Enough is enough: Inductive sufficiency guides learners' ratings of informant helpfulness

Patrick Shafto (p.shafto@louisville.edu)
University of Louisville

Hyowon Gweon (hyora@mit.edu)
Massachusetts Institute of Technology

Chris Fargen (cmfarg01@louisville.edu)
University of Louisville

Laura Schulz (lschulz@mit.edu)
Massachusetts Institute of Technology

Abstract

Much of what we learn, we learn from others. What guides learners' choice of informants? Research suggests that learners resist informants who provide incorrect information or insufficient information for accurate inference. Here we propose that learners' choices of informants are rationally guided by the extent to which evidence supports accurate inference, rather than the sheer amount of evidence provided. Extending recent research formalizing pedagogical reasoning, we propose a computational model of efficient teaching. We present an experiment on adults testing three different hypotheses about learners' preferred level of the amount of data. The results suggest that learners care about the inductive sufficiency of evidence, rather than the amount of evidence provided. We conclude by discussing the implications of these findings for cognition and cognitive development.

Keywords: Trust; Pedagogical reasoning; Bayesian model.

People face a seemingly intractable problem in learning about the world. There is an endless amount of information to learn, but relatively limited time to acquire information. Fortunately, learners are surrounded by other agents who can help them learn. However, although some people may be valuable sources of information, not all are, and learners must decide whom to ask for information. What governs learners' choices of informants?

Most prior research about learners' sensitivity to the reliability of informants has been conducted with children. Koenig and Harris (2005) found that by four years of age, children track whether informants have been correct or incorrect in the past and use this to guide their future choices of informants. Moreover, children are sensitive to parametric variations in informants' accuracy (Pasquini, Corriveau, Koenig, & Harris, 2007). Additionally, children use information about group consensus to select informants (Corriveau, Fusaro, & Harris, 2009). These results suggest that by four years of age, children can use diverse cues to establish the reliability of informants.

However, there is good reason to believe that reliability is not the only factor that influences children's epistemic trust. A recent study by Gweon, Pelton, and Schulz (2011) suggests that children not only expect teachers to provide accurate data, but also expect teachers to provide inductively sufficient data. Children gave lower ratings to a teacher who showed one function of a toy that actually had multiple functions, than to a teacher who gave the same demonstration on a toy that actually had just the one function.

Indeed, one advantage of social learning is that it reduces the amount of data required for accurate inference by allowing the learner to make inductive inferences from small amounts of data. How much evidence is enough?

We hypothesize that people do not simply use the sheer quantity of data to decide how helpful a teacher is, but instead consider the extent to which the data provided supports accurate inductive inferences. If a learner's goal is simply to acquire as much data as possible, people should always prefer a teacher who offers more data. However, if the learner's goal is specifically to acquire as much data as necessary for accurate inference, then two teachers can be considered equally helpful, even if one of them provides much less data overall. Consider the toys in Figure 1a; the toys have a number of knobs, which when pressed may or may not cause exciting effects. As a learner, you may be curious to know how many of the knobs cause an effect. You may also have past experience with toys like this, and this experience might generate expectations about how many knobs are likely to work. For instance, you may know that just a few knobs (e.g., two on average) cause effects and the rest do not (independent of the total number of knobs). If you were to learn about one of the toys, would you choose a demonstrator who exhaustively pressed all of the knobs, or one who pressed a few working knobs and stopped? (See Figure 1b.)

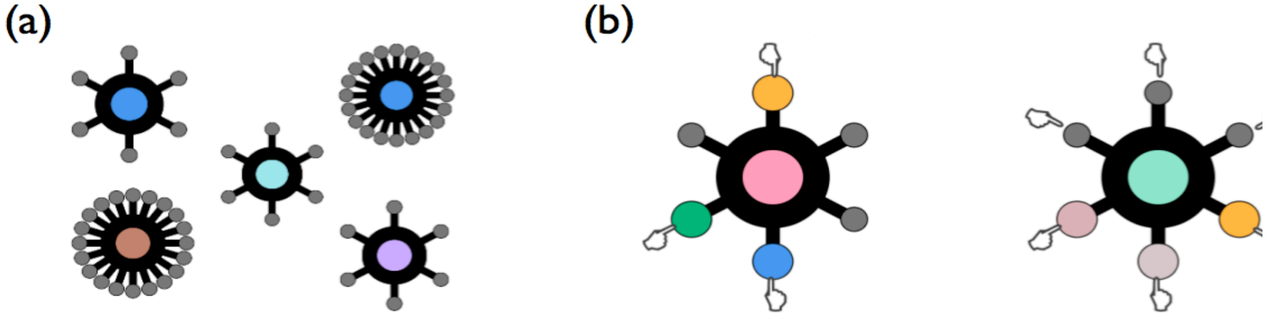

Figure 1: Figure illustrating the toys used the in the experiment, and possible demonstrations. (a) Toys could have 6 or 20 knobs, which when pressed may (or may not) lead the toy to create a sound and the button to change color. (b) Two possible sets of demonstrations. In the first case, only three knobs are pressed, all of which elicit effects. In the second case, all of the knobs are pressed, only three of which elicit effects.

Clearly, the exhaustive demonstration would provide the most data; indeed, such a demonstration would be deductively sufficient to infer the number of working knobs. If the only goal is to maximize the overall amount of data, the learner should always prefer more demonstrations to fewer demonstrations.

By contrast, if the learner is sensitive to the cost of providing data and cares most about minimizing this cost, then the learner should penalize teachers who provide data beyond what is necessary to make a reasonable guess. Rather than thinking that a teacher who provides exhaustive evidence is being helpful (e.g., by demonstrating that the remaining knobs are in fact inert and thereby marginally reducing the uncertainty), a learner who has a strong bias against incurring the cost of additional demonstrations should resist demonstrations that are consistent with the learner's beliefs. In this case, learners should prefer informants who provide fewer demonstrations.

However, if as we hypothesize, learners value data that leads to accurate inferences, learners may be satisfied with seeing just a few working knobs (since the learner can infer that informant omitted superfluous demonstrations of the inert knobs), but also may be happy to see additional demonstrations that provide maximal certainty about the toy. Such a flexible trade-off between efficiency and certainty would lead to a preference for inductively sufficient demonstrations.

Building off of recent research formalizing data selection in teaching situations, we introduce a computational model of efficient teaching. The model captures these three hypotheses about learners' choices of informants: that learners strongly prefer informants who offer as much data as possible; that learners are very sensitive to the cost of data and thus prefer teachers who offer as little data as possible, or finally, that learners care that the data supports accurate induction but are happy to acquire additional data as well. In our experiment, we ask adult participants to choose between informants who provide data that is always true but varies in quantity and informativeness. We conclude by contrasting our research with previous findings and discussing the implications for cognition, cognitive development and education.

A computational model of efficient teaching

To formalize what constitutes sufficient data, we must consider which data should be chosen and the degree to which the data increase the learners' certainty relative to the added cost of the demonstration. To do so, we adopt a Bayesian learning perspective, building off Shafto and Goodman's (2008) research formalizing teaching and learning in pedagogical settings.

In Bayesian learning, the goal for the learner is to infer the probabilities of different hypotheses, *h*, given data, *d*. The degree to which the learner believes a hypothesis after observing the data---the learner's posterior beliefs---are denoted $P(h|d)$. According to Bayes' theorem, posterior beliefs are determined by the product of the learner's prior beliefs in the hypothesis, $P(h)$, and the probability of sampling the data assuming the hypothesis is true, $P(d|h)$.

Standard approaches to learning typically assume that data are sampled randomly. However, in pedagogical contexts in which the informant is knowledgeable and the informant's goal is to help the learner infer the true hypothesis, the data are not randomly sampled, but purposefully selected. Shafto and Goodman (2008) formalized teaching and learning in such pedagogical situations. The key differences between pedagogical and random sampling are that the teacher is assumed to be knowledgeable and helpful in her choice of data, and the learner believes that the teacher is knowledgeable and helpful. Teaching is formalized as choosing data that tend to maximize the learner's probability of inferring the correct hypothesis, $P_T(d|h) \propto P_L(h|d)$, where the subscripts T and L indicate teacher and learner, respectively. Learners update their beliefs using the knowledge that teachers are choosing data purposefully,

$$P_L(h/d) \propto P_T(d/h)P_L(h). \qquad (1)$$

That is, the key difference between Equation 1 and standard approaches to learning is that in Equation 1 the learner updates her beliefs based on the assumption that the teacher chooses data to help the learner infer the correct hypothesis.

Here we propose that teachers, in addition to choosing data that is helpful, may also consider the *degree* to which additional data increase the learner's certainty. Similarly, learners may vary in how much data they expect the teacher to provide. To capture this difference, we introduce prior probabilities of choosing data. The pedagogical model can be extended to reflect this fact by introducing a term, P(d), in the teacher's choice of data,

$$P_T(d|h) \propto P_L(h|d)P(d). \qquad (2)$$

Data are assumed to have a cost, and differences in the cost of data capture the three hypotheses of interest.

The probability of data P(d) depends on two factors: the number of total demonstrations, n, and the cost of an individual demonstration, c. Intuitively, an informant may be biased toward presenting more data, less data, or may be unbiased; the learner may accordingly have different expectations of the informant. These three possibilities correspond to three qualitatively different cost parameters in our model. If the learner expects the teacher to demonstrate as much data as possible then providing fewer demonstrations incurs a higher cost. If the learner prefers to minimize the number of demonstrations, then given a choice between more data and less data, less data is more probable; the total cost of a set of demonstrations increases with the number of demonstrations. If the learner is willing to accept any evidence that it is inductively sufficient, then any amount of data is equiprobable; the total costs are constant, independent of the quantity of data. To capture these possibilities, we formalize the prior probability of the data as

$$P(d) \propto e^{-cn}. \qquad (3)$$

A negative value of c corresponds to an expectation of more data; a positive value of c corresponds to an expectation of less data; c=0 corresponds to equiprobable data.

These hypotheses generate different predictions about learners' choices of informants. In the following experiment, we investigate how learners evaluate informants. Participants are asked to make a choice between two informants, which we model using a log-likelihood ratio. To test the hypotheses, we treat the cost as a free parameter and fit it to the behavioral data. If the best-fitting cost parameter is less than 0, then this would support the hypothesis that learners prefer as much data as possible; if the best-fitting parameter is greater than 0, this would support the hypothesis that learners prefer to minimize data; if the best-fitting parameter is approximately 0, then this supports inductive sufficiency, the hypothesis that learners want enough data to make a confident inference but are happy to accept additional data.

Experiment: Who is the better teacher?

To investigate learners' choices of informants, we conducted an experiment in which two informants provided demonstrations on different toys (as in Figure 1). The experiment included two conditions. In one condition, either one, two or three knobs worked on the toys (Consistent condition); in the other condition, between 1/6 and 1/2 of the knobs worked on the toys (Proportional condition). These two conditions allowed us to generate cases in which the model makes different predictions about identical sets of evidence. (See Model predictions below.)

Each demonstration consisted of informants pressing knobs that either elicited effects or were inert. Trials varied in the number of demonstrations, as well as their composition. Additionally, toys varied in the number of total knobs (either 6 or 20). Participants used a sliding scale to indicate which informant was relatively more helpful.

This design allows us to assess the correlation between the model predictions and people's choices. In addition, we can investigate specific cases where the three accounts generate contrasting predictions.

Method

Participants. Forty-four University of Louisville undergraduates (22 per condition) participated in exchange for partial or extra credit.

Materials. Participants saw a series of novel toys, described as wugs or daxes on a computer screen. The toys had either 6 or 20 knobs extending from a central sphere (see Figure 1). Clicking on a knob caused a change in its color and size. Only some knobs elicited the effect when clicked; the rest were inert.

Design. Participants were randomly assigned to one of two conditions: Consistent or Proportional. In both conditions, participants first interacted with three six-knob-toys and then with three twenty-knob toys to learn how many knobs worked on average. In the Consistent condition, the six-knob toys had one, two or three working knobs, as did the twenty-knob toys. In the Proportional condition, the six-knob toys again had one, two, and three working knobs but the twenty-knob toys had 3, 7, and 10 working knobs.

Procedure. Participants were seated at Mac Pro Desktop computers. The experiment proceeded in two phases: training and testing. In the training phase, participants learned how many knobs tended to work by interacting with the toys, as described in the Design section. At the end of the training phase, participants indicated how many knobs worked on average for the six-knob and twenty-knob toys. These questions were checks to ensure that the training phase was successful in inducing appropriate prior beliefs about the toys. Participants were then given feedback about the actual average number of working knobs.

During the testing phase, participants were presented with a series of pairs of unique informants each performing demonstrations on a unique toy. The screen was split in half; in each half of the screen there was an informant with a toy. In the Consistent condition, the pairs were generated by

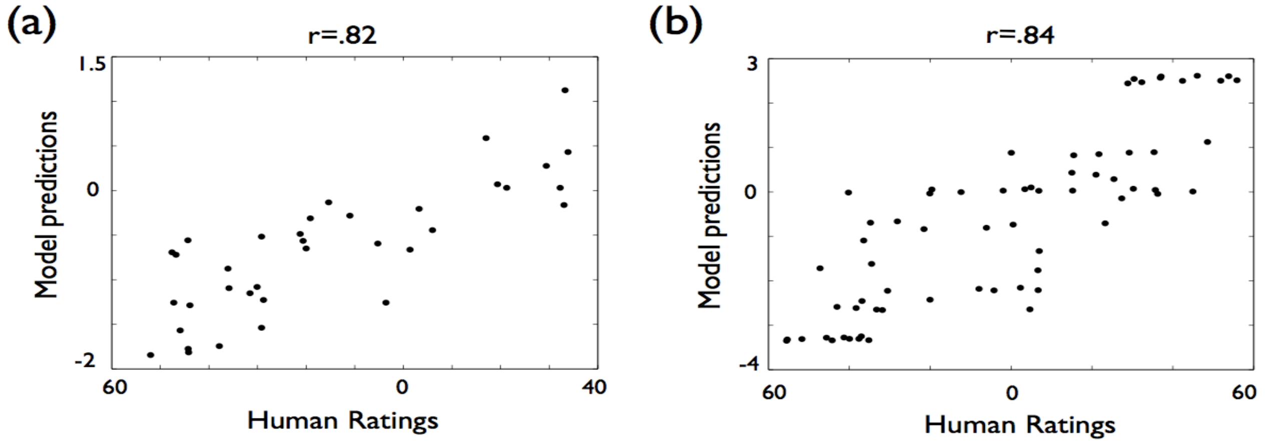

Figure 2: Correlations between human judgments in the (a) Consistent and (b) Proportional conditions. Overall, there is a strong correlation between the model predictions and people's judgments.

comparing all possible pairs of the following demonstrations. For the six-knob-toys, there were four kinds of demonstrations: 3+ 3- (show 3 working knobs and 3 inert knobs), 3+ (show 3 working knobs), 3- (show 3 inert knobs), 2+ 2- (show two working knobs and two inert knobs. For twenty-knob toys, there were five kinds of demonstrations: 3+3-, 3+, 3-, 2+2-, 10-, resulting in 36 questions total. In the Proportional condition, the pairs were generated from the following demonstrations 3+3-, 3+, 3-, 2+2- for the six-knob-toys and 3+3-, 3+, 3-, 2+2-, 10+, 10-, 7+7- and 10+10- for the twenty-knob toys, resulting in 66 total questions. To ensure that participants remembered the prior knowledge established during training, they were reminded of the average number of working knobs every ten questions.

The start of each demonstration was indicated by the appearance of a hand symbol pointing to a knob, which proceeded around the toy clockwise. The order of positive and negative examples and the locations of working knobs on the toy were determined randomly. After observing the demonstrations by the informant on the left, participants watched the informant on the right. After both sets of demonstrations, participants indicated which informant they judged as 'more helpful' using a slider that appeared in the middle, below the two toys. The order of pairs was randomized, as were the sides on which each informant appeared. After completing all of the questions for their condition, participants were debriefed and thanked.

Modeling. The prior knowledge, P(h), was set based on the demonstrations that participants observed. For simplicity, trials were assumed to be independent. For each condition, the number of working knobs (out of total number of knobs on the toy) were entered to a Beta-Binomial model with uniform parameters and the resulting distribution was the prior for the experimental judgments. This was performed separately for the six- and twenty-knob toys for each condition.

To find the parameter that best fits people's judgments, we performed a grid search over the values from -2 to 2 in increments of .02.

Results & Discussion

As an initial test of the model, we assessed the correlation between people's judgments and the model predictions for the 36 questions in the Consistent condition and the 66 questions in the Proportional condition. To do so, we fit the cost parameter separately to each of the two sets of data. The best-fitting value for the consistent condition was .02 and for the proportional condition was 0. These resulted in robust correlations between the model predictions and human judgments; they were r=.82 and r=.84 for the Consistent and Proportional conditions, respectively.

We can compare our model to a number of alternative proposals in which people's judgments might be explained by attention to more superficial aspects of the stimuli. Specifically, we investigated whether people's judgments were consistent with choosing based on the number of knobs on the toy, the number of positive examples demonstrated, the number of negative examples demonstrated, the number of knobs pressed, the percent of positive examples (out of the total knobs), and the percent of negative examples (out of the total number of knobs).

Our model provided significantly better fits to people's judgments than (a) number of knobs (Consistent: r=-.47, z=6.77, p<.0001; Proportional: r=-.47, z=9.72, p<.0001), (b) the number of knobs pressed (Consistent: r=.27, z=3.57, p<.001; Proportional: r=.58, z=3.14, p<.01), and (c) the number of negative examples demonstrated (Consistent: r=-.14, z=5.27, p<.0001; Proportional: r=.01, z=6.80, p<.0001), (d) the percent of negative examples (Consistent: r=.13, z=4.17, p<.0001; Proportional: r=.14, z=6.06, p<.0001).

However, the correlation between people's judgments and (e) the number of positive examples demonstrated were not significantly different from our model (Consistent: r=.66, z=1.48, p=.14; Proportional: r=.75, z=1.39, p=.16) and (f) the percent of positive examples (Consistent: r=.77, z=-.55, p=.58; Proportional: r=.88, z=-.87, p=.38).

To further investigate the degree to which our model and the remaining alternatives (number of positive examples and percent of positive examples) fit the data, we turn to analyses of individuals' judgments. For our model, we fit the parameter to each individual's judgments as described above. We correlated the predictions of our model, the number of positive examples, and the percent of positive examples with individual participants' judgments. Our model (Consistent: M=.52; Proportional: M=.57) predicted people's judgments better than the number of positive examples (Consistent: M=.29, t(42)=2.7, p<.01; Proportional: M=.38, t(42)=2.84, p<.01, by one-tailed t-test) and the percent of positive examples (Consistent: M=.33, t(42)=2.27, p<.05; Proportional: M=.46, t(42)=1.67, p=.05, by one-tailed t-test). These results suggest that our model provides a better explanation of people's behavior than these alternatives.

Next, we turn to the amount of data learners expect. Recall that expecting the informant to provide as much data as possible is indicated by parameter values much greater than 0, expecting the informant to provide as little data as possible is indicated by parameter values much less than 0, and expecting inductively sufficient data is indicated by parameter values near 0. For the group data, the best fitting parameters were 0 for the Consistent condition, and .02 for the Proportional condition. These parameters are most consistent with the hypothesis that learners expect data that suffices for accurate inference but exact no penalty for additional data.

The three different accounts make opposite qualitative predictions for subsets of the questions. To explore these differences we contrast the three hypotheses using parameter values of 2, -2, and 0 respectively.

The hypothesis that learners expect as much evidence as possible and the hypothesis that learners expect inductively sufficient evidence but are happy with more evidence make opposite predictions for five questions in both the Consistent and Proportional conditions. For example, in the Consistent condition, the preference for maximal evidence predicts that 2+2- should be preferred to 3+ (because there's a total of four demonstrations versus only three); whereas inductive sufficiency predicts the opposite preference.

With so few questions, it is not surprising that, although people's judgments tended toward a preference for inductive sufficiency, there were not statistically significant differences in participants' responses to these questions (Consistent: M=11.36, t(4)=1.81, p=.14; Proportional: M=4.66, t(4)=-.38, p=.72).

To separate the predictions of a preference for maximal data and inductive sufficiency, we identified the ten questions on which the predictions differed the most in each condition. We standardized the predictions of each model and chose the questions that had the largest absolute difference in predictions.

In the Consistent condition, the question with the largest difference compared a six-knob-toy with 3- and a twenty-knob-toy with 3+. Inductive sufficiency predicts a strong preference for the 3+ demonstration on the twenty-knob toy (because the learner is certain that three knobs work on the twenty-knob toy but uncertain whether 1, 2, or 3 knobs work the six-knob-toy) whereas a preference for maximal data predicts a strong preference for the 3- demonstration on the six-knob-toy (because the learner has evidence for half of the knobs on the six-knob-toy but only 3 of 20 for the twenty-knob-toy). That is, the learner might prefer the greater confidence afforded by the demonstration of the working knobs or might prefer a greater relative number of demonstrations (3- out of 6). Over the ten questions, there was a stronger correlation between inductive sufficiency and people's judgments, r=.85, than between the preference for maximal evidence and people's judgments, r=-.20, z=2.73, p<.01.

In the Proportional condition, the question with the largest difference compared 10+ versus 10- on twenty-knob-toys. Inductive sufficiency predicts a strong preference for 10+, whereas the preference for maximizing demonstrations predicts a slight preference for 10+. Over the 10 questions, there was a stronger correlation between inductive sufficiency and people's judgments, r=.90, than between maximizing demonstrations and people's judgments, r=.29, z=2.2, p<.05.

The hypothesis that learners expect as little evidence as possible and the hypothesis that learners expect inductively sufficient evidence make opposite predictions for 13 questions in the Consistent condition and 18 questions in the Proportional condition. For example, in the Consistent condition, a preference for minimal demonstrations predicts that 2+2- is preferred to 3+3- whereas inductive sufficiency predicts the opposite.

People's responses on these questions were coded as positive if consistent with the predictions of inductive sufficiency and negative if they were consistent with a preference for less data. People's judgments in the Consistent condition were in agreement with the predictions of inductive sufficiency both on average, M=32.6, t(12)=8.91, p<.0001, and in every individual case. Similarly, people's judgments in the Proportional condition were overwhelmingly in accord with the predictions of inductive sufficiency, M=34.63, t(17)=10.27, p<.0001.

General Discussion

We have proposed that learners' choice of informants is guided primarily by the degree to which evidence supports accurate inference. We presented a computational model that differentiates among the hypotheses that learners choose informants who provide as much data as possible, informants who minimize the amount of data provided, and informants who provide at least enough data to support accurate induction. The results show that people's behavior is best explained by inductive sufficiency.

Note that providing maximal data can, in simple cases, lead to deductive certainty. For finite, well-defined sets of possibilities (like those tested here), exhaustive demonstrations eliminate uncertainty. However, our results show that even on relatively small, well-defined learning problems, learners do not simply prefer informants who provide maximal amounts of data. In fact, people are just as likely to endorse much smaller sets of data, as long as the data provided suffices for accurate inductive inference. This suggests that learners are sensitive to the trade-off between the benefit of increased certainty from acquiring more data and the cost of acquiring more data; this sensitivity enables learners to decide how much data is 'enough'.

We did not find evidence for a simple preference for less data. A particularly interesting example is that people did not prefer someone who provides three working knobs over someone who provides exhaustive evidence, despite that the amount of data were twice as much in the latter. This may be due to features of our experimental design. There was relatively little reason to avoid additional demonstrations (among other things, each additional knob only took a second to press). Furthermore, our dependent measure asked learners to rate the helpfulness of the informant and learners have little reason to consider an exhaustive informant unhelpful. Finally, the additional demonstrations genuinely reduced some uncertainty: each toy was unique and the training only provided a few examples to establish the base rate of the effective knobs. Had the learners been more certain about the number of working knobs, they might have shown a stronger bias against exhaustive evidence.

We have presented evidence that learners' choice of informants is not solely guided the sheer amount of information; learners do not merely maximize the amount of data they can observe, nor do they minimize it. Learners use their inductive certainty to decide when enough is enough.

Acknowledgments

This research was partially supported by a James S. McDonnell foundation subaward to P.S.

References

Corriveau, K. H., Fusaro, M., & Harris, P. L. (2009). Going with the flow: Preschoolers prefer non-dissenters as informants. Psychological Science, 20, 372–377.

Corriveau, K. H., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and past accuracy. Developmental Science, 12, 426–437.

Gweon, H., Pelton, H., & Schulz, L. E. (2011). Adults and school-aged children accurately evaluate sins of omission in pedagogical contexts. In Proceedings of the 33rd annual conference of the cognitive science society.

Koenig, M., & Harris, P. (2005). Preschoolers mistrust ignorant and inaccurate speakers. Child Development, 76, 1261–1277.

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. Cognition, 112, 367–380.

Mischel, W., Shoda, Y., & Rodriguez, M. (1989). Delay of gratification in children. Science, 244, 933–938.

Pasquini, E. S., Corriveau, K. H., Koenig, M. A., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. Developmental Psychology, 43, 1216–1226.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012 Epistemic trust: Modeling children's reasoning about others' knowledge and intent. Developmental Science, 15, 436-447.

Shafto, P., & Goodman, N. D. (2008). Teaching games: Statistical sampling assumptions for pedagogical situations. In Proceedings of the 30th annual conference of the Cognitive Science Society.