Human Behaviour*, *8*(12), 2322–2329. https://doi.org/10.1038/s41562-024-01992-8

Palmquist, C. M., Floersheimer, A., Crum, K., & Ruggiero, J. (2022). Social cognition and trust: Exploring the role of theory of mind and hostile attribution bias in childrens skepticism of inaccurate informants. *Journal of Experimental Child Psychology*, *215*, 105341. https://doi.org/10.1016/j.jecp.2021.105341

Palmquist, C. M., & Kondrad, R. (2024). Knowledge and source type influence childrens skepticism of misinformation. *Journal of Cognition and Development*, *25*(3), 437–460. https://doi.org/10.1080/15248372.2024.2303774

Pasquini, E., Corriveau, K., Koenig, M., & Harris, P. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, *43*(5), 1216–1226. https://doi.org/10.1037/0012-1649.43.5.1216

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–321. https://doi.org/10.1016/j.cognition.2010.11.015

Pozzi, M., & Mazzarella, D. (2024). Speaker trustworthiness: Shall confidence match evidence? *Philosophical Psychology*, *37*(1), 102–125. https://doi.org/10.1080/09515089.2023.2193220

Resendes, T., Benchimol-Elkaim, B., Delisle, C., René, J.-L., & Poulin-Dubois, D. (2021). What I know and what you know: The role of metacognitive strategies in preschoolers selective social learning. *Cognitive Development*, *60*, 101117. https://doi.org/10.1016/j.cogdev.2021.101117

Ronfard, S., & Lane, J. D. (2019). Childrens and adults epistemic trust in and impressions of inaccurate informants. *Journal of Experimental Child Psychology*, *188*, 104662. https://doi.org/10.1016/j.jecp.2019.104662

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*(4), 1054–1070. https://doi.org/10.1111/1467-8624.00334

Schmid, B., Bleijlevens, N., Mani, N., & Behne, T. (2024). The cognitive underpinnings and early development of children's selective trust. *Child Development*, *95*(4), 1315–1332. https://doi.org/10.1111/cdev.14073

Sebastián-Enesco, C., Guerrero, S., & Enesco, I. (2020). What makes children defy their peers? Chinese and Spanish preschoolers' decisions to trust (or not) peer consensus. *Social Development*, *29*(2), 494–508. https://doi.org/10.1111/sode.12416

Seiver, E., Gopnik, A., & Goodman, N. D. (2013). Did She Jump Because She Was the Big Sister or Because the Trampoline Was Safe? Causal Inference and the Development of Social Attribution. *Child Development*, *84*(2), 443–454. https://doi.org/10.1111/j.1467-8624.2012.01865.x

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89. https://doi.org/10.1016/j.cogpsych.2013.12.004

Stanciu, O., & Fiser, J. (2022). Do humans recalibrate the confidence of advisers or take confidence at face value? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*, 3432–3438.

Tay, C., Ng, R., Ye, N. N., & Ding, X. P. (2024). Detecting lies through others eyes: Children use perceptual access cues to evaluate listeners beliefs about informants deception. *Journal of Experimental Child Psychology*, *241*, 105863. https://doi.org/10.1016/j.jecp.2024.105863

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*(4), 1065–1077. https://doi.org/10.1037/a0023273

Tenney, E. R., Spellman, B. A., & MacCoun, R. J. (2008). The benefits of knowing what you know (and what you dont): How calibration affects credibility. *Journal of Experimental Social Psychology*, *44*(5), 1368–1375. https://doi.org/10.1016/j.jesp.2008.04.006

Tong, Y., Wang, F., & Danovitch, J. (2020). The role of epistemic and social characteristics in childrens selective trust: Three meta-analyses. *Developmental Science*, *23*(2), e12895. https://doi.org/10.1111/desc.12895

Vanderbilt, K. E., Heyman, G. D., & Liu, D. (2014). In the absence of conflicting testimony young children trust inaccurate informants. *Developmental Science*, *17*(3), 443–451. https://doi.org/10.1111/desc.12134

Waismeyer, A., Meltzoff, A. N., & Gopnik, A. (2015). Causal learning from probabilistic events in 24-month-olds: An action measure. *Developmental Science*, *18*(1), 175–182. https://doi.org/10.1111/desc.12208

Walker, M. B., & Andrade, M. G. (1996). Conformity in the Asch Task as a Function of Age. *The Journal of Social Psychology*, *136*(3), 367–372. https://doi.org/10.1080/00224545.1996.9714014

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, *72*(3), 655–684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M., & Liu, D. (2004). Scaling of Theory-of-Mind Tasks. *Child Development*, *75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x