Parameterizing developmental changes in epistemic trust

Baxter S. Eaves Jr.
Rutgers University–Newark
Patrick Shafto
Rutgers University–Newark

Abstract

Children rely on others for much of what they learn, and therefore must track who to trust for information. Researchers have debated whether to interpret children's behavior as inferences about informants' knowledgeability only or as inferences about both knowledgeability and intent. We introduce a novel framework for integrating results across heterogeneous ages and methods. The framework allows application of a recent computational model to a set of results that span ages 8 months to adulthood and a variety of methods. The results show strong fits to specific findings in the literature trust, and correctly fails to fit one representative result from an adjacent literature. In the aggregate, the results show a clear development in children's reasoning about informants' intent and no appreciable changes in reasoning about informants' knowledgeability, confirming previous results. The results extend previous findings by modeling development over a much wider age range and identifying and explaining differences across methods.

Keywords: Bayesian models, social learning, selective trust, epistemic trust

Children face a difficult problem in learning about the world. There is much to learn and little time in which to learn it. In this context, the benefits of social learning are self-evident. Self-directed strategies are slow and cannot be used to acquire some knowledge (e.g. language). It is quicker to call upon on the knowledge of others. However, people do not always produce reliable data. A person may have inaccurate knowledge or may wish to deceive. Thus, it is necessary for people to trust informants and their information selectively (Koenig & Harris, 2005; Pasquini, Corriveau, Koenig, & Harris, 2007; Corriveau & Harris, 2009; Corriveau, Fusaro, & Harris, 2009; Chen, Corriveau, & Harris, 2012). This sort of selective trust in informants and their information is referred to as *epistemic trust*.

Research has identified informant and contextual features that cause children trust informants

This research was supported in part by NSF CAREER award DRL-1149116 and a grant from the DARPA XDATA program to P.S.

Corresponding Author

Baxter S. Eaves Jr.
Department of Mathematics & Computer Science
Smith Hall, Room 216
101 Warren Street, Newark, NJ 07102
e-mail: baxter.eaves@rutgers.edu
telephone: (502) 295-9619

and their information differently. Children trust more accurate informants (Koenig & Harris, 2005; Pasquini et al., 2007). Children are less likely to ask informants who mislabel common objects for future information than informants who label common objects correctly (Koenig & Harris, 2005) and children's preference for accurate over inaccurate informants increases with the relative accuracy between informants (Pasquini et al., 2007). Additionally, children have been shown to trust information from groups of informants over dissenters (Corriveau, Fusaro, & Harris, 2009; Chen et al., 2012) and to prefer familiar informants (Corriveau & Harris, 2009), informants with the same native accent (Kinzler, Corriveau, & Harris, 2011), informants of the same gender (Taylor, 2013) and more attractive informants (Bascandziev & Harris, 2014).

Research has also shown that children's epistemic trust develops. Older children seem to allocate their trust more flexibly than younger children (Koenig & Harris, 2005; Pasquini et al., 2007; Corriveau & Harris, 2009). The literature typically explains this development in terms of changes in the ability to monitor who is knowledgeable (Pasquini et al., 2007; Corriveau, Fusaro, & Harris, 2009; Corriveau & Harris, 2009).[1] Others have broadly argued that trust is rational (Sobel & Kushnir, 2013). An adjacent literature indicates changes in the ability to reason about deception (Couillard & Woodward, 1999; Mascaro & Sperber, 2009).

Shafto, Eaves, Navarro, and Perfors (2012) proposed a probabilistic model that formalizes epistemic trust as inferences about informants' knowledgeability (versus unknowledgeability) and helpfulness (versus deception) (see also Eaves & Shafto, 2012; Butterfield, Jenkins, Sobel, & Schwertfeger, 2008). The computational model was fit to three studies to investigate possible explanations for developmental changes in behavior. Contrary to the aforementioned qualitative accounts that attribute developmental changes to children's improving ability to monitor informants' knowledge, the results showed that the behavioral differences between three- and four-year-olds are primarily a result of a change in children's representation of informants' helpfulness. Three-year-olds' data was better explained by a model that only reasons about informants' knowledgeability, whereas four-year-olds' data was better explained by a model that reasons about both knowledge and helpfulness. Although provocative, these results are limited by the reliance on a small subset of the literature.

It would be desirable to use the computational model to generate a more integrative theoretical account of the literature on development of epistemic trust. Indeed, the model in principle should apply to findings across the literature. However, the research questions and methods used in epistemic trust research are heterogeneous. In addition to variations in age, researchers have investigated experimental features such as the modes through which informants communicate (e.g. verbal testimony, pointing, gaze), the experimental paradigm (e.g. forced-choice, looking time), and culture. Shafto, Eaves, et al. (2012) focused on a small subset of the overall literature to ensure homogeneity of tasks and ages that would allow all experiments to be explained with a simple, unified explanation, but this necessarily limits the explanatory power of the theory. Any integrative theory must deal with not only heterogeneity of tasks and ages, but *correlations* between task and age. Methods that work for very young children—such as looking time—do not work for older children, and vice versa. Indeed, the correlation between task and age, and the interpretation problems it poses, are a general problem for integrative theories of cognitive development.

In this paper we introduce a method for conducting integrative, model-driven analysis of heterogeneous experiments and apply it to the construction of an integrative account of the development of epistemic trust. The approach is based on two components: a domain-specific model of epistemic trust (Shafto, Eaves, et al., 2012) and a domain-general approach for integrative analysis (Mansinghka et al., Accepted pending revision; Shafto, Kemp, Mansinghka, & Tenenbaum, 2011). The model of epistemic trust is used to *parameterize* the conditions of heterogeneous experiments—to translate the experimental results into model parameters. Along with each parameterization, we document the methodological details of each condition—mean age, experimental paradigm, communication mode, etc. The collection of conditions, each translated into a set of model parameters

[1] For a review, see Mills, 2013.

and experimental features comprise the input into the integrative analysis. The integrative analysis infers a joint probability distribution over all relevant experimental features and model parameter values. The resulting joint distribution allows querying of conditional distributions over parameters and experimental features. From these conditional distributions we gain the ability to ask and answer fundamental questions about how features of conditions such as task and age are related to the variables in the model, e.g., how do children's beliefs about informants' helpfulness change from age 18 months, to 3 years, to 4.5 years or how are pointing versus verbal testimony reflected in children's beliefs about helpfulness.

We begin by discussing the heterogeneity in the epistemic trust literature. We then discuss the model of epistemic trust, followed by our approach to aggregating parameterized results. We then detail our methods and results, and conclude by discussing broader implications of this approach for epistemic trust and broader theories in cognitive development.

Heterogeneity in studies on the development of epistemic trust

The epistemic trust literature—as defined in terms of the scope of the computational model (see Eaves & Shafto, 2012)—is composed of many literatures each of which is interested in how learners trust informants differently in different contexts. The set of encompassed literature includes the selective trust, deception, informant expertise, and pedagogy literatures. Each of these literatures has its own conceptual, methodological, and age conventions. In this section, we briefly review each literature in turn and to offer a sense of the heterogeneity of the conceptual and methodological landscape.

The selective trust literature recounts people's different trust in informants driven by inferences about their epistemic states. As an example, (Koenig & Harris, 2005) proposed that children monitor the accuracy of informants and use prior accuracy information when choosing between and learning from informants. Preschool-aged children observed two informants label common objects (chair, ball, etc). One informant labeled all four objects correctly and the other labeled all four objects incorrectly. After three of these *accuracy trials*, unfamiliar objects were placed before the informants. The child was then asked which informant she would like to ask for the novel object's label (*ask* trial) or after having observed each informant provide a label, was asked to chose a label (*endorse* trial). Four-year-old children asked and endorsed the accurate informant most often. This result demonstrates that children's preferences for specific informants and their information is influenced by informants' accuracy. A number of other studies have reproduced this result and have shown that a single inaccuracy can shape children's informant preferences (Fitneva & Dunfield, 2010) and that children take into account not only whether an informant has been accurate or inaccurate but the relative accuracy between informants (Pasquini et al., 2007) and the magnitude of informants' errors (Einav & Robinson, 2010). Even infants appear to learn differently from reliable and unreliable informants (Tummeltshammer, Wu, Sobel, & Kirkham, 2014) and are surprised when informants mislabel common objects (Koenig & Echols, 2003). The selective trust literature also indicates that children prefer informants who are part of a consensus (Corriveau, Fusaro, & Harris, 2009; Chen et al., 2012), and who are more familiar (Corriveau & Harris, 2009) (e.g. their preschool teacher over a stranger). Another, closely related, line of research indicates that children may choose informants based on their superficial, non-epistemic, qualities such as their gender (Taylor, 2013), their attractiveness (Bascandziev & Harris, 2014), and accent (Kinzler et al., 2011). Research also suggests that selective trust is modulated by cultural factors. For example, children of different cultures are differently likely to accept seemingly unreliable information from a consensus (DiYanni, Corriveau, Nasrini, Kurkul, & Nini, 2015).

The deception literature recounts people's different trust in informants driven by inferences about knowledgeable informants' helpfulness. The deception literature is vast, addressing issues related to false belief, sarcasm and more. Here we consider only the simplest case, which is most closely related to tasks described above: informants who are knowledgeable but nonetheless provide inaccurate information. Research indicates that three-year-olds have difficulty handling deceptive

data compared with older children (Couillard & Woodward, 1999; Mascaro & Sperber, 2009). For example, three-year-olds, but not four-year-olds are repeatedly fooled by an informant who, for ten trials indicates, by way of pointing, the one of two cups under which no prize is hidden (Couillard & Woodward, 1999). In addition to age, reasoning about deception varies with communicative mode. The same study found that children's ability to choose the correct cup was improved if the informant indicated cups by placing markers on them rather than pointing at them.

The above studies focus on cases where the informant's testimony provides information about their trustworthiness. It is common to experience cases where an informant's trustworthiness is implied by social decree, as is the case with expertise. Research has investigated the development of trust in experts by pitting two informants labeled as experts in contrasting domains against each other. Children begin to correctly attribute domain knowledge fairly early, at about age four (Lutz & Keil, 2002; Aguiar, Stoess, & Taylor, 2012), and these abilities improve as children learn more about how knowledge domains are organized (Danovitch & Keil, 2004; Keil, Stein, Webb, Billings, & Rozenblit, 2008). Four-year-olds, but not three-year-olds, more often endorse novel object labels from informants who demonstrate accurate knowledge of those objects' functions and internal properties (Sobel & Corriveau, 2010). Additionally, preschoolers hold a domain-general view of ignorance and a domain-specific view of expertise (Koenig & Jaswal, 2011) and more often endorse information from nice non-experts than information from mean experts (Asheley R Landrum, Mills, & Johnston, 2013).

The pedagogy literature recounts people's different learning from informants when informants are assumed to be helpful and knowledgeable. For example, the *Natural Pedagogy* theory (Csibra & Gergely, 2009; Gergely, Egyed, & Király, 2007) asserts that children have a strong, in-born belief that all informants are helpful and knowledgeable and that relaxing this belief is a primary task in early development. Hence, the pedagogy literature looks at how children make different inferences about the world given data from teachers than they do given unintentional data (Bonawitz et al., 2011) or given data generated by self–directed strategies (Shafto & Goodman, 2008; Shafto, Goodman, & Griffiths, 2014). Recent research has demonstrated that children can identify when these assumptions do not apply and use self-directed means to fill in gaps left by poorly-performing pedagogs (Gweon, Shafto, & Schulz, 2014).

The different literatures employ different methods on different age groups. Trust-in-testimony research primarily focuses on two age groups: infants up to 18 months, and preschoolers from three to four years. Studies with preschoolers typically employ forced-choice paradigms, asking children which informants they prefer or what information they believe; and research on infants is carried out using looking-time and simple motor paradigms, observing which informants or actions infants are surprised by or which actions they imitate. Deception research typically focuses on three- and four-year-olds, but research into more subtle questions goes on well beyond those ages. Expertise research focuses on children old enough to allow the use of language to inform children about informants' expertise. Pedagogy researchers seek to evaluate children as young as possible, using ostensive cues such as gaze to cue trust. The epistemic trust literature is broad and the age groups investigated and methods employed are highly variable. To create an account of the development of epistemic trust we must not only account for performance across ages, but across fundamentally different tasks and phenomena.

Modeling epistemic trust

Leveraging theoretical work on the teleological stance (Gergely & Csibra, 2003; Dennett, 1989; Baker, Saxe, & Tenenbaum, 2009), Shafto, Eaves, et al. (2012) proposed a computational model of epistemic trust that in principle applies to all of these phenomena. The model explains epistemic trust in terms of inferences about informants' knowledgeability and helpfulness (Eaves & Shafto, 2012; Shafto, Eaves, et al., 2012; Asheley R. Landrum, Eaves, & Shafto, 2015). A trustworthy informant must both posses accurate knowledge about the world (be knowledgeable), and be willing and able to share his or her knowledge (be helpful). Knowledgeable informants

may not act consistently with their knowledge through lack of communicative skill or malicious intent; helpful informants may hold misconceptions, which may lead them to produce inaccurate information.

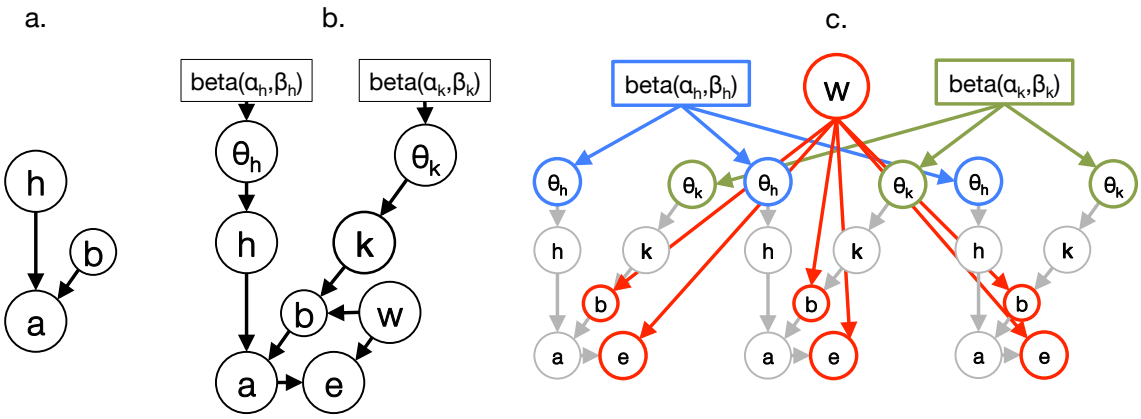


Figure 1. A graphical representation of the epistemic trust model. Informants' beliefs, b about the world, w, are determined by their knowledgeability, k. Informants' actions, a, are determined by their beliefs and their helpfulness, h. Actions on the world result in effects, e. θ_k and θ_h represent individual informants' probability of being knowledgeable and helpful, respectively. θs have beta distribution priors that represent expectations about informants in general. a) A representation of the intentional stance (Dennett, 1989) in which beliefs and desires, in this case to help or not, lead to actions. The mob) Single-informant model. c) Multi-informant model for reasoning about groups of informants. Note that beta priors on knowledgeability and helpfulness and the true state of the world, w, are shared across informants. Arrows and nodes are colored-coded for clarity.

The model is represented as a Bayesian Network (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993): a set of variables (nodes) causally linked by probabilistic relationships (edges). Edges link parent nodes to their child nodes. Figure 1a shows a graphical representation of the learner's model of how informants choose data. Informants' beliefs, b, about the world, w, are determined by their knowledgeability, k, about the world. Knowledgeable informants' beliefs align with the true state of the world; unknowledgeable informants' beliefs are determined randomly. An unknowledgeable informant's beliefs may follow a uniform distributions corresponding to a completely random guess or may follow a distribution that allows some beliefs to be less likely. For example, given the animal, *lion*, an informant should be less likely to guess the label *car*, than to guess the label *tiger*.

Informants' actions, a, are determined by their beliefs, b, and their helpfulness, h. Helpful selection of evidence is modeled using the pedagogical sampling model in Shafto et al. (2014). Helpful informants act to induce their own beliefs in learners; unhelpful informants act to induce other beliefs in learners. This is captured by the recursive equations:

$$P_{\text{learner}}(b|a) \propto P_{\text{informant}}(a|b)P(b), \tag{1}$$

$$P_{\text{informant}}(a|b) \propto \begin{cases} P_{\text{learner}}(b|a) & \text{if helpful} \\ 1 - P_{\text{learner}}(b|a) & \text{if not helpful.} \end{cases} \tag{2}$$

Informants' actions are selected conditional on their beliefs about the world. Because informants only control the action that they choose, they must consider all the possible effects of their actions. The effects are thus marginalized (summed) out. Equation 1 captures the idea that actions are

selected purposefully, with a goal (helping or deceiving), based on the informant's beliefs. Actions on the world result in effects e. The effect is determined by the true state of the world, w, and the action, a. In word learning, we do not model an effect, for unless the speaker is a wizard or has uttered some extraordinarily breathy statement, words do not themselves elicit observable effects from the world.

Prior distributions are placed on informants' helpfulness and knowledgeability, corresponding to learners' beliefs about individual informants and informants in general,

$$h|\theta_h \sim \text{Bernoulli}(\theta_h) \tag{3}$$
$$\theta_h \sim \text{beta}(\alpha_h, \beta_h) \tag{4}$$

and similarly for knowledgeability,

$$k|\theta_k \sim \text{Bernoulli}(\theta_k) \tag{5}$$
$$\theta_k \sim \text{beta}(\alpha_k, \beta_k). \tag{6}$$

The value of h and k are determined by flips of θ-weighted coins. The θs are drawn from beta distributions. These beta distributions leave the model with four free parameters: α_k, β_k, α_h, and β_h. We use the standard beta distribution parametrization, beta(α, β), which distributes probability according the function

$$f(x|\alpha, \beta) = \frac{x^{\alpha-1}\,(1-x)^{\beta-1}}{\text{B}\,(\alpha, \beta)}. \tag{7}$$

where $\text{B}(\cdot, \cdot)$ is the beta function.

Beta distributions represent the distribution of people and each informant is a draw of θ from that distribution (see Figure 1b). θ values persist across multiple demonstrations by a single informant. Keeping these rules in mind, we can link several single-informant graphs by their beta priors and by the state of the world to form a group demonstration (see Figure 1b). We can also link a number of single informant graphs by θ_k and θ_h to form successive demonstrations from a single informant. For multiple demonstrations, we need not (necessarily) link the state of the world; the state of the world is free to change from demonstration to demonstration. We can link graphs in both ways simultaneously to form successive group demonstrations.

Modeling word learning

Epistemic trust studies generally follow a similar setup. Children are introduced to one or more informants from whom they receive differing data (experience) in *familiarization* trials. Children must then choose to accept or reject information from the informant(s). For example, a child may be introduced to two informants and then observe that one informant labels common objects incorrectly while the other labels them correctly (*accuracy* trials). The child may then be presented with a novel object and asked which informant he or she would like to ask for the object's label (an *ask* trial), or similarly after having observed both informants label, the child may then be asked to label the object (an *endorse* trial). Here we discuss the process by which we model these studies.

To begin, we must make some assumptions about the world. We arbitrarily assume that at any given labeling trial there are four reasonable labels.[2] That is, $|W| = 4$ and hence there are four possible beliefs, $|B| = 4$. In word learning, each action is a label and so the number of actions (labels)

[2]We have explored the effect of increasing and decreasing the number words and found quantitative but not qualitative differences in the model output.

is equivalent to the number of world states and number of possible beliefs $|A| = |W| = |B| = 4$. We assume that the states of the world are distributed with uniform probability. No word is *a priori* more likely than any other

$$P(W) = \frac{1}{|W|}. \tag{8}$$

These assumptions result in the following relationship between the world and informants' knowledgeability and beliefs: knowledgeable informants' beliefs match the true label, w, while naive informants guess at random, uniformly from among the possible labels. The probability that an informant's belief aligns with the true state of the world is

$$P(b = w|k) = \begin{cases} 1, & \text{if } k = \text{knowledgeable} \\ 1/\left|W\right|, & \text{otherwise} \end{cases}. \tag{9}$$

As for which labels informants utter, helpful informants shall always utter the label they believe to be correct and unhelpful informants shall always utter a label they believed not to be correct,

$$P(a|h, b) = \begin{cases} 1, & \text{if } a = b \text{ and } h = \text{helpful} \\ \frac{1}{|W|-1}, & \text{if } a \neq b \text{ and } h = \text{unhelpful} \\ 0, & \text{otherwise} \end{cases}. \tag{10}$$

Again, we focus on actions and ignore effects in word learning demonstrations.

Though there are four attribute combinations based on helpfulness and knowledgeability, this formalization captures three distinct types of informant behavior. Knowledgeable and helpful informants always label correctly because they know the correct label and want the learner to know. Knowledgeable but unhelpful informants always label incorrectly because they know the correct label and do not want the learner to know. Unknowledgeable informants, regardless of whether they are helpful, may or may not label correctly because unknowledgeable informants must guess labels for objects. Unknowledgeable but helpful informants produce correct labels when they guess the correct label. Unknowledgeable and unhelpful informants produce the correct label when they guess the incorrect label and choose to produce the correct label as a foil. Thus it is difficult to determine whether an unknowledgeable informant is helpful.

In familiarization trials, the model must leverage what it knows about the world to learn about informants. In accuracy trials, informants label common objects, thus the true state of the world is known. The model can then estimate the probability with which the informant is helpful and knowledgeable.[3] This means learning the joint probability distribution for θ_k and θ_h given a and w, $p(\theta_k, \theta_h|a, w)$.

During test (ask and endorse) trials, the model must use what it has learned about the informant to learn about the world. Ask and endorse questions may seem superficially similar, but they are in fact important differences. Framed in a probabilistic context, the endorse problem is to determine the probability of each informants' label being correct given what is known about about informants in general (prior parameters) and past experience, ξ, with informant, i:

$$P(endorse_i) \propto \sum_w P(w = a|a, \alpha, \beta, \xi) \tag{11}$$

$$= \sum_{w,h,k} \iint_\theta P(w = a|a, h, k) P(h, k|\theta) P(\theta|\alpha, \beta, \xi) d\theta. \tag{12}$$

[3]Inference in the model is performed using standard approximation methods such as rejection sampling and Gibbs sampling. For details see Appendix Appendix A.

Where, for notational simplicity, we collapse similar variables and parameters such that $\theta = \{\theta_k, \theta_h\}$, $\alpha = \{\alpha_k, \alpha_h\}$, and $\beta = \{\beta_k, \beta_h\}$. The probability of endorsing informant 1 over informant 2 is,

$$P(endorse_{1,2}) = \frac{P(endorse_1)}{P(endorse_1) + P(endorse_2)}. \tag{13}$$

It is less obvious how to formalize the ask question. The question again is "who would you like to choose for information." Because one may ask an informant for a variety of reasons—i.e., because they are consistently wrong, because one wants to assess their knowledge, etc.—formalizing this question is challenging. Due to its ambiguity, we avoid modeling the ask question where we can and where we cannot we adopt the simple assumption that children choose to ask informants who are more likely to label correctly. That is,

$$P(ask) \propto \sum_{w,a} P(w = a|a, \alpha, \beta, \xi). \tag{14}$$

Inferring the label of an object given an informant's utterance is reminiscent of the referential communication and language pragmatics literature (Grice, Cole, & Morgan, 1975; Michael. C. Frank & Goodman, 2012). However, language pragmatics rely on the assumption that speakers are cooperative; the epistemic trust model does not require such a constraint to learn.

Previous developmental findings

Previous work employed this model to investigate possible explanations for developmental changes in epistemic trust. The full model and a model based on reasoning about knowledge alone were compared by searching for the parameters that best fit children's behavior in three experiments (Shafto, Eaves, et al., 2012; Eaves & Shafto, 2012). The results indicated that the knowledge-only model fit three-year-olds behavior better while the full model better fit four-year-olds'. These results are consistent with a developmental change in children's ability to reason about helpfulness.

The import of the previous modeling is limited by the fact that the model was only applied to three experiments from the literature. To broaden the scope, it is necessary to account for a wider set of heterogeneous studies. As a computational theory, the basic claim of the model is that it can explain epistemic trust behavior. That is, the model should be able to parameterize each result (locate the result in model space). Thus, it is reasonable to expect the model to predict results across the domain, regardless of the method by which they were obtained. However, it is unreasonable to expect that all methodological details are irrelevant to how the model will fit. Certain tasks may focus more on knowledge while others may focus more on helpfulness. Similarly, different methods of communicating—speaking, pointing, and marking—may elicit different degrees of trust based on past experience. While the model should explain behavior across these variations, how it explains it vis-a-vis the parameters can be expected to vary to some degree.

Of course, parameterizing the model individually in terms of each condition may not be ideal. Many free parameters raises concerns about reducing generality through over-fitting. Systematic similarities and differences among the features of experiments, such as age or communication mode, may be used as a bottom-up source of constraint on the variation in parameters. Moreover, the degree of association between the experimental features and model parameters may provide a means by which we may quantify differences in methodology or across development. We present a method for automatically identifying such similarities and differences in the next section.

Aggregation of parameterized results via Cross-categorization

How might we draw inferences about commonalities and differences among a collection of parameterized results? There are a number of possibilities, but one especially flexible and therefore attractive approach is cross-categorization. Cross-categorization (CrossCat) is a Bayesian non-parametric method for estimating the full joint probability density over tabular data (Mansinghka et

al., Accepted pending revision; Shafto et al., 2011). It simultaneously estimates dependence among variables and, among dependent variables, estimates dependence among rows. For current purposes, cross-categorization represents a method by which we can determine the probability of dependencies between the individual model parameters—such as helpfulness—and features of conditions—such as age—given a table composed of a parametrization of the results together with features of the condition. CrossCat is a more flexible tool than standard statistical approaches, such as various forms of regression, which force the user to identify which variables drive changes in others. Our goal is to learn which variables drive what kind of changes in which model parameters under what circumstances. CrossCat provides a platform to do so while seamlessly handling missing and heterogeneous data.

CrossCat is a generalization of an infinite mixture model (IMM; see Teh, Jordan, Beal, & Blei, 2006; Rasmussen, 2000; MacEachern & Müller, 1998; Neal, 2000; Anderson, 1991, for more information on IMMs) in which features' assignments to views and objects' assignments to categories within views are each inferred. Thus, CrossCat behaves as a hierarchical mixture model, where instead of assuming that there is only a single explanation for the variability over the rows, there are potentially many ways of organizing, and thus explaining the rows.

CrossCat explains a data table in terms of two main structural components: a partitioning of features (columns) into views and for each view, a partitioning of objects (rows) into categories. A view, Z, assigns the F features (columns) to $|V|$ views. The assignment of categories, V, contains $|V|$ partitions of the objects (rows), $V_0, V_1, ..., V_{|V|-1}$, such that each view $V.$ assigns the N rows to categories for the collection of features in that view. Each view models the variation in the features of that view as a mixture (those looking for a detailed treatment of cross-categorization are referred to Mansinghka et al., Accepted pending revision).

Each cross-categorization state (or sample) represents core elements of probability. The partition of features into views instantiates an inference about whether each possible pair of variables is dependent or independent. Modeling views as mixtures allows the model to identify relationships that are much more general than simple linearity. The model, therefore, allows one to generically ask key questions of interest without strong assumptions such as linearity or Gaussianity that can lead to interpretation problems. In the case of epistemic trust, for example, are age and helpfulness dependent? What is the form of that dependence? Which experiments can be explained by a common set of parameters and which require different parameters?

The data in each feature are modeled by a data-appropriate statistical model. Conjugate models are typically chosen for efficiency. For example, continuous data are modeled using a Normal distribution with a Normal-Gamma prior (Murphy, 2007; Fink, 1997), while categorical data are modeled using a Multinomial distribution with a symmetric Dirichlet prior. Many other data types can be instantiated in this framework by implementing conjugate, semi-conjugate, and non-conjugate models, as appropriate. The hyperparameters for priors are inferred to facilitate efficient inference. This produces an unusually flexible model suited to a wide variety of different types of data.

Consider a data table where each row represents a condition of an experiment and each column represents a feature of interest (experiment features or model parameters). The views would represent whether, for example, age were dependent on the helpfulness parameters by placing those features in the same view or in different views. Similarly, given a collection of samples, we could query conditional distributions to answer questions about the relationship between features. For example, we could check our previous results by asking about the relationship between age and biases toward believing informants are helpful. In this way, we use Bayesian inference to free the model from the specifics of individual studies and allow for the formulation of a general model that considers many possible hypotheses.

Method

The method consists of three steps that link a domain model (the epistemic trust model) with an analysis model (CrossCat). The domain model is used to approximate the parameter distribution

for each study, and the analysis model is used to identify trends in the parameters induced by different studies. The general method is as follows: First, select studies that can be straight-forwardly modeled with the epistemic trust model. Second, for each condition of each study, search for sets of model parameters that cause the epistemic trust model to fit the experimental data well. Third, construct and analyze a CrossCat table in which each row comprises the model parameters and experimental features of each modeled condition.

We begin by explaining the process by which studies were selected and how we determined which studies were suitable for modeling. We then describe the procedure used to search for well-fitting parameter sets. Last, we exhaustively discuss the procedure by which each study was modeled and how the model accounted for the experimental results.

Study inclusion criteria

We include for analysis studies that the epistemic trust model can capture with no extension, or simple extension by way of existing, off-the-shelf models. In previous research (Shafto, Eaves, et al., 2012), we focused on modeling three selective trust strategies (relative accuracy (Pasquini et al., 2007), familiarity (Corriveau & Harris, 2009), and consensus (Corriveau, Fusaro, & Harris, 2009)) each of which employed the ask-endorse, forced-choice paradigm and in which informants communicated either by way of verbal testimony (Pasquini et al., 2007; Corriveau & Harris, 2009) or pointing (Corriveau, Fusaro, & Harris, 2009). Different communication modes do not require extensions to the model to capture; it will be an empirical question as to how they differ in terms of the model parameters. Different paradigms necessitate minimal modifications, e.g., we model looking time as proportionate to the inverse of the probability of the event looked at. Thus, the inclusion criteria mainly focus on the informant- and information-selection strategies investigated.

There are a variety of strategies that would require involved modifications to the model and were thus omitted (see Table 1). For example, consider studies that investigate effects of domain expertise (Koenig & Jaswal, 2011). Capturing these phenomena would be quite natural within our general framework; however, expertise would require the model to be extended to capture how children believe knowledge is distributed among people. This would require multiple assumptions, and therefore expertise studies are excluded from analyses. Another group of studies uses verbal testimony by an experimenter or an additional informant to provide information about the informants. For example, some studies employ methods in which experimenters explicitly tell participants that an informant is "very mean" (Asheley R Landrum et al., 2013), "a big liar" (Mascaro & Sperber, 2009), or "a dog expert" (Koenig & Jaswal, 2011). Others have used verbal testimony about one's own beliefs, e.g., "I don't know" (Sabbagh, Wdowiak, & Ottaway, 2003; Sobel & Corriveau, 2010; Buchsbaum, Bridgers, Whalen, Griffiths, & Gopnik, 2012). Capturing the semantics of these verbal statements would require additional parameters and are therefore omitted.

We also exclude studies that investigate informant-selection strategies driven by informants' superficial qualities. For example, studies have investigated whether informants' attractiveness (Bascandziev & Harris, 2014), gender (Taylor, 2013), or accent (Kinzler et al., 2011) affect epistemic trust. While it is possible that learners attribute different knowledgeability or helpfulness to informants with certain superficial features, these features are not direct demonstrations of informants' data generation capabilities. To model how learners would learn, say, that someone dressed in a t-shirt is less trustworthy than someone dressed in a suit (McDonald & Ma, 2015) would require making assumptions about, and simulating, the types of life experiences that lead learners to acquire such biases; or worse, would require building the result of the experiment into the model.

Eight new studies met our inclusion criteria. There were three additional studies that investigated informant accuracy. (Koenig & Echols, 2003; Fitneva & Dunfield, 2010; Koenig & Harris, 2005). These studies differ in the amount of experience they provide learners—ranging from one to twelve instances of accuracy or inaccuracy—and the method employed (looking-time and forced choice).

Table 1
A list of studies excluded from analyses and the reason for exclusion.

Excluded study	**Reason for exclusion**
Sabbagh and Baldwin (2001)	**Extension**: Information from informants regarding their own ignorance
Birch and Bloom (2002)	**Extension**: *Familiarity principle* with respect to proper name referent
Robinson and Whitcombe (2003)	**Extension**: Deciding what makes an informant *better informed* and how this affects learning
Sabbagh, Wdowiak, and Ottaway (2003)	**Extension**: Information from informants regarding their own ignorance and confidence
Freire, Eskritt, and Lee (2004)	**Extension**: Information from informants regarding their own ignorance and confidence
Boseovski and Lee (2006)	**Extension**: Information from informants about the reliability of other informants
Jaswal and Neely (2006)	**Extension**: Epistemic beliefs about adults vs. children
Baum, Danovitch, and Keil (2008)	**Extension**: Quality of explanation
Corriveau, Meints, and Harris (2009)	**Extension**: Labeling vs. drawing attention
Eskritt, Whalen, and Lee (2008)	**Extension**: Relevance and quantity of information
Fusaro and Harris (2008)	**Extension**: Nonverbal information from bystanders regarding others' testimony
Kushnir, Wellman, and Gelman (2008)	**Extension**: Information from informants regarding their own ignorance and confidence. Perceptual access. Assistance from participant.
Mills and Keil (2008)	**Extension**: Impartiality and interpersonal biases
Birch, Akmal, and Frampton (2010)	**Extension**: Informant confidence
Nurmsoo and Robinson (2009)	**Extension**: Perceptual access
Poulin-Dubois and Chow (2009)	**Non-epistemic**: Informant excitement
Kinzler, Corriveau, and Harris (2011)	**Non-epistemic**: Speaker accent
Sobel and Corriveau (2010)	**Extension**: Information from informants regarding their own ignorance
Krogh-Jespersen and Echols (2012)	**Extension**: Second–label learning
Mills and Landrum (2012)	**Extension**: Informant perceptual capability and objectivity
Asheley R Landrum, Mills, and Johnston (2013)	**Extension**: Information from experimenters regarding informants' benevolence
Lane, Wellman, and Gelman (2013)	**Extension**: Information from experimenters regarding informants' honesty. Perceptual access.
Kim and Harris (2014)	**Extension**: Supernatural abilities
Boseovski and Thurman (2013)	**Extension**: Learning from informants in potentially dangerous situations

Two studies interleaved feedback between learner's guesses (Couillard & Woodward, 1999; Tummeltshammer et al., 2014). These included studies where the individual was repeatedly, implausibly incorrect (i.e. deceptive) (Couillard & Woodward, 1999) and where trust was measured in looking-time (Tummeltshammer et al., 2014), as opposed to the standard forced-choice, ask-endorse approach. To model these, we updated the model's beliefs about the knowledgeability and helpfulness of the informant by conditioning on the feedback, w, between each trial.

In addition to the consensus experiment (Corriveau, Fusaro, & Harris, 2009) modeled in our previous work, we model two additional studies investigating consensus (Chen et al., 2012; DiYanni et al., 2015). These are modeled as in Shafto, Eaves, et al. (2012), by simply considering the probability of agreement and disagreement among more than one informant.

Finally, we include a study that investigates the effect of error magnitude on epistemic trust (Einav & Robinson, 2010). This study investigated the degree of the error, and thus required extending the model with a notion of semantic similarity. We employ an existing psychological model of semantic relatedness (Griffiths, Steyvers, & Firl, 2007; Collins & Loftus, 1975). This extension allows the epistemic trust model to assess both errors and their degree. Concepts that are closer in a semantic network are more similar and errors between similar concepts are more reasonable.

These studies are heterogeneous in terms of their features: the ages of the children, the communication mode, and the experimental paradigm. Ages span 8 months to adult. Communication modes include verbal testimony, pointing, gaze, and use of markers. Paradigms include forced-choice (ask and/or endorse), and looking time. Table 2 lists the set of studies included—a total of 11 studies comprising 24 conditions.

Table 2
List of study conditions included in analyses divided into conditions. Columns list the study, the age in years of the participants, the communication mode, and the optimal parameters. (c) represent WEIRD cite children and (a) represents Asian children.

Study	strategy	Age (y)	Comm mode	Paradigm
Tummeltshammer, Wu, Sobel, and Kirkham (2014)	reliable gaze	.75	gaze	looking-time
Koenig and Echols (2003)	accuracy	1.5	verbal	looking-time
Pasquini, Corriveau, Koenig, and Harris (2007)	relative accuracy	3	verbal	ask-endorse
Pasquini, Corriveau, Koenig, and Harris (2007)		4	verbal	
Koenig and Harris (2005)	accuracy	3	verbal	ask-endorse
Koenig and Harris (2005)		4	verbal	
Corriveau, Fusaro, and Harris (2009)	consensus	3	points	ask-endorse
Corriveau, Fusaro, and Harris (2009)		4	points	
Couillard and Woodward (1999)	points v. markers	3	markers	ask-endorse
Couillard and Woodward (1999)		4	markers	
Couillard and Woodward (1999)		3	points	
Couillard and Woodward (1999)		4	points	
Corriveau and Harris (2009)	familiarity	3	verbal	ask-endorse
Corriveau and Harris (2009)		4	verbal	
Corriveau and Harris (2009)		5	verbal	
DiYanni, Corriveau, Nasrini, Kurkul, and Nini (2015)	consensus v culture	5 (c)	verbal	ask-endorse
DiYanni, Corriveau, Nasrini, Kurkul, and Nini (2015)		5 (a)	verbal	
Chen, Corriveau, and Harris (2012)	consensus	4	points	ask-endorse
Chen, Corriveau, and Harris (2012)		6	points	
Einav and Robinson (2010)	error magnitude	4-5	verbal	ask-endorse
Einav and Robinson (2010)		6-7	verbal	
Fitneva and Dunfield (2010)	accuracy	4	verbal	ask-endorse
Fitneva and Dunfield (2010)		7	verbal	
Fitneva and Dunfield (2010)		19-22	verbal	

Data preparation and model fitting

We divided the 11 studies into analysis units, which we refer to as conditions. For example, an experiment which separately reported results for three- and four-year-olds consists of two conditions. This resulted in 24 total conditions. We fit the model parameters by searching for the parameters that best reproduced the data.

Our choice of search method is dictated by the complexity of the inference problem and the heterogeneity of the studies we model. Often the distribution of an informant's helpfulness and knowledgeability cannot be calculated analytically and must be approximated. Exact calculation of probabilities requires enumerating over each unknown variable. In the case of Einav and Robinson (2010), enumerating over possible beliefs and the binary values of helpfulness and knowledgeability for four labeling trials leaves more than 10^{17} terms to evaluate. We approximate probabilities using Monte Carlo simulation (see Appendix A). In simpler situations, one may employ direct fit methods that search for local error minima by traversing the path of steepest descent. These methods require calculating the gradient of the probability space with respect to the parameters. For the same reason we cannot calculate the probabilities exactly, we cannot calculate their gradients exactly. Grid search is an alternative technique in which a finite grid of search points is placed over the parameter space and the target function is evaluated at each point. We employ a randomized version of grid search, random search (Bergstra & Bengio, 2012), in which random points in the parameter space are evaluated. In practice, grid search and random search perform similarly with respect to error, but random search offers additional flexibility in that it more easily allows us to exploit knowledge of which areas of the parameter space require more thorough search.

The random search procedure we applied involved generating a large number of parameter sets, running the model for each experiment for each parameter set, and calculating the errors between the model prediction and the empirical data. We generated 4000 parameter sets from independent exponential distributions with mean 5. That is, for each parameter in the parameter set, $\{\alpha_k, \beta_k, \alpha_h, \beta_h\}$ was drawn from $\text{Exp}(\frac{1}{5})$. We choose this specific parameter-generating distribution because it applies higher probability to lower-valued parameters but also represents higher values. Higher parameters values are more robust; small changes in high-valued parameters affect the model results less than small changes in low-valued parameters. Note that we focus only on the full, four-parameter model because previous research demonstrates that a knowledge-only model, which does not account for variable helpfulness, fails to account for development (Shafto, Eaves, et al., 2012).

We searched for parameters that minimized the summed relative error of each experiment rather than the parameters that maximize probability because the studies report different measures (e.g. proportions of participants and looking times). The relative error of two values, a and $b \neq 0$ is the absolute value of one minus their ratio $|1 - a/b|$. If a/b is 1 then $a = b$. [4] We use relative error rather than squared or absolute error because experiments' dependent measures are not always identically scaled. One experiment may report the proportion of children who asked a particular informant for information while another may report the number of seconds an infant looked at an informant. We use relative error so that error is calculated similarly regardless of the result metric employed by the study. We use the sum of error so that the error of each data point (bar in a bar chart) carries equal weight. An experiment with more bars should be weighted higher for error minimization.

To construct the cross-categorization table, we took the five[5] best-fitting parameter sets for each condition and arranged them in a table. Each row represented a single parameter set for a condition and was augmented with demographic features of the experiment. These features included

[4] We subtract 1 from this quantity because 1 is the point that represents the zero difference between a and b. We take the absolute value because we are not concerned with the direction of the error, only its magnitude. The sum relative error between two n-length vectors of values $\mathbf{a}$ and $\mathbf{b}$ is then, $\sum_{i=1}^{n} |1 - \mathbf{a}_i/\mathbf{b}_i|$.

[5] We took the top five parameter sets to capture both the best fitting parameters and variability in fit across the parameters.

the mean age of participants, communication mode, culture, and experimental paradigm. Thus, each column was a parameter or a demographic or experimental feature of interest (see Table 3).

For ease of interpretation, we converted the model's α and β parameters on knowledgeability and helpfulness to *strength* and *balance* (Kemp, Perfors, & Tenenbaum, 2007). The strength (s) and balance (b) parameterization of the beta distribution is $s = \alpha + \beta$ and $b = \frac{\alpha}{\alpha+\beta}$. Balance corresponds exactly the mean of the beta distribution and takes on values in the open interval $(0, 1)$. For example, a balance parameter on knowledgeability, b_k, closer to 1 means that the learner believes that informants are, in general, knowledgeable while a b_k closer to 0 implies that the learner believes that informants are, in general, unknowledgeable. Strength roughly corresponds to the invariance in beliefs and lies in the interval $(0, \infty)$. For example, a very high value of s_k—the strength parameter on knowledgeability—implies a very strong belief that all people are the same—either all knowledgeable or all unknowledgeable as determined by b_k.

Table 3
Structure of the prepared table used during cross-categorization. One row of five from each study is represented for demonstrative purposes. Columns correspond to strength and balance parameters for knowledgeability and helpfulness, and experiment-age identifier, age in years, communication mode, and experimental paradigm. The WEIRD *acronym (Henrich, Heine, & Norenzayan, 2010) indicates Western, Educated, Industrial, Rich, and Diplomatic.*

age	culture	comm. mode	paradigm	s_k	b_k	s_h	b_h
4.98	Asian	Verbal	Forced-choice	16.26	0.99	8.54	0.27
4.62	WEIRD	Verbal	Forced-choice	3.90	0.29	2.05	0.30
3.3	WEIRD	Marker	Forced-choice	12.92	0.83	1.41	0.95
4.05	WEIRD	Marker	Forced-choice	6.30	0.95	13.90	0.12
3.3	WEIRD	Points	Forced-choice	7.55	0.97	8.10	1.00
4.05	WEIRD	Points	Forced-choice	7.81	0.95	5.40	0.73
4.92	WEIRD	Verbal	Forced-choice	9.11	0.06	9.90	0.00
7.0	WEIRD	Verbal	Forced-choice	18.70	0.43	6.35	0.02
3.34	WEIRD	Verbal	Forced-choice	4.45	0.24	1.65	0.64
4.42	WEIRD	Verbal	Forced-choice	1.90	0.86	0.17	0.43
5.67	WEIRD	Verbal	Forced-choice	4.78	0.78	0.24	0.17
3.5	WEIRD	Points	Forced-choice	7.81	0.74	1.98	0.32
4.58	WEIRD	Points	Forced-choice	8.90	0.85	2.97	0.43
0.67	WEIRD	Gaze	Looking-time	15.32	0.90	0.52	0.27
1.5	WEIRD	Verbal	Looking-time	22.57	0.66	7.05	0.99
6.04	WEIRD	Points	Forced-choice	2.41	0.26	1.90	0.61
4.15	WEIRD	Points	Forced-choice	20.46	0.07	7.83	0.69
3.5	WEIRD	Verbal	Looking-time	15.66	0.06	25.52	0.24
4.5	WEIRD	Verbal	Looking-time	6.80	0.91	6.01	0.20
3.5	WEIRD	Verbal	Forced-choice	0.18	0.04	3.17	0.65
4.42	WEIRD	Verbal	Forced-choice	3.38	0.96	4.18	0.08
4.44	WEIRD	Verbal	Looking-time	9.52	0.77	4.16	0.50
7.29	WEIRD	Verbal	Looking-time	36.96	0.96	3.12	0.13
20.0	WEIRD	Verbal	Looking-time	1.90	0.86	0.17	0.43

Modeling individual studies

In this section we explain the procedure by which each study used for analyses was modeled and how the model captures each empirical result. This section is intended not only for those who wish to reproduce our procedure but also for those who seek an intuitive understanding of how the model works.

For each study we display results given the best-fitting parameters, and when possible, we display standard error bars given those parameters. As we have discussed, the heterogeneous nature of the literature forces individual fitting. The approach we take is distinct from the standard modeling approach in which a model's validity is measured by its fit, in which the validity of the fit is measure in terms of whether the its parameter values make intuitive sense and whether it cross-validates. In the approach we take, these concerns have no influence on the analysis. The model's ability to fit the results of individual studies is not our primary interest, but a necessary precondition for aggregating results—to include a study in analyses, the model must be able to account for its results. Our goal is to look at trends in regions of fit in which the model captures experimental results—regardless of where they are in parameter space—and to determine if these trends have implications for development.

Accuracy. Koenig and Harris' (2005) study on children's preference to ask for and endorse information from accurate sources is a seminal work in the trust-in-testimony literature. For three trials children observed two informants label common objects, e.g., a ball and a cup. One informant labeled each object correctly and the other labeled each object incorrectly. After these *accuracy* or *familiarization* trials, a novel object was placed before the informants. The child was either invited to choose the informant whom she would like to ask for the label (*ask* trial) or after having observed each informant provide his own label, the child was invited to label the object herself (*endorse* trials).

This study maps easily to inference in the epistemic trust model. We have only to account for data that does or does not match the state of the world. Participants observed novel informants, thus there is no need to account a prior bias that one informant should be more likely than the other to label correctly. Additionally, each informants' incorrect answers are equally incorrect (labeling a ball as a shoe is just as foolish as labeling a cup as a dog) therefore there is no need to account for the relative *magnitude* of errors, which we account for in a later section. Endorse questions are modeled as described in the section on modeling word learning.

During accuracy trials, children learn about their informants. The model is concerned with learning the probability distribution defining each informants' tendency toward or away from helpfulness and knowledgeability given the state of the world (the object) and the label uttered by the informant. This means collecting information about k and h given w and a.

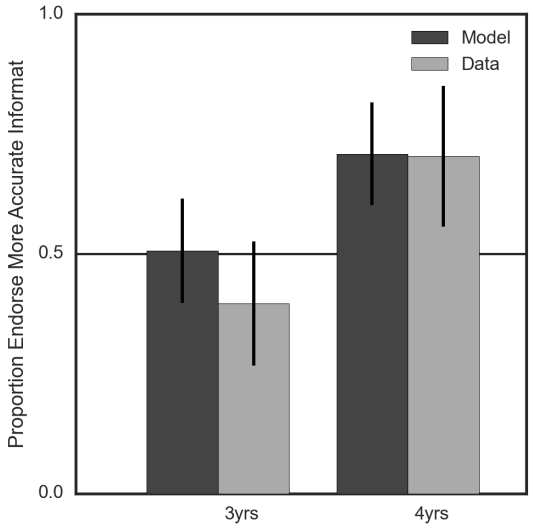


Figure 2. Model simulation results for Koenig and Harris (2005). The y-axis represents the proportion of children who endorsed the answer given by the accurate informant, or for the model, the probability of endorsing the accurate informant.

We see the model results along side the experimental results (Koenig & Harris, 2005, Experiment 1) in Figure 2. For both age groups, the model prefers to endorse the label provided by the accurate speaker. The model infers that an informant who always labels accurately is likely knowledgeable and helpful and that an informant who always labels inaccurately is not. In fact, an informant who repeatedly labels incorrectly is assumed to be knowledgeable and unhelpful—deceptive. An unknowledgeable and helpful informant will produce the correct label by correctly guessing—an informant chooses a label from a fixed set of labels of which only one (or a few) is correct.

This preference for more accurate informants has been documented after even a single encounter (Fitneva & Dunfield, 2010). In Fitneva and Dunfield (2010) children were shown an image and told a corresponding story. A sticky note occluded part of each image. The child asked two informants (children on a computer screen) what was under the card. The two informants answered differently. The sticky note was removed, revealing that one informant had been correct and the other had been incorrect. The procedure was then repeated but the child was allowed only to ask one informant. For this study we modeled ask questions. The results, averaged over three experiments can be seen in Figure 3. [6]

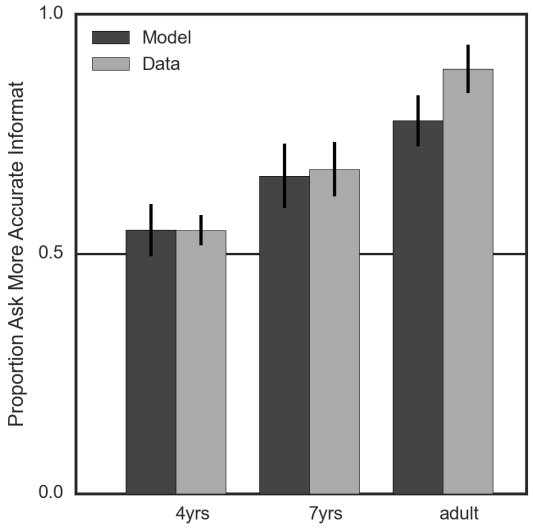


Figure 3. Model simulation results for Fitneva and Dunfield (2010). The y-axis represents the proportion of children who asked the previously accurate informant, or for the model, the probability of asking the accurate informant.

We see that the model captures people's preference for the accurate informant as well as an increasing preference with age. A theme in the literature is that the speed with which people update their beliefs about informants given data increases with age.

Relative accuracy. Informants are not deterministic. They are not always correct or always incorrect; they provide information with some amount of noise. Pasquini et al. (2007) extended the paradigm of Koenig and Harris (2005) to account for variable levels of relative accuracy between informants. Children were introduced to two informants who labeled four common objects with variable accuracy. Informants labeled either 100%, 75%, 25%, and 0% accurately, corresponding to four, three, one, and zero of four objects correctly labeled, respectively. There were four conditions 100% vs 0% accurate, 100% vs 25% accurate, 75% vs 0% accurate, and 75% vs 25% accurate. For example in the 100% vs 25% accurate condition, the child observed one informant label each object

[6]The procedure was identical for each experiment in Fitneva and Dunfield (2010), only the wording changed.

correctly and the other label only one of the four objects correctly. After accuracy trials, a novel object was placed before the child who then participated in ask and endorse trials.

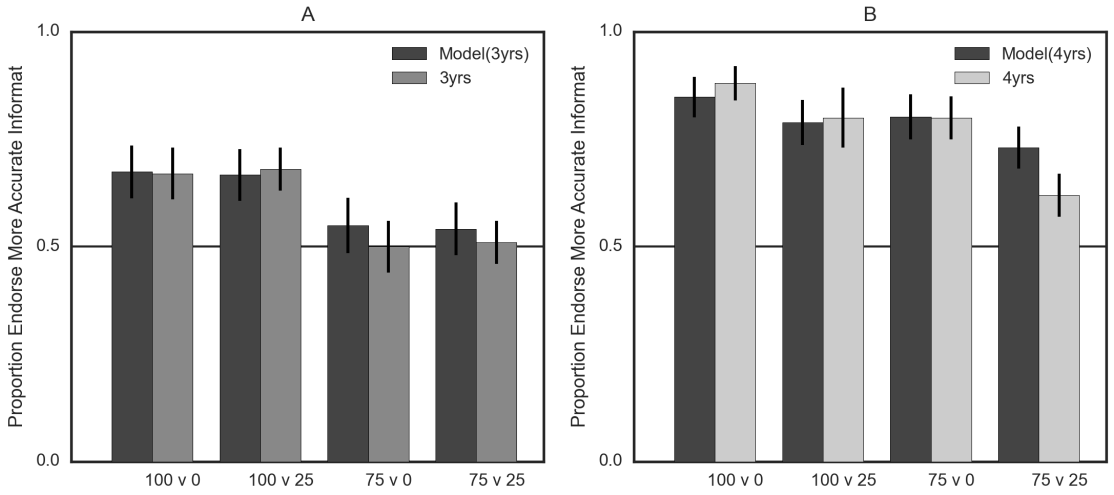


Figure 4. Model simulation results for Pasquini, Corriveau, Koenig, and Harris (2007). a) Three-year-olds. b) Four-year-olds. The y-axis represents the proportion of children who endorsed the answer given by the accurate informant, or for the model, the probability of endorsing the accurate informant. Error bars represent standard error.

The model shows a preference for the more accurate informant (Figure 4). We see a tiered effect in both three-year-olds' behavior and model prediction. In previous research, we found that 3-year-olds' behavior is best represented by a model with a strong bias toward believing all informants are helpful (Shafto, Eaves, et al., 2012). This means that the model predicts three-year-olds' inferences about informants primarily based on knowledgeably. Informants are either knowledgeable or not. An informant who always labels correctly is knowledgeable, all other informants are not. This causes difficulty in creating a grading between the different accuracy levels.

The model predictions show a rather different trend for four-year-olds. The results closely follow the data, plateauing where there is a 75% difference in relative accuracy between informants.

Familiarity. Corriveau and Harris (2009) investigated the interaction between familiarity and accuracy. For their study, Corriveau and Harris (2009) chose children's preschool teachers to play the role of familiar informants. Familiarity is formalized as prior experience. In this case specifically, because the familiar informants were teachers—not tricky uncles—we modeled familiarity as experience demonstrating helpfulness and knowledgeability. This manifests mathematically as an altered prior. This manipulation is straight forward to implement as a beta distribution posterior update. As a demonstration, assume that we have witnessed an informant be helpful twenty times and unhelpful once. Given a base prior of $beta(\alpha_h, \beta_h)$ the posterior distribution is simply $beta(\alpha_h + 20, \beta_h + 1)$. We used this procedure for both knowledgeability and helpfulness. The result is a strong bias and requires more data to override than the presumably weaker bias for an unfamiliar informant.

Before any familiarization or accuracy trials, children were given ask and endorse questions to gage their natural preference for the familiar informant (*pretest*). Children were then given four familiar object labeling trials in which the familiar informant labeled each object accurately and the novel informant labeled each object inaccurately (familiar 100%) or in which the converse occurred (novel 100%). If children hold a more biased belief that their teacher is helpful and knowledgeable, they should prefer to ask and endorse their teacher at pretest. Observing the teacher label common

objects correctly should reinforce this bias and observing her labeling them incorrectly should work to relax or reverse the bias.

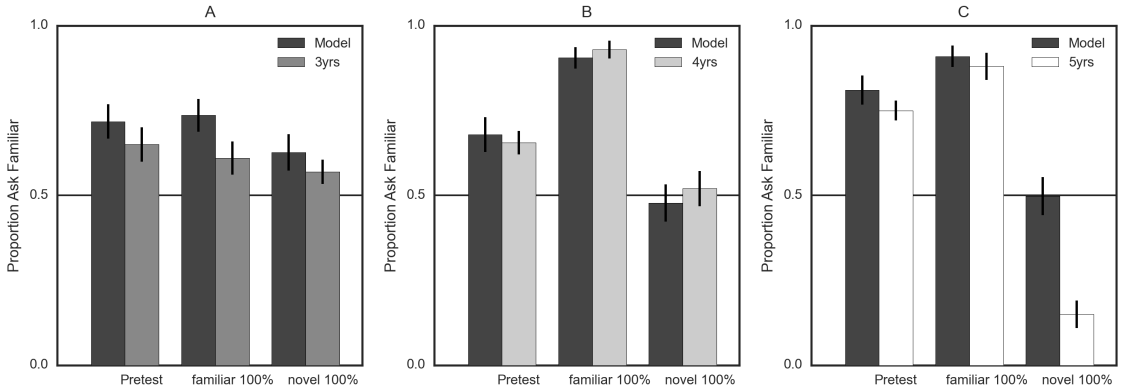


Figure 5. Model simulation results for Corriveau and Harris (2009). a) Three-year-olds. b) Four-year-olds. c) Five-year-olds. The y-axis represents the proportion of children who endorsed the answer given by the familiar informant, or for the model, the probability of endorsing the familiar informant. Error bars represent standard error.

We see in Figure 5 the model captures trends across several ages but fails to capture the sharp reversal made by five-year-olds when the familiar informant labels inaccurately in the novel 100% condition. A possible reason for this is that to minimize complexity we have applied the same familiar prior for each age group. It is reasonable to assume that children of different ages have different experiences with their teachers or handle familiarity in a more flexible way. Whether this holds true is an question for future research.

Consensus. Corriveau, Fusaro, and Harris (2009) looked at children's preferences for members of a group over rogue dissenters. For four trials, three novel objects were laid out before a group of four informants. On each trial an experimenter asked "Which is the [novel object label]", after which, each informant pointed simultaneously to an object. Three informants pointed to the same object and the other pointed to a different object. On each trial the same informants agreed and the same informant dissented. It is important to emphasize that informants testified through pointing rather than vocalization. We did not model points differently than verbal communication. After these group (pretest) trials children observed as two of the informants, one of whom had belonged to the agreeing group and the dissenter, labeled additional novel objects (test trials). Children again chose the object that they believed corresponded to the label.

We see the model results in Figure 6 (a and b). Because the objects were novel, children could not leverage their knowledge of the world to learn about informants. However, the fact that children learned from a group of informants labeling the same objects provides extra power not only for learning about novel objects but learning about informants as well. In the case of a group consensus we can exploit informant dynamics. In general, it is unlikely for multiple independent informants to repeatedly converge on the same object unless they are both helpful and knowledgeable. This leads logically to the conclusion that our dissenter is either unknowledgeable, unhelpful, or both; and that the agreeing informants are pointing at the correct object.

As a simple illustration of why this is so, let us categorize informants into two groups: reliable and unreliable. Further assume that reliable informants always point to the correct object and that unreliable informants point uniformly at random. We assume that informants are reliable and unreliable with equal probability. Given three objects to choose from, the probability that three reliable informants converge on the same object is 1, the probability that three unreliable informants converge on the same object is $\binom{3}{1}\left(\frac{1}{3}\right)^3 = \frac{1}{9}$. The probability that unreliable informants converge

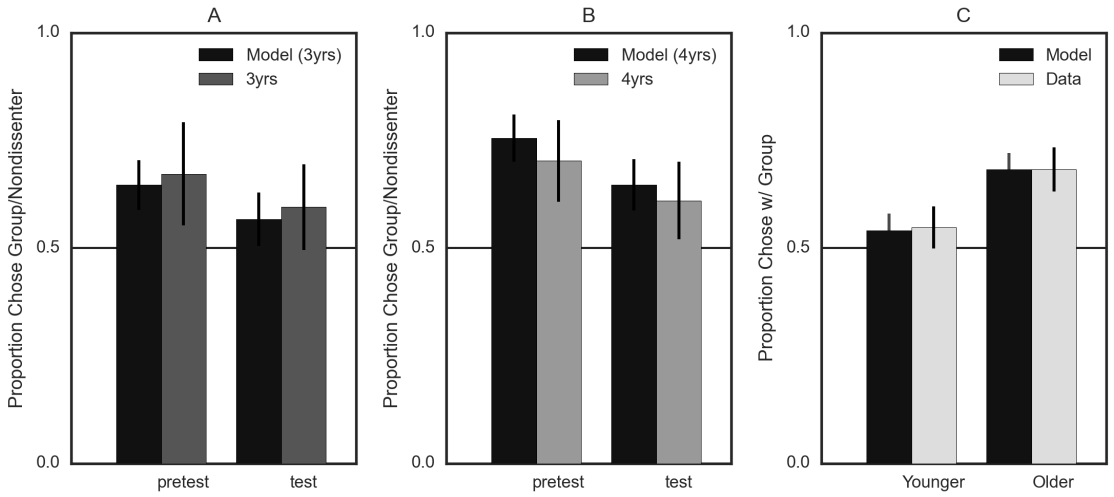


Figure 6. Model simulation results for Corriveau, Fusaro, and Harris (2009) and Chen, Corriveau, and Harris (2012). a) Corriveau, Fusaro, and Harris (2009), three-year-olds. b) Corriveau, Fusaro, and Harris (2009), four-year-olds. c) Chen, Corriveau, and Harris (2012) Younger and older groups.

on the same answer for four trials is then $\left(\frac{1}{9}\right)^4 = \frac{1}{6561}$.

Things are not so black and white in the model so this effect is softened. In the model, informants are not so neatly categorized as reliable and unreliable. There are different degrees and sources of unreliability that bring about different types of unreliability, e.g. the difference in behavior between unknowledgeable and unhelpful informants. This additional uncertainty is reflected in the results by a less distinct preference to choose with the group at pretest and the informant from the group at test. Additionally, the certainty of these inferences is dependent to an extent on prior beliefs about informants. The higher the prior toward knowledgeability and helpfulness, the higher the probability that agreeing informants are knowledgeable, helpful, and correct. This of course assumes uniform probability over labels. It is possible that there may be some wrong belief with a high prior probability that unknowledgeable informants could converge on (for example, that in the time of Christopher Columbus it was common knowledge that the Earth was flat).

We also modeled the results of Chen et al. (2012) which reproduced the pretest (group) trials of Corriveau, Fusaro, and Harris (2009) with different age groups. The model procedure was identical. The results can be seen in Figure 6c. Again, the model captures a bias toward choosing with the group, which appears to increase with age.

Culture. It is not enough to demonstrate that a model fits data; the model should fail to capture results outside of its scope. Here we demonstrate how our epistemic trust model fails to account for non-epistemic, cultural behavior.

DiYanni et al. (2015) looked at culture effects in children's deferring to consensus. Children observed three informants choose a tool to crush a cookie. The tool was either functionally affordant (hard plastic) or non-affordant (a mass of plush, fuzzy balls). Each of the three informants had a cookie in front of them. The first informant selected the affordant tool and tapped the cookie twice with it then repeated the procedure with the non-affordant tool. The cookie remained intact. The informant then held the non-affordant tool and said "This is the one I would need". This process was repeated with the other two informants. Children were then asked which tool would be best for crushing the cookie. A similar condition was conducted but with a single informant. The hypothesis was that children in both culture groups would similarly reject the advice of a single informant claiming that the non-affordant tool was best, but that for cultural—not epistemic—reasons Asian-

American children would be less likely to dissent from the group. For modeling purposes we treat this task as equivalent to labeling. The effect is the same in each case, the cookie remains intact, and can be ignored. Informants explicitly label the non-affordant tool as "the one I would need", which we interpret as a novel object labeling task in which one of the objects is "the best for crushing cookies". Children's bias for the affordant tool plays a major role and so we modeled the bias based on previous research using the same tools in which "[. . .]89% of 3-4-year-olds choose to use the Functionally-Affordant tool over the Non-Affordant tool to crush a cookie when both tools are modeled with equal intention" (DiYanni et al., 2015; DiYanni & Kelemen, 2008). The prior probability on w was left uniform because both tools are equally novel, but $P(b|\neg k, w)$ was altered such that an unknowledgeable informant should guess the affordant tool was best 89% of the time.

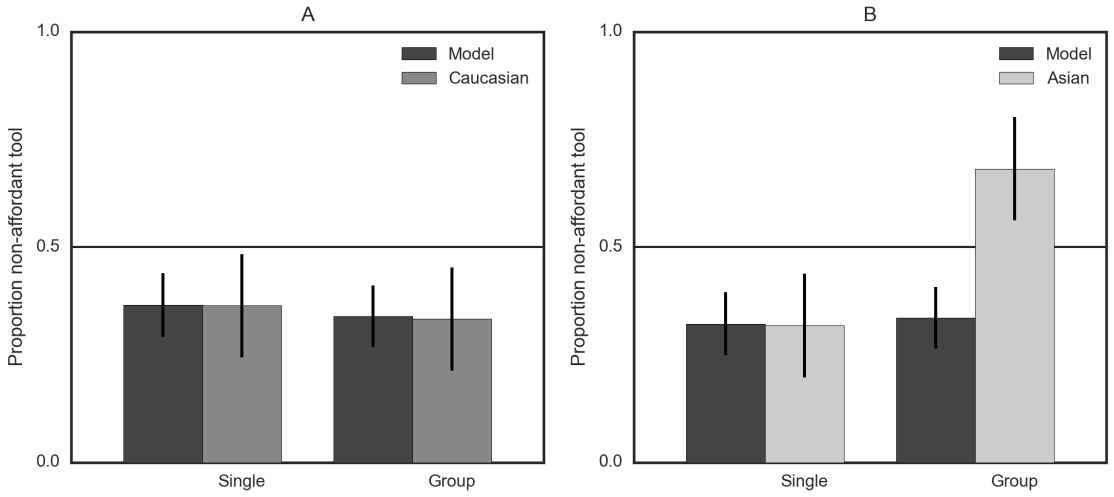


Figure 7. Model simulation results for DiYanni, Corriveau, Nasrini, Kurkul, and Nini (2015). a) Caucasian children. b) Asian Children. Error bars represent standard error.

Both groups of children were equally likely to dismiss the advice of a single informant, but Caucasian-American children more often rejected the advice of the consensus than did the Asian-American children. DiYanni et al. (2015) suggest that this result stems from a cultural stigma with respect to deviancy in the Asian community. The model can only venture to capture these results as modified prior beliefs (see Figure 7).

The model captured American-Caucasian children's disagreement with both the single informant and the group but fails to capture Asian-American children's agreement with the group. The model fundamentally fails to capture Asian-American children's behavior. The study noted that Asian-American children's conformity is likely a symptom of their avoiding appearing deviant (DiYanni et al., 2015)—not an epistemic goal.

It is important that the model fails to capture this result because the result is non-epistemic. This result illustrates that the model has limitations; it cannot explain all patterns of results. It is likely that group membership studies do not capture differential learning but simply the effect of social norms. Other research would suggest that children have no difficulty in appeasing a group of seemingly unreliable informants, but do not allow it affect their learning. Corriveau and Harris (2010) demonstrated that though children may appear to defer to a group whose consensus violates their own perceptions (in the study, the group agreed a shorter line was longer than a longer line), children rely on their own perceptions when solving a pragmatic task. Though children agreed with the group that a shorter line was the longest, children then used the longest line to construct an adequate bridge to help a bunny cross a gap.

Deceptive pointing and marking. In Couillard and Woodward (1999)'s study on children's interpretation of deceptive points, a child plays a game of Two Cup Monte with an informant. Behind a screen, the informant hides a sticker under one of two cups. The screen is taken away and the informant points to one of the cups. Children's job is to choose the cup under which the sticker is hidden. For each time children choose correctly they get to keep the sticker. This procedure repeats for ten trials. On each trial the experimenter indicates the empty cup. We assume that a point acts as a label and we assume that the informant is knowledgeable because children observe the informant place the sticker (though they do not observe under which cup). The knowledgeability bias is applied to the prior. Children receive feedback after each trial. The experiment is iterative. Each trial consists of an endorse question (choose to endorse or reject the informant's testimony) and a subsequent familiarization demonstration in which the child is given information regarding the veracity of the informant's testimony. Because the bias toward knowledgeability has been strongly influenced by the informant's hiding the sticker, children must make inferences primarily through inferences with respect to helpfulness. The informant knows the location of the sticker but does not want learners to know. Children at three-years-and-three-months of age were more often fooled by the informant than children closer to four-years of age.

The experiment was repeated with a markers condition in which the informant placed a marker to indicate a cup rather than pointing to it. Younger children were far more likely to choose the correct cup in the markers condition. We make no fundamentally different modeling assumptions to capture this result, but allow it to manifest as an alternate parameter set.

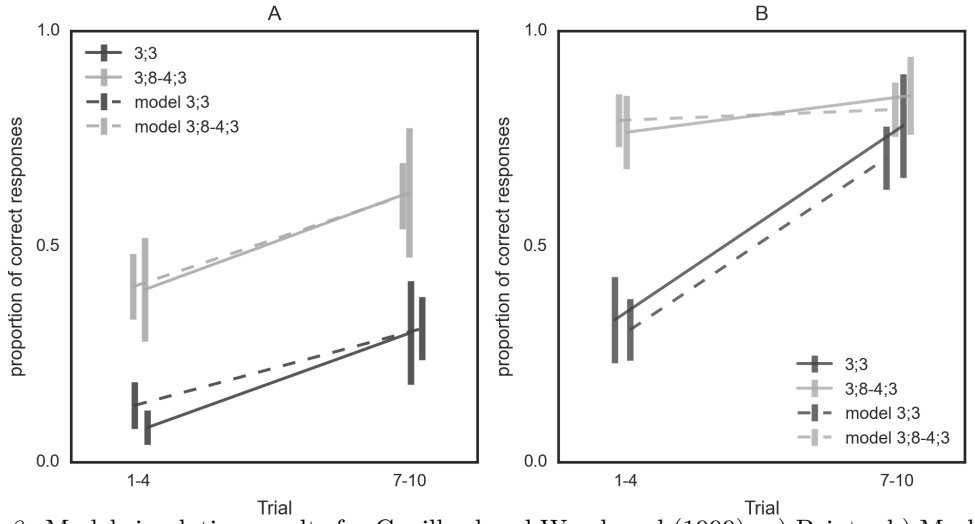


Figure 8. Model simulation results for Couillard and Woodward (1999). a) Points. b) Markers. The x-axis shows the trial number collapsed into blocks. The y-axis displays the proportion of children who choose the cup opposite the cup indicated by the informant, or for the model, the probability that the marker is in the cup opposite the cup indicated by the informant.

Figure 8 shows the proportion of children who chose the correct cup (the cup not indicated by the informant) averaged across the first four and last four trials. We see that the model captures the rate of learning. At each trial the learner is given extra information about the informant which it uses to learn about the world. The informant is reliably inaccurate. An informant who repeatedly labels incorrectly is likely deceptive. Because a deceptive informant never labels correctly, the model infers that the opposite cup is more likely. Younger children have a stronger belief that informants are helpful. A stronger belief requires more data to overcome, thus we see that younger children more often choose with the informant, though they choose with the informant less as trials progress.

Error magnitude. Einav and Robinson (2010) looked at the effect of error magnitude on children's informant preferences. For example, labeling a lion as a tiger is a smaller magnitude error than labeling a lion as a mouse or a clock. The structure of the study was nearly identical to that of Pasquini et al. (2007). Children observed two informants label common animals for four trials. On each trial after the first, both informants labeled incorrectly but one informant produced higher magnitude errors. For example, given the labels "dog", "tiger", "horse", and "butterfly", the more accurate informant provided the labels "dog", "lion", "cow", and "bee", while the less accurate informant either provided the labels "dog", "mouse", "fish", and "cat" (animal-animal condition) or "dog", "clock", "fork", and "car" (animal-object condition).

Some words are more prevalent than others. If one was asked to provide a word starting with the letter 'A' one may be more likely to respond 'Apple' than 'Appendectomy'. To capture that some labels are more inappropriate in response to certain cues, we must formalize a meaningful relationship between words. Griffiths et al. (2007) had success using semantic networks and *pagerank* (Page, Brin, Motwani, & Winograd, 1999; Sloman, Love, & Ahn, 1998).

The lexicon can be organized into a network where associated words share links. We can represent a network containing n words as a $n \times n$ matrix $\mathbf{L}$ where $\mathbf{L}_{ij}$ is 1 if there is a link from word j to word i and 0 otherwise. Pagerank captures that important words have more incoming links and that importance travels along these links. Pagerank is thus recursively defined: important nodes have more links incoming from important nodes. If $\mathbf{M}$ is a matrix where $\mathbf{M}_{ij}$ is the total proportion of importance that travels through $\mathbf{L}_{ij}$, then

$$\mathbf{M}_{ij} = \mathbf{L}_{ij} \Big/ \sum_{k=1}^{n} \mathbf{L}_{kj}, \tag{15}$$

and Pagerank is the solution for $\mathbf{r}$ in the recursive equation,

$$\mathbf{r} = \mathbf{M}\mathbf{r}. \tag{16}$$

Now that we have defined a prior probability distribution on cues, $p(\text{cue})$, we must define a sampling distribution (likelihood) for labels given cues, $p(\text{label}|\text{cue})$ which is exactly $P(b|\neg k, w)$:[7] the probability of an unknowledgeable informant believing a particular label given the cue, w. For this we apply the idea of *spreading-activation* (Collins & Loftus, 1975) in which activation—which is directly analogous to importance—flows from node to node in the network. We can construct an activation-based sampling distribution by assuming that the probability of a label given a cue is determined by the minimal path length from the cue to the label in the network. That is, the closer the label is to a cue in a network, the higher its probability. More formally, if we assume that activation decays at the same rate across every edge, then for the set of edges, D, that defines the minimal path from *cue* to *label*, the probability of *label* given *cue* is,

$$P(\text{label}|\text{cue}) \propto \gamma^{|D|}, \tag{17}$$

where $|D|$ is the number of links in the path ($|D| = 0$ if $label = cue$) and $\gamma \in [0, 1]$ is a decay constant capturing that activation decreases as a function of distance. We arbitrarily chose $\gamma = .5$, which corresponds to losing half of the signal at each jump. This formalization of the belief probabilities implies that low-magnitude errors are most indicative of a helpful, unknowledgeable informant while high-magnitude errors are most indicative of unhelpful informants. A knowledgeable informant knows the correct label, an unknowledgeable informant is likely to guess a close label; in both cases, unhelpful informants will choose a label to lead learners away from their own beliefs: a label distant from the true label or distant from a close label.

The network used here was constructed from the University of South Florida free association norms database (Nelson, McEvoy, & Schreiber, 2004), which comprises free associations for 5019

[7] $\neg k$ is the negation of k or *not knowledgeable*.

cue words. We only included words that were both cues and responses, leaving 4870 words. Links were created from targets to responses. We used the python package *NetworkX* (Hagberg, Swart, & Chult, 2008) to construct the network, find minimal paths, and calculate pagerank. This allowed us to model the study using the exact words used in the study rather than word analogs as we did in the previous studies. For example, given this model we can ask for the probability that an informant is knowledgeable and helpful given that she labeled a lion as a tiger, $P(k, h|a = \text{tiger}, w = \text{lion})$, instead of asking about a label indicies, $P(k, h|a = 0, w = 1)$, or simply whether a label does not match the true state of the world, $P(k, h|a \neq w)$. It should be noted that the free-association database records responses given to text cues and not visual cues, which were used case in the study.

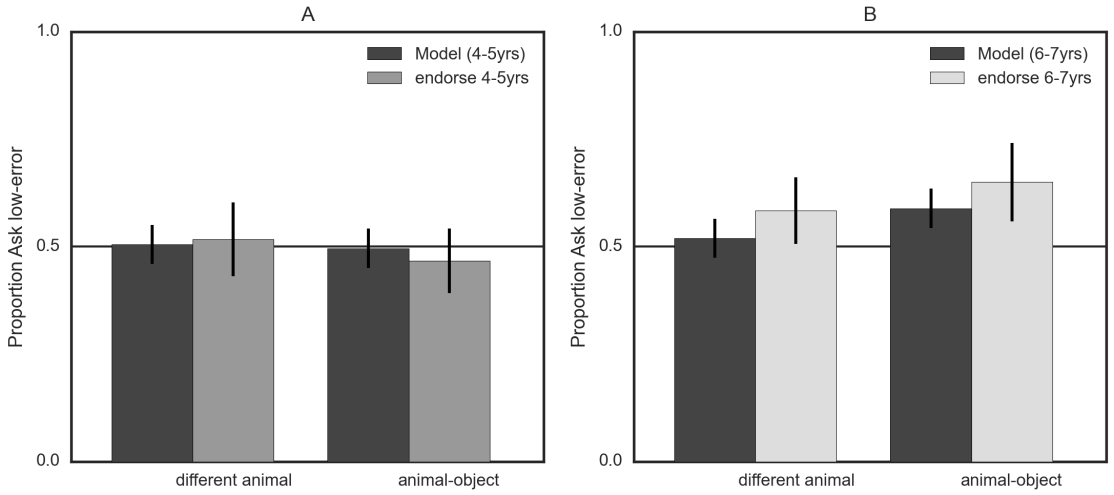


Figure 9. Model simulation results for Einav and Robinson (2010). a) Four- and five-year-olds. b) Six- and Seven-year-olds. The x-axis displays the accuracy condition. The y-axis shows the proportion of children who endorsed the answer given by the lower-magnitude-error informant, or for the model, the probability of endorsing the lower-magnitude-error informant.

The experimental results (see Figure 9) indicate that four- and five-year-olds do not exhibit a preference for either informant, but six- and seven-year-olds prefer informants who produce lower-magnitude errors. Higher magnitude errors are a better indication of naivety or unhelpfulness than lower magnitude errors. Unknowledgeable, helpful informants should guess a label close to the target and then produce a label that is close to the guessed label.

Looking time. The epistemic trust model is easily adapted to account for looking time paradigms. The primary hurdle is the mapping from probability to looking time. We assume that the time spent looking at an event is inversely proportional to the probability of that event. We are aware of recent work that suggest looking time follows a U-shaped function whereby infants look longer at moderately improbable events and less at extremely probable or improbable events (Kidd, Piantadosi, & Aslin, 2012). Recent work has successfully modeled this phenomenon (Piantadosi, Kidd, & Aslin, 2014), but adopting this model requires more than doubling the number of free parameters in our model, which we believe adds unjustifiable complexity.

We model Koenig and Echols (2003, Study 1) in which 18-month-olds observe novel informants label common objects, displayed on a screen, either correctly (*true* labels condition) or incorrectly (*false* labels condition) for twelve trials. At each trial the number of seconds infants looked at the informant, the object, and their parents (on whose lap they sat) was recorded. We model only the time spent looking at the informant because the model most fluidly produces the probability of an informant producing a specific label given a specific target. Koenig and Echols (2003) report

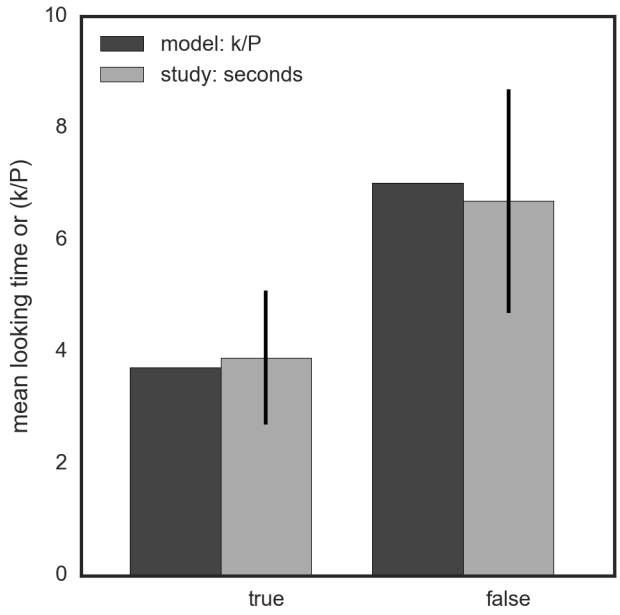


Figure 10. Model simulation results for Koenig and Echols (2003). On the Y axis is the mean time in seconds infants spent looking at the informant across trials and for the model, the mean inverse probability of the informants actions across trials.

the mean looking time over trials. We report the mean inverse probability scaled arbitrarily. It is important to note that the parameter fit for this particular experiment was achieved by minimizing the error of the *proportion difference* between the time spent looking at each informant in both the accurate and inaccurate conditions. For example, if infants in the true labels condition looked at the informant for an average of 4 seconds and infants in the false labels condition looked at the informant for an average of 7.5 seconds, the proportion difference is $7.5/4 = 1.875$. If the mean inverse probabilities for the true and false labels conditions are 1.2 and 3.8, respectively, then the relative error is $|1 - (3.8/1.2)/(7.5/4)| = 0.69$. We use this method because we are interested only in the trend from one condition to the other; we make no attempt to find the scaling constant that maps inverse probability to seconds. In this way, we can capture the trend without adding complexity.

Apart from the looking-time modifications, the rest of the workings are identical to those we used to model Pasquini et al. (2007). The results can be seen in Figure 10. We plot seconds beside inverse probability arbitrarily scaled. The model captures that an informant labeling common objects correctly is less surprising than an informant labeling common objects incorrectly.

Gaze following. Tummeltshammer et al. (2014, Experiment 1) investigated 8-month-olds' learning from informants using a gaze-following paradigm. The researchers employed eye-tracking technology to record infants' eye movements in response to gazes made by *reliable* and *unreliable* faces. For each face type, infants participated in four blocks of four familiarization trials. In each trial, a woman's head appeared in the center of a black screen. In each of the four corners of the screen were empty boxes (squares). At the beginning of each trial the head looked at the infant, said "Wow, look!", and turned to look at one of the four corners, at which time an animal noise sounded and its respective animal appeared in one of the boxes. Reliable faces always preemptively looked at the box in which the animal appeared and unreliable faces preemptively looked at the box in which the animal appeared only 25% of the time. Each square had a distinct animal and the heads only ever looked at two of the four boxes, that is, there were two boxes in which an animal

never appear and which were never looked at. After familiarization trials, infants participated in two different kinds of target trials: *test* and *generalization*. On *test* trials, the head looked at a box it had previously look at. After a short delay an animal sound played but no animal appeared, instead the corner boxes flashed. The same procedure repeated for *generalization* trails but the head looked at one of the boxes it had never looked at before—the hypothesis, in both cases, being that if such young infants are sensitive to informant reliability, infants who observed the reliable head should be more likely to follow its gaze. In both target trial types, infants looked at the box indicated by the reliable informant far more than the others boxes. Infants looked at the box indicated by the unreliable informant at chance.

From a modeling standpoint this study was difficult to capture, not because there is something about it that is inherently difficult to capture, but because the information supplied in the publication does not provide sufficient information to account for all the relevant details.[8] Before the experiment began, infants participated in a number of calibration trials during which objects appear in the corners and center of the screen. It is possible that these trials affected infants' beliefs about where objects should appear on the screen and hence their learning during familiarization. As an illustration: assume that during calibration infants cumulatively observe ten objects appear in each of the four corners. We capture the likelihood of an object appearing in a given corner with multinomial distribution with Jeffery's prior,

$$P(\text{corner}) \sim \text{Dirichlet}\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right), \tag{18}$$

which is the probabilistic way of establishing a loose, uniform belief that objects are equally likely to appear in any of the four squares. After calibration and posterior probability updates we have

$$P(\text{corner}) \sim \text{Dirichlet}\left(\frac{1}{2} + 10, \frac{1}{2} + 10, \frac{1}{2} + 10, \frac{1}{2} + 10\right), \tag{19}$$

which amounts to a very rigid uniform belief and which slows future updating—that is to say that each subsequent observation less affects the predictive probability of a specific event. Assuming that infants update their beliefs about objects and corners on each trial, an infant who receives the above calibration trials will attribute a predictive probability of 0.362 to an object appearing in one of the two never-before-indicated boxes on generalization trials where an infant with no calibration trials would attribute a probability of only 0.056 to the same event. We ignored this sort of posterior updating because the study provides insufficient data and, as we have demonstrated, subtle differences in calibration assumptions can lead to dramatically different results. We assume that infants held a uniform probability over objects to corners for the duration of the experiment. It should be noted that there was a qualitative difference in infants' behavior in the two target trials that could be explained by updating beliefs about objects and corners. It appears that infants followed the reliable head's gaze to the cued box more in generalization trials than they did in test trials and followed the unreliable head's gaze less in generalization trials than they did in test trials (see Figure 11). If infants are looking for the box with the animal and an unreliable informant looks toward a box in which an animal has never appeared, children should look less because at baseline it is unlikely for an animal to appear there. A reliable head's gaze, to an extent, overrides the low prior probability of an animal appearing in that corner.

Another issue is trial ordering. Just as beliefs about corners and objects propagate across trials, so too do beliefs about informants. The study was conducted using a between-subjects design. The order of the boxes in which the animals appears and—we assume—the order of the trials during which the unreliable informant looked at the correct object were counterbalanced. It is computationally intractable to average over many orderings for an experiment of so many trials, and because we do not have the exact trial orders of each participant, we cannot use approximation

[8] We requested, but were not able to obtain data from the authors

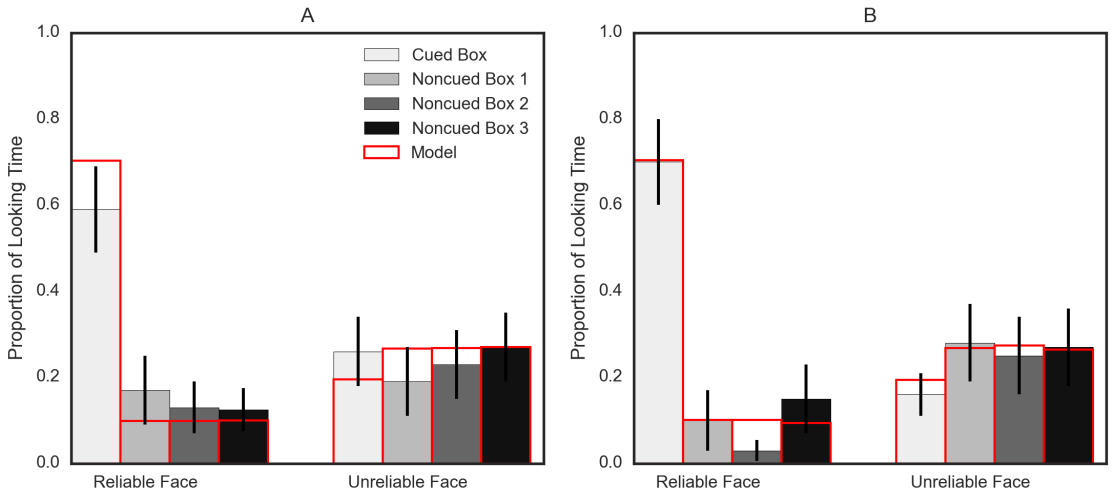


Figure 11. Model simulation results for Tummeltshammer, Wu, Sobel, and Kirkham (2014). Error bars represent standard error.

methods to capture individuals' behaviors (e.g. win-stay, lose-shift Bonawitz, Denison, Gopnik, & Griffiths, 2014). We modeled each condition—reliable and unreliable—separately and chose a single order for the unreliable condition (that the face looked at the correct box on the second trial of each block).

Infants' likelihood of looking at the box indicated by the face was modeled using the same process as modeling an endorse trial. The infant should expect an animal to appear in the box indicated by the face if the face is likely to correctly label (via its gaze) that box as "the box that is going to have the animal in it". In Figure 11 we report the model results[9].

We see that the model captures infants' preference to follow the reliable face's gaze and to look other than where the unreliable face gazes. Again, there is a qualitative (though not statistically significant) difference in the results for the test and generalization trials for unreliable faces. Infants seems to look uniformly in the test trials (Figure 11a) and seem to look other than where the unreliable face looks in generalization trials (Figure 11b). Because we have ignored posterior updating with respect to object locations, these two target trials are indistinguishable to the model.

Results

An especially novel aspect of this work is in integrating results across experiments. We proceed by conducting an analysis using CrossCat. Once CrossCat has inferred a joint probability distribution over the data table, the bulk of the work is done; we need only ask CrossCat what it has learned. We refer readers who are more familiar with significance testing and who may wonder why we chose not to use analogous significance test to Appendix C for a concrete example.

The first question relates to dependence among the variables. Previous research has debated what explains changes in children's behavior with age: changes in reasoning about knowledgeability, or changes in reasoning about helpfulness? Figure 12 (left) presents a *dependence probability matrix* where each row and column entry, (i, j), represents the probability that variables i and j share a dependence (for details on calculating dependence probability and conditional distributions under CrossCat, see Appendix B or [Mansinghka et al., Accepted pending revision]). Pairs of variables for which changes in one tend to be associated with changes in the other are said to be dependent.

[9]Tummeltshammer et al. (2014) did not report their means and did not provide them on request so we used the *ruler-to-bar-chart method* to approximately measure them.

As a reference point, the expected dependence probability (before effects of the data), derived from the CRP with parameter α where $\alpha \sim \text{Exp}(1)$ is roughly 0.596 (for the full derivation of this quantity see Appendix section B). The dependence probability matrix is used as a way to explore which variables have interesting relationships. The higher the dependence probability between two variables, the more likely it is that the variables are mutually predictive. Because CrossCat learns a joint distribution over the entire dataset, we can try to predict any variable using any other variable but if the dependence probability between those variables is low, the two variables may not hold much information about each other; and if the dependence probability between two variables is zero, they have zero mutual information. The dependence probability matrix gives us a way to quickly determine which variables are likely to have interesting relationships that warrant more in-depth exploration.

In our model, the dependence probability between columns is generally high. The lower right-hand area of the matrix shows that the strength parameter for helpfulness and age are highly dependent and that both helpfulness parameters are highly dependent with communication mode. In contrast, both knowledgeability parameters show minimal evidence for dependence with age. Thus, the model indicates that age-related changes in behavior on epistemic trust tasks are related to changes in children's reasoning about helpfulness.

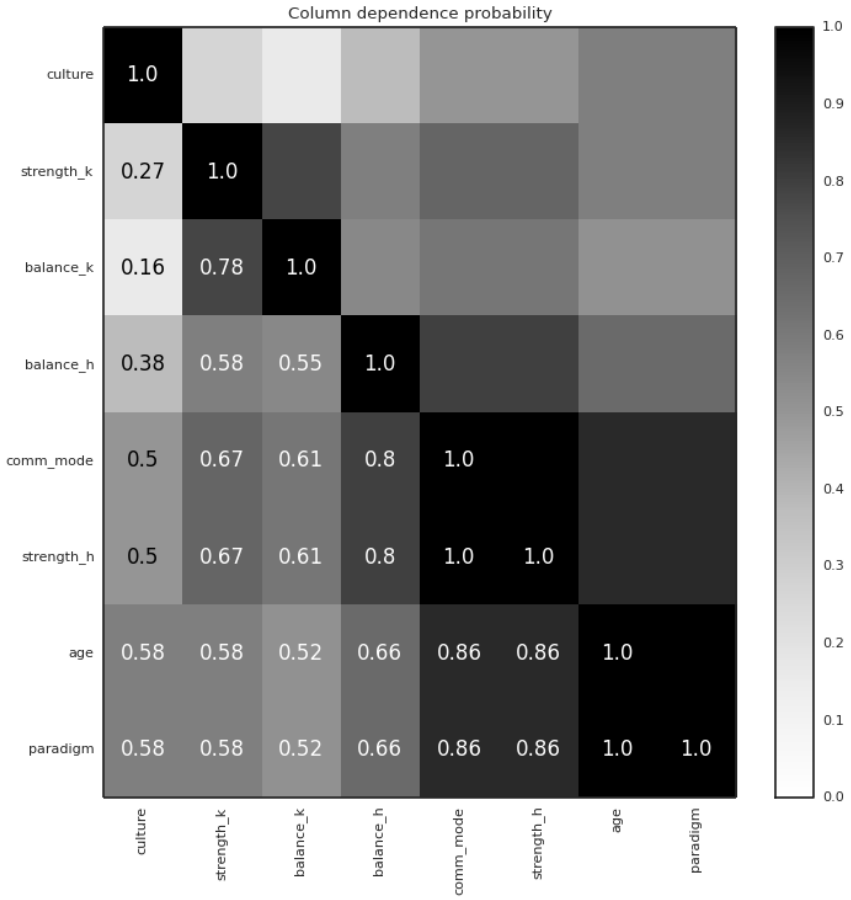


Figure 12. The dependence probability matrix resulting from cross-categorization. Each cell, $[i, j]$ of the table represents the probability of dependence between columns i and j. Probability is represented by shade. The lighter the shade, the lower the probability of dependence. Numerical dependence probabilities values are displayed in their respective cells.

Because the dependence probability matrix suggests a dependency exists between the helpfulness variables and age, we may investigate the form of these dependencies. How does children's reasoning about helpfulness change with age? We can form predictions about one variable based on different values of a second variable. To investigate the relationship between age and helpfulness we compute the distributions for the strength and balance parameters on helpfulness given a set of age groups, i.e. $P(s_h|age = \{1.5, 3.5, 4.5, 5.5\})$ and $P(b_h|age = \{1.5, 3.5, 4.5, 5.5\})$. The resulting distributions are multimodal, so we display the full distributions rather than report standard summary statistics, which are largely useless in this case. For example, the mean and variance of the data are *sufficient* to summarize normally-distributed data because a single normal distribution is parametrized in terms of a mean and a variance, but they are not sufficient to describe data from a mixture of many normal distributions.

Figure 13 (a,b) shows the results for balance and strength, respectively. The mass of balance for helpfulness (see Figure 13a) for 18-month-olds rests heavily to toward 1 indicating that the model explains their behavior via an assumption that people are in general helpful. From 18 months through 5.5 years there is a shift through a more uniform (flat) distribution to a peak at a more neutral position. This suggests that the data are explained by an increasing belief that not everyone is helpful.

We see a similar trend in the strength of helpfulness. Younger ages have higher mean strengths, which, together with the balance parameter result, indicates more rigid beliefs that everyone is helpful. With age, the strength relaxes to a lower value. Lower strength indicates greater flexibility, indicating a non-rigid belief that people are either helpful or not. Thus, the model captures younger children's behavior by attributing higher, more rigid prior biases toward helpfulness.

We calculated similar distributions for knowledgeability parameters but saw no marked age differences (see Figure 13 c and d). The shapes of the distributions for each age group are essentially the same, suggesting no evidence for developmental changes in reasoning about knowledgeability.

The dependence probability matrix (Figure 12) showed that communication mode was dependent with the helpfulness parameters. Previous empirical research has observed differences in behavior based on different communication modes. For example, Couillard and Woodward (1999) found that children who received communication in the form of marker placement were less susceptible to informants' misinformation that those who were communicated to through finger points (Jaswal, Croft, Setia, and Cole [2010] found similar results exploring different communication modes). Querying the helpfulness parameter distributions given different communicative modes allows investigation of how the model captures differences across communication modes. Figure 14 shows the conditional distribution of helpfulness parameters based on each communication mode. The results show that the model explains behavior resulting from communication using markers differently than the others communication modes. Marker placement (in green) is captured with a bimodal distribution and further investigation reveals that the each mode corresponds to an age group (see Figure 14c). The high-balance mode corresponds to three-year-olds and the low-balance mode corresponds with four-year-olds. The other communications modes induce more unimodal distributions. This is broadly consistent with the idea that labeling, pointing, and gaze are ostensive cues that may be strongly associated with helpful communication (Gergely et al., 2007; Topal, Gergely, Miklosi, Erdohegyi, & Csibra, 2008). However, given that this result is based on a single study (Couillard & Woodward, 1999), some caution is warranted in this interpretation of the differences in epistemic trust using ostensive and non-ostensive cues.

Discussion

The model predicts that development is driven by changes in children's understanding of helpfulness in part because we have modeled studies that explicitly demonstrate the development of the understanding of helpfulness (e.g. Couillard & Woodward, 1999). Couillard and Woodward (1999) provided children with demonstrations of an informant behaving inconsistently with her knowledge, which is only possible in the epistemic trust model if helpfulness is represented. An

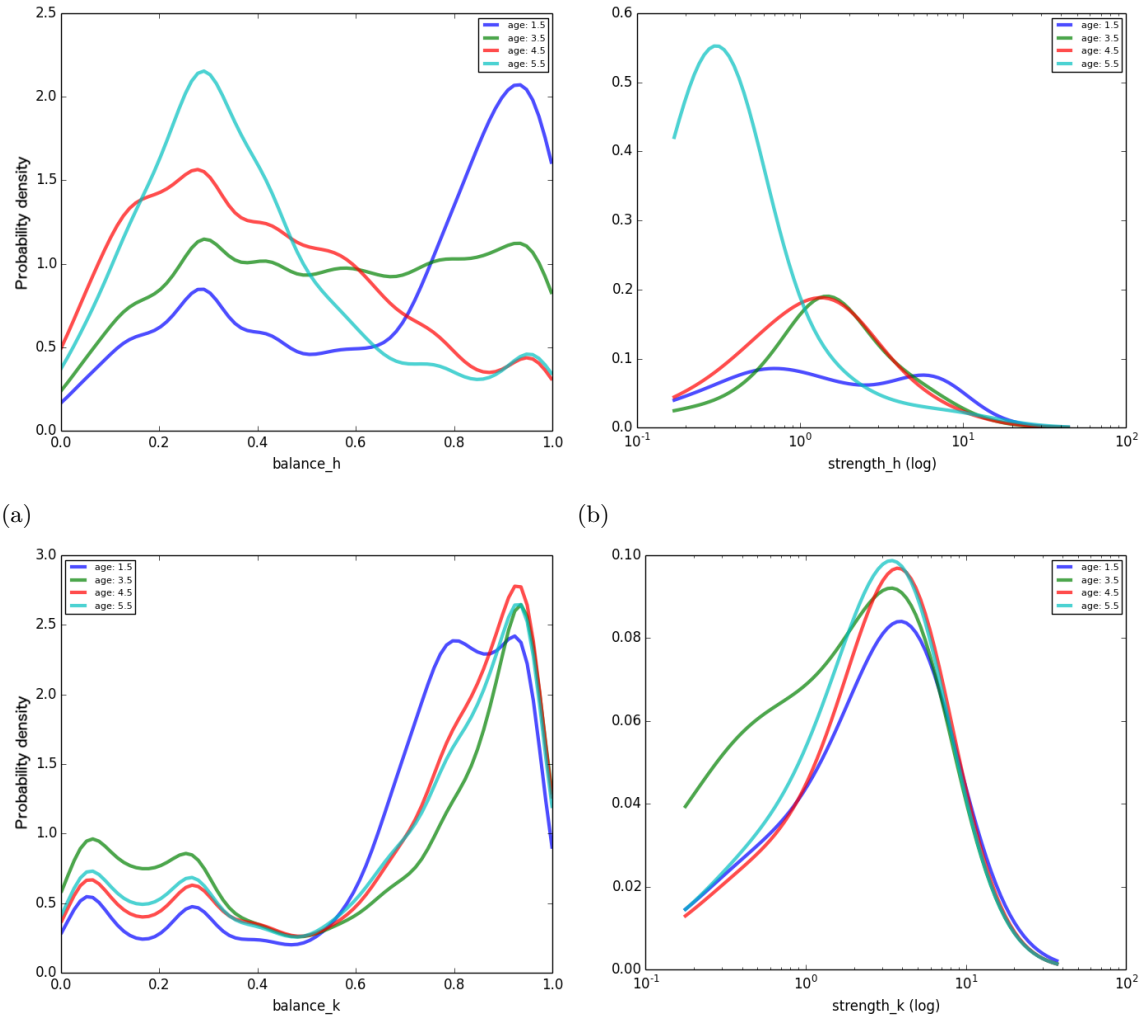
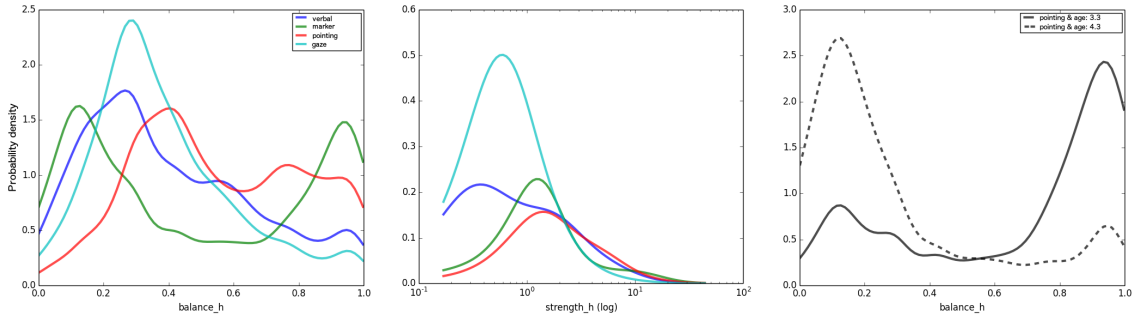


Figure 13. The conditional probability distributions of helpfulness and knowledgeability parameters given age. The distributions for ages 1.5 (blue), 3.5 (green), 4.5 (red), and 5.5 (teal). (a) Averaged conditional probability distribution of helpfulness's balance parameter. (b) Averaged conditional probability distribution of helpfulness's strength parameter. (c) Averaged conditional probability distribution of knowledgeability's balance parameter. (d) Averaged conditional probability distribution of knowledgeability's strength parameter.

informant who knows that the sticker is under cup A, but indicates cup B must not be helpful in conveying her knowledge. The more flexibly children represent helpfulness, the quicker they can learn to choose the opposite cup.

Younger children's slower updating in response to inaccurate labels may also be attributed to a lack of understanding of variable helpfulness. Older children update their trust more quickly than younger children. A four-year-old who observes an accurate label from informant A but an inaccurate label from informant B, is more likely to prefer informant A than a a three-year-old (cf. Koenig & Harris, 2005; Pasquini et al., 2007; Fitneva & Dunfield, 2010). In the epistemic trust model, helpfulness is a more predictive informant attribute than knowledgeability. This means that



(a) (b) (c)

Figure 14. The marginal conditional distribution of helpfulness's balance distribution (a) and strength distribution (b) given each communication mode. In blue: verbal, in green: marker placement, in red: pointing; in teal: gaze. (c) The distributions of helpfulness's balance parameter given that the informant communicated via marking and the informant for 3.3-year-olds (solid line) and 4.3 year-olds (dashed line).

knowing only about informants' helpfulness provides more information about the veracity of their testimony than knowing only about their knowledgeability. Given two informants with unknown knowledgeability, an known unhelpful informant will produce correct labels less often than a known helpful informant. Assuming that there are n possible labels for an object and that the probability of guessing the correct label is $1/n$, the helpful informant will produce the correct label $n+1$ times more often than the unhelpful informant. Under the same assumptions, but not knowing the informants' helpfulness, the known knowledgeable informant will produce the correct label only $n/2$ times more often than the known unknowledgeable informant. Thus knowing an informant's helpfulness reduces one's surprise at the outcome of a label more so than knowing an informant knowledgeability. [10]

Any predictions made by the model will reflect these properties. The model indicates that younger children represent helpfulness, but are highly biased to believe that all informants are helpful. This implies that children can learn that informants can act in ways inconsistent with their model (relax their biases); thus the more a child observes informants acting unhelpfully, the better that child should perform on helpfulness-oriented tasks. This leads to the prediction that a child with more experience with unhelpful informants should perform better on epistemic trust tasks. For example, younger children who attend preschool or daycare, or have older siblings should perform similarly to older children who spend more of their time around only their caregivers. This suggests that researchers should collect more demographic information and conduct analyses grouped by experiential variables rather than age.

General Discussion

Research in cognitive development routinely emphasizes the importance of other people in learning about the world. While a considerable amount of research has investigated the bases on which children decide epistemic trust, precise theories of the basic phenomenon and how it develops have been limited. Researchers have interpreted their results in terms of updating beliefs about informants' knowledge (Pasquini et al., 2007; Corriveau, Fusaro, & Harris, 2009; Corriveau & Harris, 2009), theorists have discussed whether epistemic trust is rational (Sobel & Kushnir, 2013), and philosophers have formalized accounts based on reasoning about informants' knowledgeability only (Bovens & Hartmann, 2004).

[10] As defined in terms of conditional entropy. The conditional entropy between two random variables X and Y, $H(X \mid Y)$ is the amount of information needed to describe X if Y is known, $H(X \mid Y) = \sum_{i,j} p(x_i \mid y_i)p(y_i)\log p(x_i \mid y_i)$. $H(X \mid Y_1) < H(X \mid Y_2)$ implies that knowing Y_1 tells us more about X than does knowing Y_2.

More recently, computational (Shafto, Goodman, & Frank, 2012; Butterfield et al., 2008), theoretical (Sperber et al., 2010), and empirical accounts (Mascaro & Sperber, 2009; Heyman & Legare, 2013; Koenig & E, 2014) have proposed that a complete theory of epistemic trust requires reasoning about both informants' knowledgeability *and* intent. Shafto, Eaves, et al. (2012) proposed a computational model and applied it to three studies from the literature, finding that an account based on knowledge and intent best explained four-year-olds' behavior. They also found that there were developmental changes in reasoning, and that these changes were in reasoning about intent rather than knowledgeability. However, the import of this evidence is limited by the need to limit consideration to only three studies, which ensured uniformity in methods, ages, etc. necessary for the model fitting.

We have proposed a computational framework for integrating results from heterogeneous studies and used it to model the development of epistemic trust. The framework is based on parameterizing results in model space and analyzing the parametrized results alongside demographic features of the studies, allowing heterogeneous studies to be included and the heterogeneity to be analyzed without requiring arbitrary assumptions from the analyst as to how to partition the data. Our results confirm and quantify previous arguments claiming that reasoning about both knowledgeability and intent play a role in epistemic trust and developmental differences are attributable to changes in reasoning about informants' intent. Reasoning about informant's knowledgeability is found to be relatively constant. Our results extend previous findings, but in a much broader age range—18 months to 5.5 years—and quantify gradual change in reasoning about informants' intent across that time period. Our results also extend previous findings by explaining why different modes of communicating used in experiments lead to different results. Consistent with previous theoretical accounts (Csibra & Gergely, 2006) and empirical observations (Couillard & Woodward, 1999), different modes of communication induce different expectations about how the data are selected.

Our approach represents a proposed solution to a vexing problem in cognitive development: developing coherent theoretical accounts that explain changes in behavior over time despite the confounded relationship between age and methodology. Standard practice in cognitive development circumvents this problem by focusing on identifying the youngest age at which children can succeed on a conceptual problem. This avoids the problem of covariance between age and task by prioritizing methods that apply at the youngest ages. However, this approach limits the relevance of resulting theory by prioritizing questions of competence over questions of performance.

Instead of focusing only on tasks that demonstrate competence at the youngest ages, we used the computational theory to parametrize the complete set of results that are explainable with that theory. We then used computational tools to make explicit the relationship between the model's parametrization and the demographics of the experiment. This approach formalizes developmental theorizing in a way that supports inferences about the youngest ages that children may succeed on a task, as well as relationships across behavior on different tasks, at different ages.

While this approach provides a more comprehensive, computationally precise account of the development of epistemic trust, there are limitations. Most notably, we have considered 11 studies from the literature. Although the epistemic trust literature is in principle, much larger, including more studies would have required additional assumptions and/or free parameters. The evidence is too sparse to constrain these choices. Currently, much of the focus of epistemic trust research is on documenting new paradigms that cause children to allocate trust differently. The method we have outlined will be most informative given more systematic analyses of phenomenon, in which studies are more mutually informative, e.g. paradigms that are slight adjustments of other paradigms or that investigate interactions between paradigms. Furthermore, empirical research focusing on quantitative, as well as qualitative, results would provide richer data for testing computational theories of epistemic trust on cognitive development.

Theoretical and empirical accounts of cognitive development emphasize the important role of other people in children's learning about the world. We have proposed a computational theory and an approach for integrating results across heterogeneous methods and ages. The results indicate

developmental changes in reasoning about informant's intent and differences across tasks. Although we believe our approach to be the most precise and comprehensive account of the development of epistemic trust, there are many ways in which it is likely too simple to explain the richness of development. Continued empirical research is necessary toward the goal of developing a complete computational theory of the development of epistemic trust. Developmentalists are vital to this effort and can contribute in two ways. First, by filling gaps in the literature by reproducing existing results in different age and culture groups, and by extending existing paradigms to account for more nuanced phenomenon (much in the same way Pasquini et al. [2007] did for Koenig and Harris [2005]). And second, by experimentally evaluating the model assumptions.

Acknowledgments

This research was supported in part by NSF CAREER award DRL-1149116 and the DARPA XDATA program to P.S.

References

Aguiar, N. R., Stoess, C. J., & Taylor, M. (2012). The development of children's ability to fill the gaps in their knowledge by consulting experts. *Child development*, *83*(4), 1368–81.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bascandziev, I. & Harris, P. L. (2014, March). In beauty we trust: Children prefer information from more attractive informants. *The British journal of developmental psychology*, *32*(1), 94–9.

Baum, L., Danovitch, J. H., & Keil, F. C. (2008). Children's sensitivity to circular explanations. *Journal of experimental child psychology*, *100*(2), 146–155.

Bergstra, J. & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, *13*(1), 281–305.

Birch, S. A., Akmal, N., & Frampton, K. L. (2010). Two-year-olds are vigilant of others' non-verbal cues to credibility. *Developmental science*, *13*(2), 363–369.

Birch, S. A. & Bloom, P. (2002). Preschoolers are sensitive to the speaker's knowledge when learning proper names. *Child development*, *73*(2), 434–44.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014, July). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive psychology*, *74C*, 35–65.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011, September). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–30.

Boseovski, J. J. & Lee, K. (2006, May). Children's Use of Frequency Information for Trait Categorization and Behavioral Prediction. *Developmental psychology*, *42*(3), 500–13.

Boseovski, J. J. & Thurman, S. (2013, March). Evaluating and Approaching a Strange Animal: Children's Trust in Informant Testimony. *Child development*, *85*(2), 824–34.

Bovens, L. & Hartmann, S. (2004). Bayesian epistemology. *OUP Catalogue*.

Buchsbaum, D., Bridgers, S., Whalen, A., Griffiths, T. L., & Gopnik, A. (2012). Do I know that you know what you know? Modeling testimony in causal inference. In *Proceedings of the 35th annual meeting of the cognitive science society*.

Butterfield, J., Jenkins, O. C., Sobel, D. M., & Schwertfeger, J. (2008, November). Modeling Aspects of Theory of Mind with Markov Random Fields. *International Journal of Social Robotics*, *1*(1), 41–51.

Chen, E. E., Corriveau, K. H., & Harris, P. L. (2012). Children trust a consensus composed of outgroup members–but do not retain that trust. *Child development*, *84*(1), 269–82.

Collins, A. & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6).

Corriveau, K. H., Fusaro, M., & Harris, P. L. (2009, March). Going with the flow: preschoolers prefer nondissenters as informants. *Psychological science*, *20*(3), 372–7.

Corriveau, K. H. & Harris, P. L. (2009, April). Choosing your informant: weighing familiarity and recent accuracy. *Developmental science*, *12*(3), 426–37.

Corriveau, K. H. & Harris, P. L. (2010, March). Preschoolers (sometimes) defer to the majority in making simple perceptual judgments. *Developmental psychology*, *46*(2), 437–45.

Corriveau, K. H., Meints, K., & Harris, P. L. (2009, June). Early tracking of informant accuracy and inaccuracy. *British Journal of Developmental Psychology*, *27*(2), 331–342.

Couillard, N. & Woodward, A. (1999). Children's comprehension of deceptive points. *British Journal of Developmental Psychology*, *17*(4), 515–521.

Csibra, G. & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In *Processes of change in brain and cognitive development. attention and performance xxi* (pp. 249–274).

Csibra, G. & Gergely, G. (2009, April). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–53.

Danovitch, J. H. & Keil, F. C. (2004). Should You Ask a Fisherman or a Biologist?: Developmental Shifts in Ways of Clustering Knowledge. *Child development*, *75*(3), 918–31.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

DiYanni, C., Corriveau, K. H., Nasrini, J., Kurkul, K., & Nini, D. (2015). The role of consensus and culture in children's imitation of inefficient actions. *Journal of experimental child psychology*, *137*, 99–110.

DiYanni, C. & Kelemen, D. (2008). Using a bad tool with good intention: young childre's imitation of adults' questionable choices. *Journal of experimental child psychology*, *101*(4), 241–261.

Eaves, B. S. & Shafto, P. (2012). Unifying pedagogical reasoning and epistemic trust. *Advances in child development and behavior*, *43*, 295–319.

Einav, S. & Robinson, E. J. (2010, July). Children's sensitivity to error magnitude when evaluating informants. *Cognitive Development*, *25*(3), 218–232.

Eskritt, M., Whalen, J., & Lee, K. (2008). Preschoolers can recognize violations of the Gricean maxims. *British Journal of Developmental ...* *26*(3), 435–443.

Fink, D. (1997). A Compendium of Conjugate Priors. (1994), 1–47.

Fitneva, S. & Dunfield, K. (2010, September). Selective information seeking after a single encounter. *Developmental psychology*, *46*(5), 1380–4.

Frank, M. C. [Michael. C.] & Goodman, N. D. [Noah D.]. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998.

Freire, A., Eskritt, M., & Lee, K. (2004). Are Eyes Windows to a Deceiver's Soul? Children's Use of Another's Eye Gaze Cues in a Deceptive Situation. *Developmental psychology*, *40*(6), 1093–1104.

Fusaro, M. & Harris, P. L. (2008, September). Children assess informant reliability using bystanders' non-verbal cues. *Developmental science*, *11*(5), 771–7.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6), 721–741.

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

Gergely, G., Egyed, K., & Király, I. (2007, January). On pedagogy. *Developmental science*, *10*(1), 139–46.

Grice, H. P., Cole, P., & Morgan, J. L. (1975). Syntax and semantics. *Logic and conversation*, *3*, 41–58.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007, December). Google and the mind: predicting fluency with PageRank. *Psychological science*, *18*(12), 1069–76.

Gweon, H., Shafto, P., & Schulz, L. E. (2014). Children consider prior knowledge and the cost of information both in learning from and teaching others Hyowon Gweon. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 565–570).

Hagberg, A., Swart, P., & Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th python in science conference (scipy2008)* (SciPy, pp. 11–15). Pasadena, CA USA.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences*, *33*(2-3), 61–83.

Heyman, G. D. & Legare, C. H. (2013). Social Cognitive Development: Learning from Others. In D. E. Carlston (Ed.), *The oxford handbook of social cognition* (pp. 749–766). New York, NY: Oxford University Press.

Jaswal, V. K., Croft, A., Setia, A., & Cole, C. (2010). Young children have a specific, highly robust bias to trust testimony. *Psychological Science*, *21*(10), 1541–1547.

Jaswal, V. K. & Neely, L. a. (2006, September). Adults don't always know best: preschoolers use past reliability over age when learning new words. *Psychological science*, *17*(9), 757–8.

Keil, F. C., Stein, C., Webb, L., Billings, V. D., & Rozenblit, L. (2008, March). Discerning the Division of Cognitive Labor: An Emerging Understanding of How Knowledge Is Clustered in Other Minds. *Cognitive science*, *32*(2), 259–300.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007, May). Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, *10*(3), 307–21.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012, January). The Goldilocks effect: human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, *7*(5), e36399.

Kim, S. & Harris, P. L. (2014, April). Children prefer to learn from mind-readers. *The British journal of developmental psychology*, 1–13.

Kinzler, K. D., Corriveau, K. H., & Harris, P. L. (2011, January). Children's selective trust in native-accented speakers. *Developmental science*, *14*(1), 106–11.

Koenig, M. & E, S. (2014). Characterizing children's responsiveness to cues of speaker trustworthiness: two proposals. In E. J. Robinson & S. Einav (Eds.), *Trust and skepticism: children's selective learning from testimony*. Cambridge, UK: Psychology Press.

Koenig, M. & Echols, C. H. (2003). Infants' understanding of false labeling events: the referential roles of words and the speakers who use them. *Cognition*, *87*, 179–208.

Koenig, M. & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child development*, *76*(6), 1261–77.

Koenig, M. & Jaswal, V. K. (2011). Characterizing children's expectations about expertise and incompetence: halo or pitchfork effects? *Child development*, *82*(5), 1634–47.

Krogh-Jespersen, S. & Echols, C. H. (2012, March). The influence of speaker reliability on first versus second label learning. *Child development*, *83*(2), 581–90.

Kushnir, T., Wellman, H. M., & Gelman, S. A. (2008, June). The role of preschoolers' social understanding in evaluating the informativeness of causal interventions. *Cognition*, *107*(3), 1084–92.

Landrum, A. R. [Asheley R.], Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: a theoretical framework. *Trends in Cognitive Sciences*, *19*(3), 109–111.

Landrum, A. R. [Asheley R], Mills, C. M., & Johnston, A. M. (2013, July). When do children trust the expert? Benevolence information influences children's trust more than expertise. *Developmental science*, *16*(4), 622–38.

Lane, J. D., Wellman, H. M., & Gelman, S. a. (2013). Informants' traits weigh heavily in young children's trust in testimony and in their epistemic inferences. *Child development*, *84*(4), 1253–68.

Lutz, D. J. & Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child development*, *73*(4), 1073–84.

MacEachern, S. N. & Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of computational and graphical statistics*, *7*(2).

Mansinghka, V., Shafto, P., Jonas, E., Petschulat, C., Gasner, M., & Tenenbaum, J. B. (Accepted pending revision). Crosscat: a fully bayesian nonparametric method for analyzing heterogeneous, high dimensional data. *Journal of Machine Learning Research*.

Mascaro, O. & Sperber, D. (2009, September). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*(3), 367–80.

McDonald, K. P. & Ma, L. (2015). Dress nicer= know more? young children's knowledge attribution and selective learning based on how others dress. *PloS one*, *10*(12), e0144424.

Mills, C. M. (2013, March). Knowing when to doubt: developing a critical stance when learning from others. *Developmental psychology*, *49*(3), 404–18.

Mills, C. M. & Keil, F. C. (2008). Children's developing notions of (im) partiality. *Cognition*, *107*(2), 528–551.

Mills, C. M. & Landrum, A. R. [Asheley R]. (2012). Judging judges: how do children weigh the importance of capability and objectivity for being a good decision maker? *British Journal of Developmental Psychology*, *30*(3), 393–414.

Murphy, K. P. (2007). Conjugate bayesian analysis of the gaussian distribution. *Technical Report*.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, *9*(2), 249–265.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Nurmsoo, E. & Robinson, E. J. (2009, January). Identifying unreliable informants: do children excuse past inaccuracy? *Developmental science*, *12*(1), 41–7.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web., 1–17.

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007, September). Preschoolers monitor the relative accuracy of informants. *Developmental psychology*, *43*(5), 1216–26.

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge Univ Press.

Piantadosi, S. T., Kidd, C., & Aslin, R. (2014, February). Rich analysis and rational models: inferring individual behavior from infant looking data. *Developmental Science*.

Poulin-Dubois, D. & Chow, V. (2009, December). The Effect of a Looker's Past Reliability on Infants' Reasoning About Beliefs. *Developmental psychology*, *45*(6), 1576–82.

Rasmussen, C. (2000). The infinite Gaussian mixture model. *Advances in neural information processing*, (11), 554–560.

Robinson, E. J. & Whitcombe, E. (2003, February). Children's suggestibility in relation to their understanding about sources of knowledge. *32*(1), 54–55.

Sabbagh, M. A. & Baldwin, D. (2001). Learning words from knowledgeable versus ignorant speakers: links between preschoolers' theory of mind and semantic development. *Child development*, *72*(4), 1054–70.

Sabbagh, M. A., Wdowiak, S. D., & Ottaway, J. M. (2003, November). Do word learners ignore ignorant speakers? *Journal of Child Language*, *30*(4), 905–924.

Shafto, P., Eaves, B. S., Navarro, D. J., & Perfors, A. (2012, May). Epistemic trust: modeling children's reasoning about others' knowledge and intent. *Developmental science*, *15*(3), 436–47.

Shafto, P. & Goodman, N. D. [Noah D]. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the thirtieth annual conference of the cognitive science society*.

Shafto, P., Goodman, N. D. [Noah D], & Frank, M. C. [Michael C]. (2012, June). Learning From Others: The Consequences of Psychological Reasoning for Human Learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014, March). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71C*, 55–89.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011, July). A probabilistic model of cross-categorization. *Cognition*, *120*(1), 1–25.

Sloman, S., Love, B., & Ahn, W. (1998, April). Feature centrality and conceptual coherence. *Cognitive Science*, *22*(2), 189–228.

Sobel, D. M. & Corriveau, K. H. (2010). Children monitor individuals' expertise for word learning. *Child development*, *81*(2), 669–79.

Sobel, D. M. & Kushnir, T. (2013, October). Knowledge matters: how children evaluate the reliability of testimony as a process of rational inference. *Psychological review*, *120*(4), 779–97.

Sperber, D., Cléement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, O., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search*. Lecture Notes in Statistics. New York, NY: Springer New York.

Taylor, M. G. (2013, August). Gender influences on children's selective trust of adult testimony. *Journal of experimental child psychology*, *115*(4), 672–90.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006, December). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.

Topal, J., Gergely, G., Miklosi, A., Erdohegyi, A., & Csibra, G. (2008). Infants' perseverative search errors are induced by pragmatic misinterpretation. *Science*, *321*(September), 1831.

Tummeltshammer, K. S., Wu, R., Sobel, D. M., & Kirkham, N. Z. (2014, July). Infants Track the Reliability of Potential Informants. *Psychological science*, (July).

Appendix A
Monte Carlo estimation in the epistemic trust model

Gibbs sampling

To estmiate distributions in the model, we employ Gibbs sampling (see S. Geman & Geman, 1984; Gelman et al., 2013), a Marko chain Monte Carlo (MCMC) method well-suited to use in Bayesian networks. It works by re-sampling each node conditioned on the values of every other node in the network. However, it is most often the case that a node is *not* dependent on every other node, but only a few, thus the terms in which the target node does not appear cancel out. Further more, we can exploit conditional dependence to simplify further. A node is conditionally independent of all other nodes in a network given its *Markov blanket*: the nodes comprising its parents, children, and children's parents. The Markov blanket of the knowledgeability node can be seen shaded in gray in Figure A1. The conditional probabilities and distributions on each variable are thus:

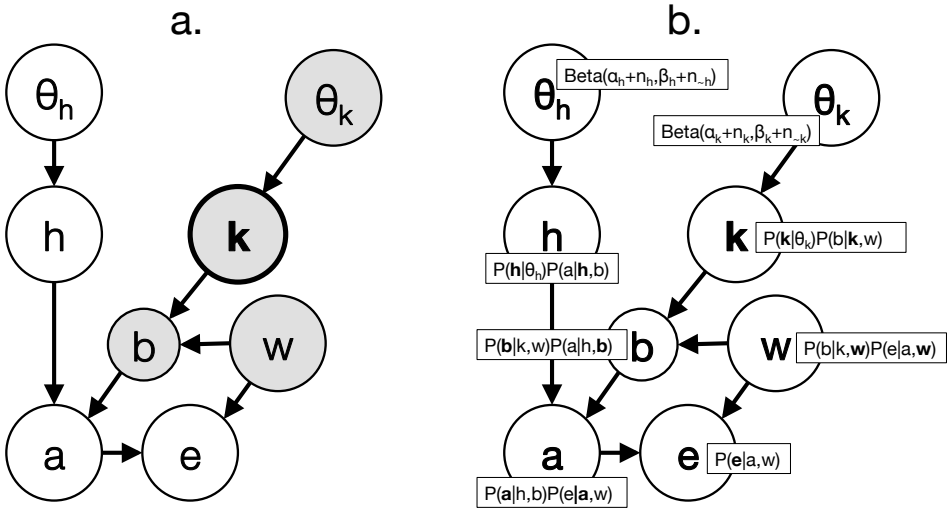


Figure A1. Example Markov blanket for epistemic trust model and Gibbs sampling conditional probabilities. a) Markov blanket (shaded in gray) for knowledgeability, k. b) Gibbs sampling conditional probabilities and distribution superimposed on their respective variables.

$$\begin{aligned}
\theta_k &\sim \text{beta}(\alpha_k + n_k, \beta_k + n_{\neg k}), && (20)\\
\theta_h &\sim \text{beta}(\alpha_h + n_h, \beta_h + n_{\neg h}), && (21)\\
w &\sim p(w)p(b|k,w)p(e|a,w), && (22)\\
k &\sim p(k|\theta_k)p(b|k,w), && (23)\\
h &\sim p(h|\theta_h)p(a|h,b), && (24)\\
b &\sim p(b|k,w)p(a|h,b), && (25)\\
a &\sim p(a|h,b)p(e|a,w), && (26)\\
e &\sim p(e|a,w), && (27)
\end{aligned}$$

where n_h and $n_{\neg h}$ are the number of trials in which the informant has be helpful and unhelpful, and where n_k and $n_{\neg k}$ are the number of trials in which the informant has be knowledgeable and unknowledgeable.

The sampler state is set to some random value, fixing observed nodes to their observed values. Then, for a predetermined number of iterations, the Gibbs sampler updates each unobserved node in random order. For example, if we observe an action and an effect, we set the a and e nodes and update all other nodes while keeping a and e static. We then collect or count as we did with rejection sampling subject to some caveats.

Samples generated by a Gibbs sampling algorithm are not independent. They depend on the previous state. To mitigate effects of sample interdependence we ignore a certain number of samples between each collection. This process is known as *lag* or *thinning*. For the same reason, we must throw out a large number of samples before collecting the first. The sampler state may have been initialized to a value that is not representative of the target distribution and it make take the sampler some time to walk its way to the target region. Another concern is Gibb samplers' propensity to get stuck in local maxima. Imagine a bimodal probability distribution with two distant peaks. In order for the sampler to cross the gap from peak to peak, it must cross a large space of low probability. It is common practice to average samples over multiple independent instances (*chains*) of Gibbs sampler runs to smooth the between–chain variability due to local maxima.

Appendix B
Cross-categorization details

A cross-categorization state consists of the following parts:

1. α_s: the CRP concentration parameter for the assignment of columns to views
2. $\alpha_v = \{\alpha_v^0, \alpha_v^1, \ldots, \alpha_v^{|V|}\}$: the CRP concentration parameter for each of the $|V|$ views' assignments of rows to categories.
3. $Z = \{z_0, z_1, \ldots, z_{F-1}\}$ where $z \in \{0, 1, \ldots, |V|\}$: the assignment of the F features (columns) to the $|V|$ views.
4. $V = \{V_0, V_1, \ldots, V_{|V|-1}\}$ where $V_i = \{v_i^0, v_i^1, \ldots, v_i^{N-1}\}$ and where $v_i \in \{0, 1, \ldots, |K_i| - 1\}$: the assignment of the the N rows in view i to the $|K_i|$ categories in view i.
5. $\mathbf{\Theta} = \{\theta_0, \theta_i, \ldots, \theta_{F-1}\}$ where θ_f^k is the data model for feature category (components) k of feature f: the data models for each feature. For example if feature f is modeled with a Normal distribution then $\theta_f^k = \{\mu_k, \rho_k\}$, the mean and precision of the category k.
6. $\Phi = \{\phi_0, \phi_1, \ldots, \phi_{C-1}\}$: the prior distributions for each colum. For example if column c is modeled via a normal distribution, ϕ_c may represent a Normal-Gamma prior, $\phi_c = \{m, r, s, \nu\}$.
7. $G_0 = \{G_0^0, G_0^1, \ldots, G_0^{F-1}\}$: the hyper prior distributions on each $\phi \in \Phi$.
8. H_S: the prior distribution on the CRP concentration parameter for the assignment of features to views.
9. H_V: the prior distribution on the CRP concentration parameter for the assignment of rows to components.

The score (un-normalized probability) of a cross-categorization state, S, is,

$$\text{score}(S) = P(\alpha_S|Z, H_S) \prod_{i=0}^{|V|-1} \left(P(V_i|\alpha_v^i) P(\alpha_v^i|H_V) \prod_{k=0}^{|K_i|-1} \prod_{f:V_f=v} \int_{\theta_k} P(X_f^k|\theta) P(\theta_f^k|\phi_f) d\theta_f \right) \prod_{f \in F} P(\phi|G_0^f), \tag{28}$$

where X_f^k is the data in feature f assigned to component k.

In a single cross-categorization sample, the conditional probability of a value, x, in column i given a value, y, in column j is

$$P(x|y) = \begin{cases} \sum_{c \in C_v} \frac{n_c}{n+1+\alpha_v} P(x|X_c) P(y|Y_c) + \frac{\alpha_v}{n+1+\alpha_v} P(x) P(y) & \text{if } z_i = z_j \\ \sum_{c \in C_v} \frac{n_c}{n+1+\alpha_v} P(x|X_c) + \frac{\alpha_v}{n+1+\alpha_v} P(x) & \text{if } z_i \neq z_j, \end{cases} \tag{29}$$

where n is the number of objects in the table, C_v is the set of categories belonging to view v, α_v is the CRP concentration parameter for veiew v, n_c is the number of objects assigned to category c, and X_c and Y_c are the data in X and Y assigned to category c. Note that if $z_i \neq z_j$—columns i and j are not in the same view—then $P(x|y) = P(x)$ because columns i and j are independent. For conditional distributions over multiple models, we employ model-averaging. Conditional distributions are averaged over samples:

$$P(x|y) = \frac{1}{|S|} \sum_{s \in S} P_s(x|y), \tag{30}$$

where S is the set of samples, s is an individual sample, and $P_s(x|y)$ is the conditional probability of x given y under sample s.

Crosscat offers a measure or the dependence between pairs of columns by way of *dependence probability*. Given S samples, the dependence probability between i and j is defined as being proportional the the number of samples in which columns i and j belong to the same view and in which the view to which i and j belong has more than one cluster. Formally:

$$P(dep) \equiv \frac{|\{s \in S;\, z_i^s = z_j^s,\, K_{z_i}^s > 1\}|}{|S|}. \tag{31}$$

For cross-categorization, each feature must be assigned an appropriate probability distribution. All zero-bounded continuous features (s_k, s_h, and age), were assigned Lognormal likelihood functions with the standard conjugate Normal-Gamma prior; balance parameters were assigned Normal likelihood functions with Normal-Gamma prior; and the categorical variables (communication mode, paradigm, and culture) were assigned Multinomial likelihood functions with the conjugate symmetric Dirichlet prior. We used a custom python implementation of cross-categorization[11]. We collected 64 samples after 500 iterations of inference. That is, we initialized 64 independent Markov chains of the sampler, ran the sampler for 500 iterations, and conducted analyses using the 64 independent states.

Expected dependence probability between cross-categorization columns

First, we derive the probability, under the Chinese Restaurant Process (CRP), that two items will be assigned to the same component. Because the CRP is an exchangeable process, in the limit it may be described as i.i.d. This means that we need only be concerned with the probability that the *first two* items are assigned to the same component. The first item is always assigned to its own component, the second item is assigned the the same component with probability $\frac{1}{1+\alpha}$, where α is the CRP concentration parameter. Thus the probability that any two columns, i and j, belong to the same components is,

$$P(z_i = z_j|\alpha) = \frac{1}{1+\alpha}. \tag{32}$$

In our implementation of cross-categorization, α is given an exponential prior with mean 1. That is,

$$\alpha \sim \text{Exp}(1). \tag{33}$$

We must calculate the expected expected dependence probability across the prior. That is,

$$E[Y] = E\left[\frac{1}{1+\alpha}\right]. \tag{34}$$

We derive the cdf of this distribution:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) && (35)\\ &= P\left(\frac{1}{y} - 1 \leq \alpha\right) && (36)\\ &= 1 - F_X\left(\frac{1}{y} - 1\right). && (37)\\ & && (38) \end{aligned}$$

Differentiating leaves us with the pdf:

[11] Our implementation, *BaxCat*, can be found at https://github.com/BaxterEaves/BaxCat

$$f_Y(y) = \frac{1}{y^2}\exp\left(1 - \frac{1}{y}\right). \tag{39}$$

The expected dependence probability between two columns is

$$P(z_i = z_j) = E\left[Y\right] = \int_0^1 y f_Y(y) dy = \int_0^1 \frac{1}{y}\exp\left(1 - \frac{1}{y}\right) dy \approx 0.596. \tag{40}$$

Appendix C
Why regression fails

Disregarding CrossCat's ability to infer the existence of dependencies between variables, one might wonder why use CrossCat, rather than linear regression, to determine the nature of the dependencies. The majority of epistemic trust research evaluates young children. Of the studies we included in our analyses, only one was done with adults. This study creates, what one using tradition meta-analysis methods might consider, outliers in the age-versus-model-variables scatter plots (see Figure C1). Regression is sensitive to outliers, but if we want to create a continuous account of development, we must include these results.

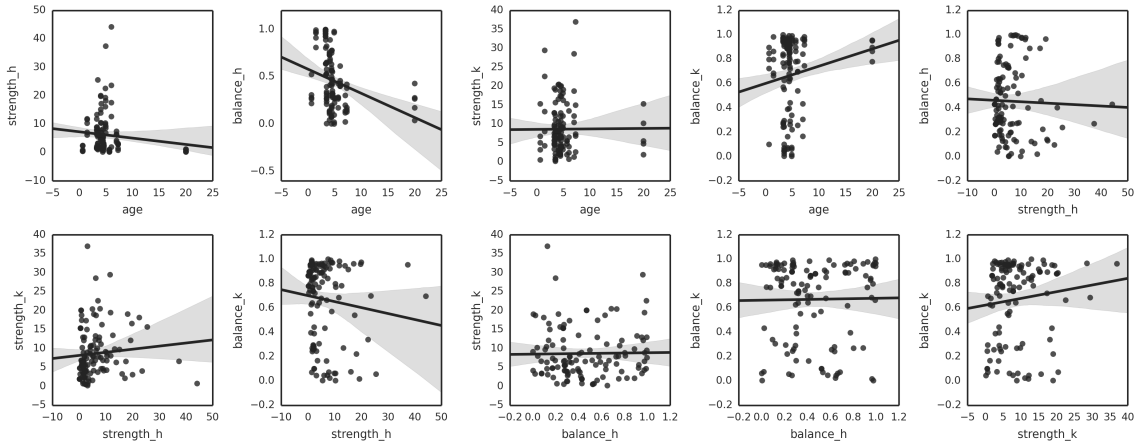


Figure C1. Pairwise plots of age and model parameters for the full dataset. Regression lines are shown with their 95% confidence intervals in gray.

Assuming that we ignore a valuable part of our data and remove the outliers (see Figure C2), we see that remaining data violate most of the assumptions made by standard linear regression. The data are nonlinear and heteroscedastic. One could first look at the pair plots and choose a more appropriate regression method for each pair, but each of these decisions introduces arbitrariness to the model and reduces its generality. CrossCat neither assumes linearity nor homoscedasticity and has no problem dealing with outliers.

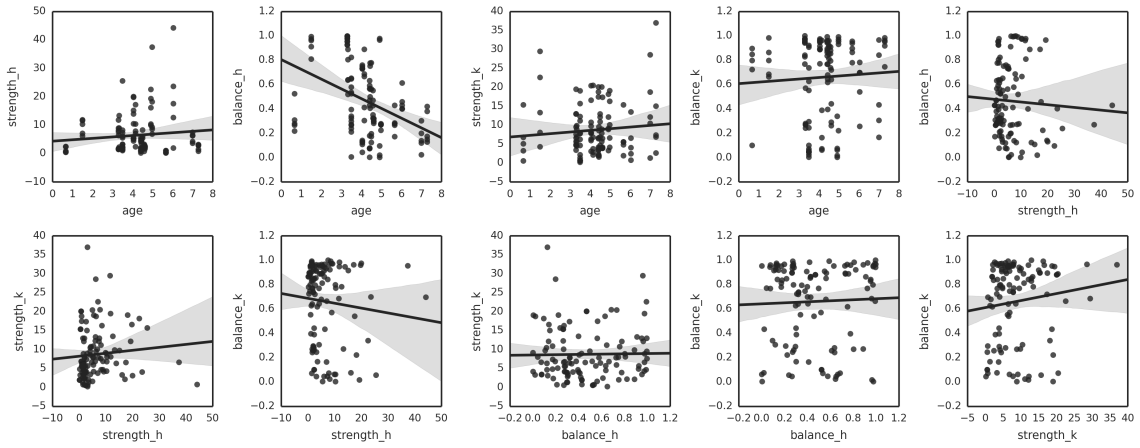


Figure C2. Pairwise plots of age and model parameters for data in which the age is less than 15. Regression lines are shown with their 95% confidence intervals in gray.