More than true: Developmental changes in the use of inductive strength for selective trust

Asheley R. Landrum (LandrumAR@gmail.com)
Annenberg Center for Public Policy, University of Pennsylvania
Philadelphia, PA 19104 USA

Joshua Cloudy (Joshua.Cloudy@Louisville.edu)
Department of Psychological and Brain Sciences, University of Louisville
Louisville, KY 40292 USA

Patrick Shafto (patrick.shafto@gmail.com)
Department of Mathematics and Computer Science, Rutgers University – Newark
Newark, NJ 07102 USA

Abstract

When learning from others, it is important to take a critical stance—evaluating both the informants themselves as well as the content of their claims. In addition to accuracy, one can evaluate claims based on quality. The current study investigates developmental change in learners' evaluations of evidence that varies in quality—inductive strength based on typicality or diversity. We found that while younger children track which informant provides which examples, they do not have clear preferences for the informant who provides stronger examples. Older children, on the other hand, are in the middle of a developmental transition. They rate informants who provide inductively strong examples as more trustworthy, but only reliably choose the informant who provides diverse examples.

Keywords: selective trust; induction; social cognition; cognitive development

Introduction

Much of what we know we learn from other people. Although learning from others is often an efficient means of gathering information (e.g., Csibra & Gergely, 2009), it is not infallible. Depending on an informant's knowledgeability and intentions, information presented by an informant may vary from accurate and helpful to inaccurate or even misleading. Thus, it is important for learners to consider an informant's trustworthiness when drawing inferences from examples he or she has chosen (Landrum, Eaves, & Shafto, 2015; see also Sperber et al., 2010).

Research has demonstrated that learners as young as 4-years-old engage in this reciprocal process of trusting to learn and learning to trust. For instance, Koenig & Harris (2005) show that preschoolers track whether informants provide accurate labels for a series of common objects (e.g., "ball", "cup"), and later use this information to determine which of those informants to trust for providing labels for novel, unfamiliar objects (e.g, "wug", "loma"). In this paradigm, children are leveraging their prior knowledge of known object labels to learn whom to trust for new information. Having inferred that this informant is trustworthy, children will trust the informant to continue to provide accurate labels, even when the children are no longer able to confirm that this is the case (e.g., when the informant is labeling unfamiliar objects).

A major difficulty in real-life learning situations, however, is that information presented by informants is rarely as clearly right or wrong as in these labeling paradigms. In fact, information presented by informants can be accurate but misleading, causing learners to hold false beliefs. For instance, if a teacher demonstrates only one function of a multifunction toy, leaners could draw the conclusion that there is only one function (see Bonawitz, Shafto et al., 2011). Work by Gweon and colleagues, for example, demonstrates that even 6- and 7-year-old children recognize that informants must not only be correct but also provide an appropriate *quantity* of information. In the study, children rated under-informative teachers (i.e., teachers who demonstrated only 1 function of a 4-function toy) lower than informative teachers (i.e., teachers who demonstrated 1 function of a 1-function toy). In addition, the study showed that children compensated for under-informative teachers by exploring a new toy more broadly (Gweon, Pelton, Konopka, & Schulz, 2014). Empirical and modeling work with undergraduates shows that learners are willing to trust informants who omit some information, as long as the informants have provided enough information to support accurate inference (e.g., Shafto, Gweon, Fargen, & Schulz, 2012).

Building on this, it is also possible that learners consider the *quality* of information—such as the typicality or diversity of examples—when determining if informants are trustworthy. Voluminous research shows that children as young as 4 and 5 are sensitive to the quality of information when making inductive generalizations (e.g., Bjorklund, Thompson, & Ornstein, 1983; Gelman & Markman, 1986; Heit, 2000; Heit & Hahn, 2001; Osherson, Smith, Wilkie, Lopez & Shafir, 1990). Thus, it is possible that learners may be able to leverage these inductive abilities to infer informant credibility, even when informants provide neither

incorrect nor differing quantities of information. The current study aims to address this question.

In order to evaluate informant trustworthiness based on sample quality, a learner has to be able to differentiate between sets of examples based on inductive strength. Prior work by Rhodes, Brickman, and Gelman (2008) demonstrated a developmental shift in the ability to choose between samples varying in quality. In their study, participants were presented with a diverse and a non-diverse set of examples and were asked which of the two sets they would prefer to use to learn about a domain. Depending on condition, the diverse and non-diverse example sets either both included all typical or all atypical instances (i.e., all-typical and all-atypical conditions) or one included typical instances while the other included atypical instances (i.e., diverse-typical condition and diverse-atypical condition). Notably, children were not told about the origin of the examples, and therefore only evaluated sample composition—not informants. The authors found that whereas adults chose the diverse sample set across all conditions, 6-year-olds generally chose the sample set containing typical instances, and 9-year-olds chose the diverse samples more in the all-typical or all-atypical conditions than they did in the other two conditions. The authors suggest that children appreciate premise typicality at a young age, but the ability to recognize the benefit of diversity does not develop until later.

Current Study

The current study examines developmental variation in the use of example quality for *evaluating informants*. In the study, participants were presented with a situation in which they were interested in answering anatomical questions about several animal categories. In order to learn about these animals, they could examine example animals from the categories to study, similar to Rhodes et al. (2008). However the current study differs from Rhodes et al. (2008) in two important ways. First, the current study presents the examples as being selected by informants. Thus, whereas the Rhodes et al. (2008) study was focused on the composition of samples, the current study *additionally* investigated inferences that are made about *informants*. Following standard methods in the selective trust literature, participants were first asked whose examples they *endorse* and were then asked which informant they would prefer to *ask* to help them solve a new problem. The *endorse* trials provide a close analog to the methods of Rhodes et al.; participants were asked to choose between one set of examples and the other. The *ask* trials, in contrast, provide insight about how participants evaluate informants; participants were asked which of the two informants to ask without having provided examples to evaluate. We also assess participants' explicit perceptions of informant trustworthiness by asking them to rate each informant's knowledgeability and helpfulness. Prior empirical and modeling work suggests these are the two important dimensions of trustworthiness (e.g., Shafto, Eaves, Navarro, & Perfors, 2012). We average these two ratings together to get an estimate of perceived trustworthiness.

Second, we differ from Rhodes et al. (2008) in the way that typicality and diversity were presented. Whereas the Rhodes et al. study crossed the two factors, we isolated typicality and diversity, allowing us to investigate their development independently and in a format closer to prior work in the inductive generalization literature.

Our study had two goals. First, we aimed to investigate potential developmental changes in the ability to use example typicality and diversity for (a) choosing between informant claims and (b) trusting informants for new information. Second, we aimed to determine whether the provision of typical and diverse examples leads to informants being explicitly perceived as more trustworthy.

Method

Participants

Child-aged participants were recruited and tested at the Kentucky Science Center. Children were randomly assigned to participate in one of the two conditions: the *typical* condition or the *diverse* condition. Because we collected data from a wide age range, we divided the participants within each condition using a median split into younger children and older children. In the *typical* condition, 42 children participated: 21 younger children (M_{age}=6.47 years, SD=1.39) and 21 older children (M_{age}=10.93 years, SD=2.22). In the *diverse* condition, 46 children participated: 23 younger children (M_{age}=5.35 years, SD=1.29) and 23 older children (M_{age}=10.03 years, SD=1.83). In addition, 40 adults were tested via Amazon's Mechanical Turk (MTurk) and were each paid 50 cents for participation.

Procedure

As stated above, participants were randomly assigned to one of two between-subjects condition: the *typical* condition or the *diverse* condition. For both conditions, the experiment consisted of three sections: the endorse trials, the ask trials, and the explicit rating items, and was completed on a tablet computer using Qualtrics survey software. To begin, participants were introduced to the following prompt:

> *We are going to pretend that you are a scientist who is trying to learn about the insides of some animals. In order to learn these things, you need to pick good example animals to look at and learn about. Two people have agreed to help pick out example animals for you. For each question, your job is to decide who picked out the example animals that will help you learn best*[1].

Following the prompt, the participants completed the three sections of the experiment. The first section consisted of four *endorse* trials. For each trial, each informant presented a set of examples which participants were asked to choose between. In the typical condition, one informant

[1] The prompt wording was adjusted slightly for adults to be more age appropriate (e.g., "insides of some animals" was changed to "anatomical properties of certain animals.")

presented a typical example (i.e., the typical informant) and the other informant presented an atypical example (i.e., the atypical informant). For the category DOG, for instance, the typical informant selected a yellow labrador retriever, and the atypical informant selected a Chinese crested (see Figure 1). In the diverse condition, one informant presented diverse examples (i.e., the diverse informant) and the other informant provided non-diverse examples (i.e., non-diverse informant). For the category DOG, for instance, the diverse informant selected three different dog breeds, and the non-diverse informant selected three instances of one dog breed (see Figure 1). The order of the four endorse items was randomized between participants using the survey's item randomizing function. Which informant appeared on the right or left side of the screen (and which informant image was red or blue) was counterbalanced between participants. To help participants track the informants, the informants appeared on the same side of the screen for a given participant for the duration of the experiment. For the endorsement trials, we measured how many times participants endorsed the examples provided by the typical or diverse informant out of the four trials (i.e., number of endorsements).

Figure 1. Example item from the *typical* condition (top) and from the *diverse* condition (bottom).

Immediately after the endorsement trials, two check questions were asked to ensure that participants remembered which informant provided which examples. Which informant participants were asked about first (diverse/typical or non-diverse/atypical) was randomized between participants. Importantly, participants were provided with feedback to whether they were correct or incorrect and then were shown the examples provided by each informant as a reminder. In general, participants had little difficulty with the check items: 61% of younger children and 93% of older children got both check questions correct and 25% of younger children and 6% of older children got one right. Only 6 younger children answered no check questions correctly (5 were in the diverse condition). Again, as participants were reminded which informants provided which examples after answering the check questions, no participants were excluded from the analyses.

The second section of the experiment consisted of four *ask* trials in which participants needed to determine whom to ask for new example animals. For instance, participants were told, "Now you want to know if pigs have something inside called *kervicas*. Who do you want to ask for example pigs?" For these items, participants only saw the informant image (with no examples). The order of the four ask items was randomized between participants. The number of times the participants requested examples from the typical/diverse informant (i.e., number of requests) was measured. Note that participants did not receive any information from the informants during these trials.

The third and final section of the experiment asked participants to rate both informants on the two dimensions of trustworthiness: knowledgeability and helpfulness. For the knowledge rating items, participants were asked to rate each informant on a sliding scale from 0 to 10 (allowing for two decimals) how much they think each informant knows about animal insides (i.e., animal anatomy). For the helpfulness rating items, participants were asked to rate each informant on a sliding scale from 0 to 10 (allowing for two decimals) how helpful they thought each person's examples were. The order of the two rating items (knowledge and helpfulness) was randomized between participants, as was which informant they rated first (i.e., the typical/diverse informant or the atypical/non-diverse informant). The knowledgeability and helpfulness ratings for each informant were averaged together to create a *trustworthiness* rating score.

At the end of the experiment, children were awarded a certificate and a small toy.

Results

Example Typicality

We began by investigating developmental variation in the use of example *typicality*. In addition, we examined whether there were differences between participants' responses to the *endorse* items and the *ask* items. Then, we investigated whether participants in each age group preferred the typical informant by comparing their selections of the typical informant to chance. Finally, to determine whether participants of each age group saw the two informants as varying in trustworthiness, we examined the participant's trustworthiness rating scores of the two informants.

Developmental variation in the use of example typicality for endorsing and asking informants. First, to investigate developmental variation in the use of example typicality for endorsing and asking informants, we conducted a mixed-design ANOVA with item-type (endorse, ask) as a within-subjects factor and age group (younger children, older children, adults) as a between-subjects factor. We found a

main effect of item-type: participants generally *endorsed* (M=3.03, SD=0.92) the typical informant's examples more often than they *asked* (M=2.63, SD=1.19) him for new examples, $F(1, 59)=7.74$, p=.007, η^2=.116. This effect did not differ significantly by age (no item type by age group interaction), $F(2, 59)=0.42$, p=.66, η^2=.014. However, post-hoc comparisons of endorsing versus asking for each age group show that the difference between endorsing and asking was only significant for the youngest age group ($M_{endorse}$=2.81, SD=0.93, M_{ask}=2.24, SD=1.09, p=.036, d=.56). See Figure 2.

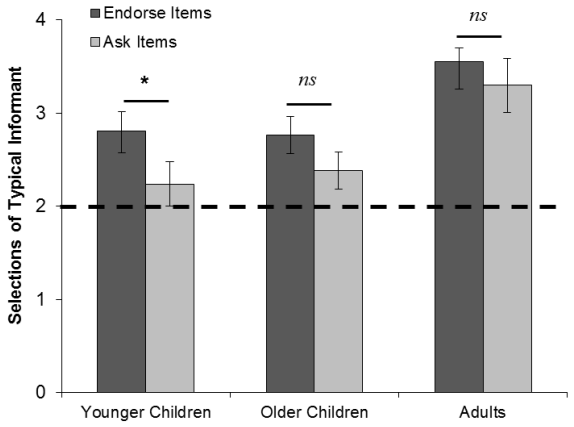

Figure 2. Participants' selections of the typical informant. Asterisks denotes differences between endorse and ask items. *p<.05

The ANOVA also revealed a developmental change in the use of example typicality, collapsed across item type (i.e., main effect of age group), $F(2, 59)=7.80$, p=.001, η^2=.209. Bonferroni corrected follow-up tests show that adults (M_{adults}=3.43, SD=0.89) selected the typical informant more frequently than both younger children ($M_{younger}$=2.52, SD=0.83, p=.002, d=1.09) and older children (M_{older}=2.57, SD=0.73, p=.004, d=1.09). Younger and older children did not vary significantly from one another (p=1.00, d=0.07).

Preferences for the typical informant. Comparing participants' preferences for the typical informant to chance, we find that all age groups *endorsed* the typical informant above chance levels (Younger: M=2.81, SD=0.93, $t(20)=4.00$, p=.001, d=1.79; Older: M=2.76, SD=0.94, $t(20)=3.7$, p=.001, d=1.65; Adult: M=3.55, SD=0.69, $t(20)=10.10$, p<.001, d=4.59). In contrast, only adults *asked* the typical informant above chance levels, while older children trended towards doing so (Younger: M=2.23, SD=1.09, $t(20)=1.00$, p=.329, d=0.45; Older: M=2.38, SD=0.92, $t(20)=1.90$, p=.072, d=0.85; Adult: M=3.30, SD=1.30, $t(19)=4.47$, p<.001, d=2.05).

Developmental variation in trustworthiness ratings. Next, to examine developmental differences in participants' ratings of informant trustworthiness, we conducted an ANOVA on informant trustworthiness rating scores where informant (typical, atypical) was a within-subjects variable and age group was a between-subjects variable. Importantly, we found a main effect of informant, such that participants rated the typical informant as more trustworthy (M=7.38, SD=2.22) than the atypical informant (M=5.58, SD=2.46), $F(1, 59)=14.72$, p<.001, η^2=.20. This effect did not vary by age (no informant by age group interaction), $F(2, 59)=1.65$, p=.202, η^2=.053. That said, post-hoc comparisons of the ratings for the typical and atypical informants for each age group show that this difference is significant for older children and adults, but not for the youngest age group (Younger: $t(20)=0.61$, p=.550, d=0.24; Older: $t(20)=3.08$, p=.006, d=0.92: Adult: $t(19)=4.14$, p=.001, d=1.41). See Figure 3.

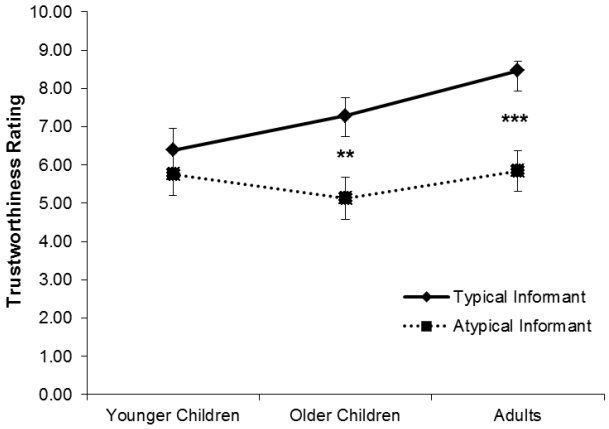

Figure 3. Trustworthiness ratings for the Typical and Atypical informants shown. Asterisks denote differences in perceived trustworthiness between the two informants. **p<.01, ***p<.001.

The ANOVA also revealed an overall main effect of age, $F(2, 59)=3.87$, p=.026, η^2=.116. Bonferroni-corrected post-hoc t-tests show that adults (M_{adults}=7.16, SD=1.19) gave the informants higher ratings than younger children ($M_{younger}$=6.07, SD=2.60, p=.039, d=0.55) and trended towards doing so compared to older children (M_{older}=6.21, SD=1.71, p=.087, d=0.66). Younger children's and older children's ratings did not differ from one another significantly (p=1.00, d=0.07).

Summary. We find developmental variation in the use of example typicality for making inferences about informants. While even the youngest children are able to endorse typical versus atypical examples provided by informants, there were differences in whether participants used this information to determine whom to request future examples from. Adults preferred to request for examples from an informant who had previously provided typical examples, and older children trended towards doing so. Younger children, however, showed no preference. These findings were supported in the trustworthiness ratings data; older children and adults saw the typical informant as more

trustworthy than the atypical informant, but younger children did not.

Example Diversity

The analysis for the diverse condition was the same as the analysis for the typical condition except where otherwise noted.

Developmental variation in the use of example diversity for endorsing and asking informants. Unlike for example typicality, there was no main effect of item type; participants did not *endorse* (M=2.92, SD=1.13) the diverse informant more often than they *asked* (M=2.91, SD=1.12) that informant, $F(1, 64)=0.02$, $p=.093$, $\eta^2<.001$. Also, there was not an age by item interaction, $F(2, 64)=1.41$, $p=.252$, $\eta^2=.042$. There was, however, a main effect of age, $F(2, 64)=23.94$, $p<.001$, $\eta^2=.428$. Bonferroni-corrected comparisons show that adults (M_{adult}=3.79, SD=0.46) selected the diverse informant more often than both younger ($M_{younger}$=2.22, SD=0.39, $p<.001$, d=3.78) and older children (M_{older}=2.74, SD=1.16, $p<.001$, d=1.20). See Figure 4.

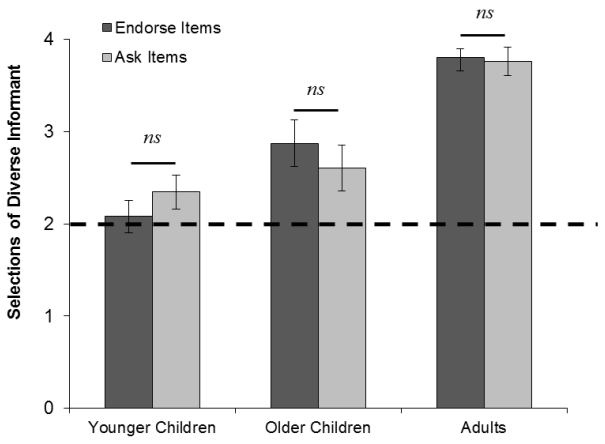

Figure 4. Participants' selection of the diverse informant.

Preferences for the diverse informant. Comparing participants' preferences for the diverse informant to chance, we find that only older children and adults *endorsed* the diverse informant above chance levels (Younger: M=2.08, SD=0.79, $t(22)=0.53$, $p=.604$, $d=.22$; Older: M=2.87, SD=1.25, $t(22)=3.33$, $p=.003$, d=1.42; Adult: M=3.80, SD=0.40, $t(20)=20.61$, $p<.001$, d=9.22). Similarly, older children and adults *asked* the diverse informant above chance levels, and the younger children trended towards doing so (Younger: M=2.35, SD=0.89, $t(22)=1.89$, $p=.073$, $d=.80$; Older: M=2.61, SD=1.20, $t(22)=2.44$, $p=.023$, d=1.40; Adult: M=3.76, SD=0.70, $t(20)=11.53$, $p<.001$, d=5.16).

Developmental variation in trustworthiness ratings. Next, to examine developmental differences in participants' ratings of trustworthiness, we conducted an ANOVA on informant trustworthiness rating scores where informant (typical, atypical) was a within-subjects variable and age group was a between-subjects variable. Importantly, like with the *typical* condition, we found a main effect of informant, such that participants rated the diverse informant as more trustworthy (M=7.68, SD=2.13) than the non-diverse informant (M=4.97, SD=2.63), $F(1, 64)=37.15$, $p<.001$, $\eta^2=.367$. Unlike the *typical* condition, however, this effect *did* vary by age, $F(2, 64)=6.34$, $p=.003$, $\eta^2=.165$. Bonferroni-corrected comparisons suggest that the difference between the trustworthiness ratings for the two informants was smaller for the younger children ($MD_{younger}$=0.51) than for both older children (MD_{older}=4.00, $p=.005$) and adults (MD_{adult}=3.62, $p=.018$). Older children's and adults' ratings did not differ significantly from one another (p=1.00). Moreover, older children and adults saw the diverse informant as significantly more trustworthy than the non-diverse informant, but younger children did not (Younger: $t(22)=0.55$, $p=.586$, $d=.19$; Older: $t(22)=5.539$, $p<.001$, d=1.90; Adult: $t(20)=5.88$, $p<.001$, d=1.73). See Figure 5.

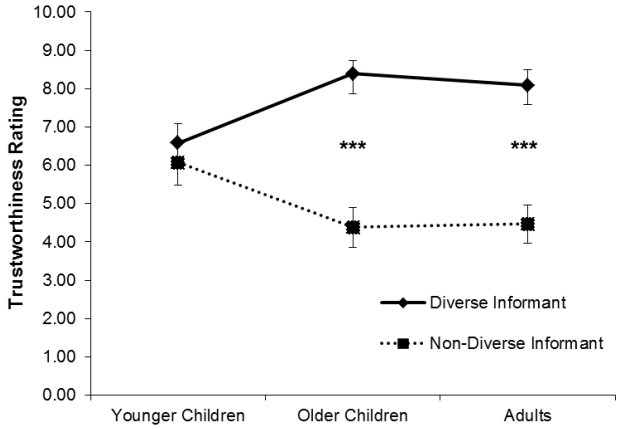

Figure 5. Trustworthiness ratings for the Diverse and Non-Diverse informants shown. Asterisks denote differences in perceived trustworthiness between the two informants. ***p<.001.

Summary. Unlike for the typical condition, participants in the diverse condition *endorsed* the diverse informant just as often as they *asked* that informant. There were, however, overall age differences in whether or not participants preferred the diverse informant: older children and adults preferred the diverse informant above chance levels, and younger children did not. Like in the typical condition, these findings were supported in how participants rated the informants; older children and adults saw the diverse informant as more trustworthy than the non-diverse informant, but younger children did not.

Discussion

Researchers suggest that tracking informant trustworthiness is important, but most of the research to date has focused on cases in which tracking trustworthiness is arguably least important—when you already know the answer. In most real-life learning situations, learners are not likely to have the answers, and thus must use evaluative methods of trustworthiness other than assessing veracity.

Recent research investigating the tracking of information *quantity* provides a good start. This work shows that even young children recognize when an informant is omitting information that may lead to false beliefs (e.g., Gweon et al., 2014). However, even in these paradigms, children have been able to rely on their prior knowledge and experience to recognize when someone is being truthful, but uninformative. So, how can a learner evaluate the content of information when he or she lacks knowledge? Our research demonstrates that learners can use inferences about the *quality* of information. In these instances, the inferences learners make about trustworthiness are bootstrapped off more general background knowledge about a domain as opposed specific knowledge about the correctness of the evidence provided. Thus, the current study provides an initial indication of how children might infer who to trust in more general cases.

Our results show that, whereas young children are not reliably using example typicality and diversity when determining whom to trust, older children seem to be in the middle of a developmental transition. They (either above chance or trending toward) prefer the informant providing stronger examples, and they explicitly rate those informants as more trustworthy. Thus, we find a similar developmental pattern to Rhodes et al (2008) where around 9 years of age, children are beginning to consider the benefits of typical and diverse examples. However, our work takes a step further—demonstrating that in addition to recognizing the benefits of typical and diverse examples, children around 9 years of age can leverage example quality to evaluate informants' trustworthiness.

Acknowledgments

The authors would like to thank the Kentucky Science Center and the Computational Cognitive Science Lab and CoCo Kids Lab team members at the University of Louisville. This research was supported in part by National Science Foundation CAREER grant DRL-1149116 to PS.

References

Bjorklund, D. F., Thompson, B. E., & Ornstein, P. A., (1983). Developmental trends in children's typicality judgments. *Behavior Research Methods & Instrumentation, 15*(3), 350-356. Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy. *Cognition, 120*(3). 322-330.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. Trends in Cognitive Science, 13, 148-153.

Corriveau, K. H., Fusaro, M., & Harris, P. L. (2009). Going with the flow: Preschoolers prefer nondissenters as informants. *Psychological Science, 20*(3), 372-377.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*(3), 183-209.

Gweon, H., Pelton, H., Konopka, J. A., & Schulz, L. E. (2014). Sins of omission: Children selectively explore when teachers are under-informative. *Cognition, 132*(3), 335-341.

Gweon, H., Shafto, P., & Schulz, L.E. (2014). Children consider prior knowledge and the cost of information both in learning from and teaching others. *Proceedings of the 36th annual conference of the Cognitive Science Society*

Heit, E. (2000). Properties of inductive reasoning. *Psychonomical Bulletin & Review, 7*(4), 569-592.

Heit, E., & Hahn, U. (2001). Diversity-based reasoning in children. *Cognitive Psychology, 43*(4), 243-273. doi:10.1006/cogp.2001.0757

Kinzler, K. D., Corriveau, K. H., & Harris, P. L. (2011). Children's selective trust in native-accented speakers. *Developmental Science, 14*(1), 106-111.

Koenig, M. A., & Harris, P. A. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, *76*(6), 1261–1277.

Landrum, A. R., Eaves, B. S., Jr., & Shafto, P. (2015). Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Science*.

Osherson, D. N., Smith, E. W., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*(2), 185-200.

Rhodes, M., Brickman, D., & Gelman, S. A. (2008). Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. *Cognition, 108*(2), 543-556.

Shafto, P., Gweon, H., Fargen, C., & Schulz, L. (2012). Enough is enough: Inductive sufficiency guides learners' ratings of informant helpfulness. *Proceeding of the 34th annual conference of the Cognitive Science Society*.

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D., (2010). Epistemic Vigilance. *Mind & Language, 25*(4), 359-393.

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science, 10*(3), 288-297.