# Do I know that you know what you know? Modeling testimony in causal inference

**Daphna Buchsbaum**[1], **Sophie Bridgers**[1], **Andrew Whalen**
**Elizabeth Seiver, Thomas L. Griffiths, Alison Gopnik**
{daphnab, sbridgers, awhalen,seiver,tom_griffiths, gopnik}@berkeley.edu
Department of Psychology, University of California, Berkeley, Berkeley, CA 94720 USA

## Abstract

We rely on both our own observations and on others' testimony when making causal inferences. To integrate these sources of information we must consider an informant's statements about the world, her expressed level of certainty, her previous accuracy, and perhaps her apparent self-knowledge – how accurately she conveys her own certainty. It can be difficult to tease apart the contributions of all these variables simply by observing people's causal judgments. We present a computational account of how these different cues contribute to a rational causal inference, and two experiments looking at adults' inferences from causal demonstrations and informant testimony, focusing on cases where these sources conflict. We find that adults are able to combine social information with their own observations, and are sensitive to the reliability of each. Adults are also sensitive to the accuracy, certainty, and self-knowledge of the informant, a result confirmed by comparing predictions from models with and without these variables.

## Introduction

People face challenging causal learning problems on a daily basis. They have a variety of information they can use to help solve these problems, including directly observed patterns of cause and effect, and social data such as others' statements about existing causal relationships. Having multiple sources available should enhance our causal reasoning, but integration can be difficult, especially when sources disagree. If an informant's causal statements contradict our causal observations, which source should we trust? Informants can be ignorant, mistaken, even deceptive, so one might think we should always trust what we see over what we hear. Yet the world is unpredictable: Observing a phenomenon once does not mean we will reliably observe it again. How do we evaluate these sources and determine which to rely on? We attempt to better understand how people combine different sources of information when making causal inferences, how they integrate their own observations with conflicting testimony, and how this affects their future evaluation of the social informant.

Informant testimony is a key type of social information that guides learning across domains, but the role of testimony in causal learning has not been extensively explored. Research on how we incorporate information from the social context into our causal judgments has shown that both children and adults are skilled causal learners, and can use information from social demonstration to inform their causal inferences (e.g., Kushnir, Wellman, & Gelman, 2008; Sobel & Sommerville, 2009; McGuigan, Makinson, & Whiten, 2011). This work has investigated how we learn by observing other people (e.g., Goodman, Baker, & Tenenbaum, 2009; Buchsbaum, Gopnik, Griffiths, & Shafto, 2011), and by observing different types of people (e.g., Kushnir et al., 2008), but has not examined in detail how we make inferences about their credibility based on their causal statements. Here, we explore how people combine information from causal observations and testimony, both to make causal judgments and to evaluate the informants themselves.

Recently, there has been a growing literature on how people, especially children, evaluate informants (e.g., Borckardt, Sprohge, & Nash, 2003; Corriveau, Meints, & Harris, 2009; Koenig & Harris, 2005), including how children integrate their prior knowledge with informant testimony (e.g., Jaswal, 2010; Jaswal & Markman, 2007). Integrating testimony with our observations is particularly challenging because multiple aspects of informants and their testimony contribute to the value of the information they provide, and to how much they should be trusted in the future. These aspects include the level of certainty informants express, their past accuracy, and their self-knowledge – how well their certainty reflects their true knowledge (for an exploration of a similar idea in the context of eye-witness testimony see Tenney, Small, Kondrad, Jaswal, & Spellman, 2011). The difficulty of combining information from what someone says and what we see is especially apparent when an informant's statements appear to be incorrect. It could be that the informant is actually right, and our own observations were inaccurate. Alternatively, the informant may be a knowledgeable person who has simply misspoken, or she could truly be clueless. Finally, an informant could express certainty or uncertainty about her knowledge, shedding a different light on her inaccuracy. Deciding whether an informant has erred, and if so, why, and whether to trust her in the future is therefore a complex problem.

Bayesian modeling provides a mechanism for explicitly representing the contributions of different sources of information to judgments about causal structure and informant credibility. Previous work has used such models to explore the role of social observations in causal learning (e.g., Goodman et al., 2009), and to evaluate the role of informant knowledgeability and helpfulness (Shafto, Eaves, Navarro, & Perfors, in press). Here, we present a model that helps us evaluate the roles of observed cause and effect patterns, as well as an informant's expressed certainty, current and past accuracy, and awareness of her own knowledge level, when making a causal inference.

In this paper, we first review a study exploring how preschoolers combine information from informant testimony with conflicting information from observed causal data. Next, we introduce a computational model of causal inference from testimony that explicitly represents the roles of informant certainty, accuracy, and self-knowledge, as well as direct causal observations, allowing us to assess the

[1] These authors contributed equally to this work.

contributions of each to a rational causal inference. We then present a series of adult experiments motivated by both the model and the child experiments. Finally, we conclude by discussing how predictions by models including some or all of these variables provide us with further insight into our ability to learn from multiple sources, and the information we use to determine when to trust what other people say.

## Children's Causal Inferences from Testimony

Bridgers, Buchsbaum, Seiver, Gopnik, and Griffiths (2011) presented preschoolers with either an informant who claimed to know which of two blocks was better at activating a machine or an informant who claimed to be guessing, and with observed statistical data that contradicted the informant's claim. The study investigated which source of information (the person or the data) children would rely on, as well as how likely children would be to trust the informant in a new situation. Though both informants made incorrect predictions, the naïve informant demonstrated more self-knowledge because she knew she did not know, while the knowledgeable informant was unaware she was mistaken.

Results from this study imply that preschoolers are sensitive to the certainty and accuracy of an informant – they were more likely to trust the informant's endorsement over the data when the informant was knowledgeable than when she was naïve, and were more likely to trust the knowledgeable informant before her inaccurate statements than afterwards. However, these results also suggest that children may not be as sensitive to an informant's level of self-knowledge since children were as likely to trust the knowledgeable informant (who was mistaken in her certainty) as they were to trust the naïve informant (who was correctly uncertain) in a new situation.

Intuitively, an informant's certainty, past accuracy, *and* self-knowledge should all be useful indicators of her credibility. However, it is challenging to infer the influence of each of these variables and of the causal data simply by examining people's resulting inferences. For example, does children's failure to differentiate between the informants in Bridgers et al. (2011) mean they lack a concept of self-knowledge altogether or that they weigh other cues to reliability more heavily? Given children's performance, would adults trust a previously inaccurate but uncertain informant over one who was certain and inaccurate, or would they also use a simpler strategy, for instance mistrusting anyone who was previously incorrect? A computational model of how people combine information from both observed data and an informant to determine the likelihood of a causal relationship could help clarify the factors impacting people's resolution of the conflict, and their decision of whether or not to trust the informant in the future.

## Modeling Causal Inference From Testimony

People may take into account a variety of social information when making causal inferences from testimony. As noted earlier, there is evidence that both children and adults are sensitive to an informant's expressed certainty and previous accuracy. People may also have pre-existing assumptions about how knowledgeable others tend to be, and how often others make mistakes in their assertions. Finally, there is some evidence that at least adults are sensitive to others' self-knowledge (Tenney et al., 2011).

This social information also interacts with the individual's own causal observations. It can therefore be difficult to determine which of these variables contribute to people's resulting causal inferences. We present an explicit model of how these variables could interact. We then evaluate the roles of these different variables by comparing people's causal judgments to those that would be normative under our model, as well as under simpler models that do not explicitly represent the informant's knowledge and self-knowledge.

Our model is defined in terms of observed variables representing causal outcomes, statements by an informant about the causal strengths of potential causes, and about her level of certainty about her causal knowledge. The model also has hidden variables representing the actual causal strengths of the potential causes, the informant's general level of knowledgeability, her specific knowledge of the individual causes, and her level of self-knowledge – how well she knows what she knows. We capture the complex relationships among these variables in a graphical model (see Figure 1).

In this model, we assume that all the variables are binary valued, as they were presented to children in the Bridgers et al. (2011) experiments. Each cause $c$ has a causal strength $w_c$, such that $p(e_c = 1 \mid c, w_c) = w_c$ for effects $e_{c,i}$ where $p(w_c = \rho) = \gamma$ and $p(w_c = 1-\rho) = 1-\gamma$. Here, $\rho$ is some relatively high probability of effect, corresponding to the causal strength "almost always makes it go" in Bridgers et al. (2011), with $1-\rho$ corresponding to "almost never makes it go."

The informant's prediction $r_c$ about the causal strengths of each cause depends on the true causal strength $w_c$, and on her knowledge of the cause $k_c$. Here, we assume that $k_c \in \{0,1\}$, corresponding to two possible states of knowledge of a cause: guessing and knowing. If $k_c = 1$ (the informant knows about the causal strength of c) then $p(r_c = w_c \mid k_c = 1, w_c) = 1-\varepsilon$, meaning an informant with knowledge of cause $c$ will predict the true value of $w_c$ with probability $1-\varepsilon$, but with small error probability $\varepsilon$ will report the incorrect value. On the other hand, if $k_c = 0$ and the informant is guessing about cause $c$, then $p(r_c = w_c \mid k_c = 0, w_c) = p(r_c = w_c \mid k_c = 0) = 0.5$, that is, the informant will choose uniformly at random between the two possible causal strengths.

We assume that the probability of the informant knowing about a particular cause depends on the informant's global knowledgeability $g \in \{0,1\}$, with the informant having probability $\kappa$ of being globally knowledgeable. If $g = 1$ then the informant is globally knowledgeable and $p(k_c = 0 \mid g = 1) = 1-\tau$ and $p(k_c = 1 \mid g = 1) = \tau$, that is, the informant is knowledgeable about cause $c$ with some relatively high probability $\tau$. Conversely, if $g = 0$ and the informant is globally ignorant then $p(k_c = 0 \mid g = 0) = \tau$ and $p(k_c = 0 \mid g = 0) = 1-\tau$.

Finally, we need to represent the informant's statement

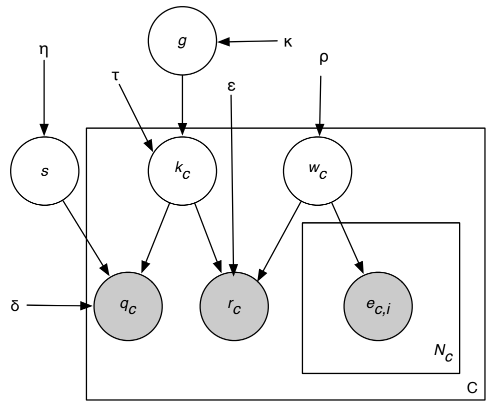

Figure 1: Causal testimony graphical model

about her knowledge of cause $c$. The informant's statement $q_c$ depends on her knowledge $k_c$, and her level of self-knowledge $s$. We assume that $s \in \{0,1\}$, corresponding to two possible states of self-knowledge: accurate and inaccurate. If $s = 1$ (the informant has accurate self-knowledge) then $p(q_c = k_c \mid s = 1, k_c) = 1 - \delta$, meaning the informant will accurately report her level of knowledge $k_c$ with probability $1 - \delta$, but with small error probability $\delta$ will report her level of knowledge inaccurately.

If $s = 0$ and the informant has inaccurate self-knowledge then $p(q_c = k_c \mid s = 0, k_c) = p(q_c = k_c \mid s = 0) = 0.5$, that is, the informant will choose uniformly at random when stating her knowledge of the causal system. We assume that any given informant has probability $\eta$ of having accurate self-knowledge, and $1 - \eta$ of having inaccurate self-knowledge.

We assume $p(\text{choose}\, c) \propto p(\text{effect} \mid c, \text{obs})$, meaning people choose causes in proportion to how likely they think they are to produce the effect, given their observations (including the informant's statements). This is computable from the model and the dependencies defined in our graphical model (see Figure 1). To evaluate our model, we conducted a series of experiments with adults, exploring whether they can successfully use an informant's certainty, accuracy, and self-knowledge when making causal inferences.

## Experiment 1: Adult Inferences from Testimony

We investigate how adults resolve a conflict between an informant's explanation of how a causal system works and actual demonstrations of that system, closely following the procedure of Bridgers et al. (2011). We hypothesized that like preschoolers, adults would be sensitive to the certainty and accuracy of the informant, and be more likely to trust an informant who claimed to be knowledgeable over one who claimed to be naïve, and less likely to trust a previously inaccurate informant. However, unlike children, we predicted that adults would be sensitive to an informant's level of self-knowledge, and would be more likely to extend their trust to a previously incorrect informant who had claimed ignorance than a previously incorrect informant who had claimed knowledgeability.

### Methods

**Participants** A total of 204 participants were recruited: 100 were UC Berkeley undergraduates who received course credit and 104 were Mechanical Turk workers who were compensated \$0.50. Participants were randomly assigned to one of two experimental groups: the Knowledgeable condition ($n$ = 103) or the Naïve condition ($n$ = 101).

**Stimuli** The experiment was a web-administered survey involving text and pictures. An image of a brown-haired woman was the informant, and an image of a blonde woman was her assistant. The machine was an image of a green box with a black top. The activated machine had a yellow top and musical notes were placed around it. The blocks were a green rectangle, a pink disk, an orange cube, and a blue cylinder.

**Procedure** First, a woman named Ann (the *informant*) introduced a machine that could light up and play music when certain blocks were placed on top. She then introduced two different blocks and explained that one block almost always activated the machine (the *endorsed* block), while the other block almost never did (the *unendorsed* block). In the Knowledgeable condition, the informant claimed that she really knew which block was better at activating the machine, while in the Naïve condition, the informant claimed that she was just guessing. Besides this difference, the procedure was identical across conditions. Ann then said she needed to leave, and her assistant Jane continued the experiment.

Jane first asked participants to rate how likely each block would be to activate the machine on a scale from 0 (definitely will not) to 10 (definitely will) (the *prior* rating). Jane demonstrated each block on the machine, providing probabilistic evidence that contradicted Ann's claim: the endorsed block only activated the machine 2/6 times, while the unendorsed block activated it 2/3 times.[1] Participants were then again asked to rate how likely each block would be to activate the machine (the *causal* rating).

Finally, Ann returned with two new blocks, and in both conditions, claimed she *knew* that one block almost always activated the machine and that the other almost never did. Ann then left once more, and Jane asked the participants to rate how likely they thought these new blocks were to activate the machine (the *generalization* rating).

### Results and Discussion

For a summary of the results see Table 1. We analyzed causal efficacy ratings with a $2 \times 3 \times 2$ repeated measures ANOVA, with endorsement (endorsed or unendorsed), and rating phase (prior, causal, generalization) as the within subject variables, and knowledge condition (Knowledgeable or Naïve) as the between subjects variable. There was a main effect of endorsement – adults rated the endorsed block more highly across phases and conditions ($F$(1, 1206) = 77.69, MSE = 372.1, $p < 0.001$). There was also an effect of endorsement

[1] This pattern of probabilistic data is the same as was used in Experiment 3 of Kushnir and Gopnik (2007).

× condition ($F(1, 1206) = 73.19$, MSE = 350.5, $p < 0.001$), with the endorsed block rated higher in the Knowledgeable condition across phases, and of endorsement × phase ($F(2, 1206) = 62.18$, MSE = 297.8, $p < 0.001$), with the rating of the endorsed block decreasing and of the unendorsed block increasing in the causal phase. Finally, there was a significant three-way interaction of endorsement × phase × condition ($F(2, 1206) = 13.89$, MSE = 66.5, $p < 0.001$), indicating that the degree to which the ratings change between phases varied by the claimed knowledge level of the informant and whether the block was endorsed.

We explored the particulars of these findings via planned t-tests. In the prior phase of both conditions, adults were more likely to give the endorsed block a higher rating (paired t-tests. Knowledgeable: $t(102) = 17.76$, $p < 0.001$. Naïve: $t(100) = 5.02$, $p < 0.001$) though participants in the Knowledgeable condition gave the endorsed block a higher rating than those in the Naïve condition (two sample t-test, $t(202) = 5.72$,$p < 0.001$). These results suggest that before seeing any data, adults in both conditions were likely to trust the informant's testimony, but were also sensitive to the certainty expressed by the informant.

In the causal phase, there was no difference in adults' ratings of the endorsed and unendorsed blocks in the Knowledgeable condition (paired t-test, $t(102) = 1.39$, $p = 0.17$), while adults in the Naïve condition gave the unendorsed block a higher rating (paired t-test, $t(102) = 7.18$, $p < 0.001$). Adults in the Knowledgeable condition gave the endorsed block a higher rating than adults in the Naïve condition (two sample t-test, $t(202) = 2.00$, $p < 0.05$) and vice versa for the unendorsed block (two sample t-test, $t(202) = 3.62$, $p < 0.001$). The fact that in the causal phase adults thought the two blocks had approximately equal causal efficacy in the Knowledgeable condition but rated the unendorsed block more highly in the Naïve condition suggests that participants were responding to both the observed statistical data and the claimed knowledge level of the informant.

Finally, in the generalization phase, adults in both conditions gave the endorsed block higher ratings (paired t-tests. Knowledgeable: $t(102) = 13.21$, $p < 0.001$. Naïve: $t(102) = 8.23$, $p < 0.001$). Unlike the previous phases, there was no difference between conditions in ratings of the endorsed block (two sample t-test, $t(202) = 0.82$, $p = 0.41$).

We can also compare ratings in the two no-data phases – prior and generalization – to capture how participants' evaluation of the informant might have changed after receiving evidence about her accuracy in the intervening causal phase. Adults' ratings decrease between prior and generalization in the Knowledgeable condition (paired t-test $t(102) = 3.30$, $p < 0.01$), while they increase in the Naïve condition (paired t-test, $t(100) = 2.12$, $p < 0.05$). This difference suggests that adults may actually be sensitive to an informant's self-knowledge, increasing their trust in an informant who was incorrect but uncertain in the past over an informant who was incorrect but certain. Our results in the generalization phase imply that adults are willing to trust both informants more or less equally regardless of their level of self-knowledge. However, due to the stochastic nature of the data, participants may have made excuses for the discrepancy between the data and the knowledgeable informant's endorsement, possibly appealing to hidden causes such as a faulty machine part that would explain away the conflict. In Experiment 2, we contrasted an informant's testimony with deterministic data to see if increasing the apparent inaccuracy of the informant would reveal a use of informant self-knowledge.

Table 1: Mean ratings and standard errors for Experiment 1

| Mean Rating (std err) | Endorsed | Unendorsed |
|---|---|---|
| Prior Naïve | 6.46 (.24) | 4.16 (.25) |
| Prior Knowledgeable | 8.22 (.19) | 1.94 (.19) |
| Causal Naïve | 4.09 (.18) | 6.29 (.20) |
| Causal Knowledgeable | 4.66 (.22) | 5.18 (.23) |
| Gen Naïve | 7.20 (.25) | 3.27 (.25) |
| Gen Knowledgeable | 7.45 (.18) | 2.71 (.20) |

Table 2: Mean ratings and standard errors for Experiment 2

| Mean Rating (std err) | Endorsed | Unendorsed |
|---|---|---|
| Prior Naïve | 6.58 (.31) | 3.68 (.32) |
| Prior Knowledgeable | 8.61 (.18) | 1.47 (.18) |
| Causal Naïve | 1.15 (.35) | 9.45 (.11) |
| Causal Knowledgeable | 2.82 (.54) | 8.0 (.45) |
| Gen Naïve | 7.0 (.44) | 3.47 (.47) |
| Gen Knowledgeable | 5.88 (.52) | 4.15 (.51) |

## Experiment 2: Deterministic Data

We replicated Experiment 1 but with deterministic data, to explore how changing the strength of the data might impact adults' inferences. We predicted that adults would weight conflicting deterministic data more heavily than conflicting probabilistic data, and would therefore prefer the unendorsed block more often in the causal phase. We also predicted that the stronger data would exaggerate the knowledgeable informant's lack of self-knowledge, leading adults to consider the naïve informant's testimony as more reliable than the knowledgeable informant's in the generalization phase.

### Method

**Participants** A total of 74 participants recruited from Mechanical Turk were compensated $0.50 and randomly assigned to the Knowledgeable condition ($n$ = 34) or the Naïve condition ($n$ = 40).

**Stimuli** Stimuli were identical to those in Experiment 1.

**Procedure** The procedure was the same as in Experiment 1 except that the endorsed block activated the machine 0/6 times, while the unendorsed block activated it 6/6 times.

### Results and Discussion

We performed the same analyses for Experiment 2 as we did for Experiment 1. Due to limited space, we discuss only the most relevant results here. Results are summarized in Table 2.

Adults in both conditions of Experiment 2 gave lower ratings to the endorsed block in the causal phase as compared to Experiment 1 (two sample t-tests. Knowledgeable: $t(135) = 3.77$, $p < 0.001$. Naïve: $t(139) = 7.90$, $p < 0.001$). Adults thus were sensitive to the increased strength of the data in the second experiment, and were less likely to trust the knowledgeable informant's claim over the data than participants who observed probabilistic data.

In the generalization phase of Experiment 2, adults rated the endorsed block in the Naïve condition as more causally efficacious than in the Knowledgeable condition, though this effect was only marginal ($t(72) = 1.69$, $p < 0.10$). Thus, adults in the Knowledgeable condition were less inclined to trust the informant's endorsement than adults in the Naïve condition. This finding suggests that differences in self-knowledge *do* impact adults' evaluations of informants since adults appeared to place more confidence in the statement of the person whose prior certainty reflected their accuracy.

Comparing across experiments, we found that adults in the Knowledgeable condition of Experiment 1 gave the endorsed block a higher generalization rating than those in Experiment 2 (two sample t-test. $t(135) = 3.77$, $p < 0.001$). On the other hand, there was no difference in adults' generalization ratings of the endorsed block in the Naïve condition (two sample t-test. $t(135) = 0.41$, $p = 0.68$) across experiments. As predicted, increasing the strength of the conflicting data magnified the knowledgeable informant's inaccuracy. However, since the naïve informant claimed ignorance, this change did not affect how adults evaluated future information from this informant. In general, trust in the knowledgeable informant decreased with increasingly conflicting data (i.e. from situations of no conflict (prior phases) to situations of prior conflicting data (generalization phases) to situations of directly conflicting data (causal phases)).

Finally, comparing between phases of Experiment 2, in the Knowledgeable condition, adults' ratings of the endorsed block decrease between prior and generalization phases (paired t-test, $t(33) = 4.99$, $p < 0.001$). Thus, adults are likely to initially trust the endorsement of an informant, whereas they are less likely to extend that trust to a new situation (the generalization phase) after observing evidence that contradicts the informant's prior claim. Conversely, in the Naïve condition there was no difference between adults' prior and generalization phase ratings (paired t-test, $t(39) = 0.79$, $p < 0.43$). Thus, adults' evaluation of the credibility of the naïve informant does not appear to have changed after observing the conflicting data. This further suggests adults are sensitive to self-knowledge when determining the usefulness of an informant's statement: They found a previously uncertain, inaccurate informant more trustworthy than a previously certain, inaccurate one.

One alternative account is that adults consider the knowledgeable informant to be deceptive rather than having poor self-knowledge. However, if adults suspect the knowledgeable informant is deceiving them by saying the opposite of what she knows, then we would expect them to go against her endorsement in the generalization phase, rating the unendorsed block as more causally efficacious than the endorsed block. However, while adults' ratings of the endorsed block decrease in the generalization phase, they are still greater than their ratings of the unendorsed block, implying that adults view this informant as less reliable rather than intentionally dishonest.

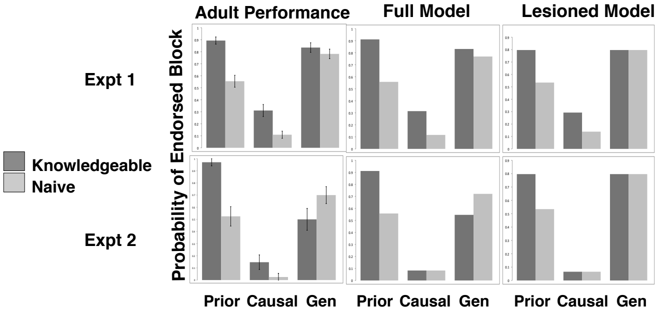

Figure 2: Adult performance vs model predictions

## Modeling People's Inferences

We can use our model to further test whether an informant's expressed certainty, accuracy, and apparent self-knowledge inform adults' causal judgments. We first evaluate the model by fitting it to data from Experiment 1. In order to be consistent with Bridgers et al. (2011), where children were asked to choose the better block, we assumed that for each set of ratings adults would choose the block they had rated as most likely to be effective. We then optimized the log likelihood of these choices under the model. The best fitting model corresponded to model parameters of $\rho = 0.91$, $\gamma = 0.40$, $\varepsilon = 0.01$, $\delta = 0.1$, $\tau = 0.91$, $\kappa = 0.74$ and $\eta = 0.93$, (Pearson's $r$=0.999, $p < 0.001$). However, a reasonable range of values around these settings also fit the data well. Of particular note are the values of $\kappa$ and $\eta$, corresponding to a belief that most people have good general knowledge, but a substantial minority are relatively clueless, and that almost everyone is aware of their own knowledge level, but a small number of people tend to inaccurately assess what they know.

We tested the generalization of these model parameters by looking at how well they predict adult performance in Experiment 2. The parameters fit to the Experiment 1 data also provide a good fit to the results of Experiment 2 (Pearson's $r$=0.9664, $p < 0.001$). This suggests that our model is accurately capturing human performance, so we can use it to tease apart the contributing variables to adult causal inferences.

We then conducted a nested model comparison, examining whether including the informant's global knowledgeability and their self-knowledge adds explanatory value to the model, by creating a series of "lesioned" models, lacking global knowledge $g$ and self knowledge $s$. This approach controls for the additional free parameters in the more complex model. Removing global knowledge corresponds to a model that assumes that all informants have the same probability of knowing about all causes, and that an informant knowing about one cause does not make her any more likely to know about others. Removing self-knowledge corresponds to a model where informants' statements of certainty always reflect their true knowledge – if they say they know something, then they must really know it.

Compared to a model lacking both global knowledge and self-knowledge variables, adding global knowledge to the model resulted in a marginally significant ($\chi^2(1) = 3.29, p = .07$) improvement in model fit. Adding self-knowledge on its own did not improve model performance ($\chi^2(1) = 0.392, p = 0.53$), however adding both self-knowledge and global knowledge variables significantly improved model fit over having only global knowledge ($\chi^2(1) = 22.04, p < 0.001$), or having neither ($\chi^2(1) = 25.30, p < 0.001$) (see Figure 2).

Qualitatively, the addition of global knowledge and self-knowledge only modestly improves the model fit to Experiment 1, their biggest effect is on the fit to Experiment 2. Of particular interest in Experiment 1, the full model and the "lesioned" model (without both variables) appear to make similar predictions about adults' performance in the generalization phase, suggesting that contrary to our initial intuitions, even with a concept of self-knowledge it may still be rational to extend trust equally to both informants. However, these two models make different predictions for generalization performance in Experiment 2, with the full model more accurately capturing adults' inferences. This supports the interpretation that in the first experiment, participants continued to extend trust to the knowledgeable informant not because they lacked a concept of self-knowledge but by explaining away the informant's apparent incorrectness, inferring that the ambiguous data could have been "unlucky." However, in Experiment 2, where the data more strongly supports the inference that the informant was incorrect, these model results suggest that it requires both a concept of general knowledge ("if this person was wrong before, they're more likely to be wrong again"), and self-knowledge ("they said they 'knew' before and they didn't, why should I think they know now?" vs. "they said they didn't know, so it's okay that they were wrong"), in order to infer that the naïve informant is more deserving of trust in the generalization phase.

Overall, our nested model comparison demonstrates that adults take into account the informant's past performance when deciding how much to weight their current testimony, and in particular that adults are sensitive to both the apparent knowledgeability of the informant and their level of self-knowledge.

## Conclusion

We examined how people combine an informant's statements about a causal system with direct observations of that system, and how this influences their evaluation of the informant's knowledgeability and credibility. Together, Experiments 1 and 2 suggest that adults are weighting and integrating evidence from both observed data and the informant in their causal inferences, and that their trust in the informant is moderated by the degree to which the informant's testimony conflicts with the data. Adults were sensitive to both the informant's certainty and accuracy, and to how well the informant's certainty reflected her accuracy. These findings support our intuition that self-knowledge is a valuable cue adults can use to determine the trustworthiness of an informant's testimony. The close fit of the model to adult performance, and its ability to generalize from Experiment 1 to Experiment 2, confirms that adults are rationally integrating their direct observations with testimony from a social informant when making causal inferences. Our model also strongly suggests that representing self-knowledge is necessary to making these inferences, and that adults could not be using a simpler strategy such as only tracking previous inaccuracy. Overall, these results provide us with further insight into how we learn from and evaluate the sources of information available to us and in particular, revealing that knowing that you do not know can be just as important as knowing that you know.

**Acknowledgments.** This work was supported by grant number FA-9550-10-1-0232 from the Air Force Office of Scientific Research, the NSF Graduate Research Fellowship, NSF Grant BCS-1023875, and the McDonnell Foundation Causal Learning Initiative.

## References

Borckardt, J., Sprohge, E., & Nash, M. (2003). Effects of the inclusion and refutation of peripheral details on eyewitness credibility. *Journal of Applied Social Psychology*, *33*(10), 2187–2197.

Bridgers, S., Buchsbaum, D., Seiver, E., Gopnik, A., & Griffiths, T. L. (2011). Which block is better at making the machine go?: How children balance their trust in an informant vs. the data. *Poster presented at Biennial Meeting of the Cognitive Development Society*.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331-340.

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

Goodman, N. D., Baker, C. L., & Tenenbaum, J. B. (2009). Cause and intent: Social reasoning in causal learning. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Jaswal, V. K. (2010). Believing what you're told: Young children's trust in unexpected testimony about the physical world. *Cognitive Psychology*, *61*, 248-272.

Jaswal, V. K., & Markman, E. M. (2007). Looks aren't everything: 24-month-olds'willingness to accept unexpected labels. *Journal of Cognition and Development*, *8*(1), 93-111.

Koenig, M., & Harris, P. (2005). The role of social cognition in early trust. *Trends in Cognitive Sciences*, *9*(10), 457–459.

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

Kushnir, T., Wellman, H., & Gelman, S. (2008). The role of preschoolers' social understanding in evaluating the informativeness of causal interventions. *Cognition*, *107*(3), 1084–1092.

McGuigan, N., Makinson, J., & Whiten, A. (2011). From over-imitation to super-copying: Adults imitate causally irrelevant aspects of tool use with higher fidelity than young children. *British Journal of Psychology*, *102*, 1-18.

Shafto, P., Eaves, D., Navarro, D. J., & Perfors, . A. (in press). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*.

Sobel, D. M., & Sommerville, J. (2009). Rationales in children's causal learning from others' actions. *Cognitive Development*, *24*(1), 70–79.

Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., & Spellman, B. A. (2011). Accuracy, confidence, and calibration: How young children and adults assess credibility. *Developmental Psychology*, *47*(4), 1065-1077.