Cognition 222 (2022) 104999

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

The effects of information utility and teachers' knowledge on evaluations of under-informative pedagogy across development

Ilona Bass [a,*], Elizabeth Bonawitz [b], Daniel Hawthorne-Madell [c], Wai Keen Vong [d], Noah D. Goodman [c], Hyowon Gweon [c]

[a] Department of Psychology, Harvard University, Cambridge, MA 02138, United States
[b] Graduate School of Education, Harvard University, Cambridge, MA 02138, United States
[c] Department of Psychology, Stanford University, Stanford, CA 94305, United States
[d] Center for Data Science, New York University, New York, NY 10011, United States

ARTICLE INFO

Keywords:
Pedagogy
Bayesian models
Social evaluation
Causal learning

ABSTRACT

Teaching is a powerful way to transmit knowledge, but with this power comes a hazard: When teachers fail to select the best set of evidence for the learner, learners can be misled to draw inaccurate inferences. Evaluating others' failures as teachers, however, is a nontrivial problem; people may fail to be informative for different reasons, and not all failures are equally blameworthy. How do learners evaluate the quality of teachers, and what factors influence such evaluations? Here, we present a Bayesian model of teacher evaluation that considers the utility of a teacher's pedagogical sampling given their prior knowledge. In Experiment 1 ($N = 1168$), we test the model predictions against adults' evaluations of a teacher who demonstrated all or a subset of the functions on a novel device. Consistent with the model predictions, participants' ratings integrated information about the number of functions taught, their values, as well as how much the teacher knew. Using a modified paradigm for children, Experiments 2 ($N = 48$) and 3 ($N = 40$) found that preschool-aged children (2a, 3) and adults (2b) make nuanced judgments of teacher quality that are well predicted by the model. However, after an unsuccessful attempt to replicate the results with preschoolers (Experiment 4, $N = 24$), in Experiment 5 ($N = 24$) we further investigate the development of teacher evaluation in a sample of seven- and eight-year-olds. These older children successfully distinguished teachers based on the amount and value of what was demonstrated, and their ability to evaluate omissions relative to the teacher's knowledge state was related to their tendency to spontaneously reference the teacher's knowledge when explaining their evaluations. In sum, our work illustrates how the human ability to learn from others supports not just learning about the world but also learning about the teachers themselves. By reasoning about others' informativeness, learners can evaluate others' teaching and make better learning decisions.

1. Introduction

Receiving instruction from others, as in both formal and informal pedagogy, is a powerful way to learn about the world. Rather than having to go through trial-and-error or discover new knowledge solely from exploration and observation, learners can rely on knowledgeable, well-intentioned individuals to teach them about the world. By learning from teachers, learners not only learn more effectively and efficiently, but also learn things that would be too risky or even impossible to learn by themselves (Bridgers, Jara-Ettinger, & Gweon, 2020; Gweon, 2021). Indeed, pedagogical practices are widespread across human societies (Hewlett, Fouts, Boyette, & Hewlett, 2011). Even infants are sensitive to pedagogically communicated information from adults (Csibra & Gergely, 2009; Geraghty, Waxman, & Gelman, 2014), and young children readily draw inferences that go beyond the face value of evidence in pedagogical contexts (Bonawitz et al., 2011).

Despite its power, however, learning from pedagogy comes with a potential hazard: Not everyone is equally knowledgeable and helpful. They may provide inaccurate, insufficient, or even deliberately misleading information for learners. Thus, the ability to evaluate others'

* Corresponding author at: William James Hall, Room 1310, 33 Kirkland St., Cambridge, MA 02138, United States.
E-mail addresses: ibass@fas.harvard.edu (I. Bass), elizabeth_bonawitz@gse.harvard.edu (E. Bonawitz), d.j.hawthorne@alumni.stanford.edu (D. Hawthorne-Madell), waikeen.vong@nyu.edu (W.K. Vong), ngoodman@stanford.edu (N.D. Goodman), hyo@stanford.edu (H. Gweon).

https://doi.org/10.1016/j.cognition.2021.104999
Received 7 March 2021; Received in revised form 12 November 2021; Accepted 22 December 2021
Available online 12 January 2022

quality as teachers is a critically important skill for effective social learning. By being able to recognize and evaluate teachers who failed to provide informative evidence, learners can avoid learning from them and protect themselves from ineffective pedagogy. While there is a host of prior developmental work showing that even young children are sensitive to inaccurate informants and selectively trust accurate informants (Birch, Vauthier, & Bloom, 2008; Corriveau & Harris, 2009; Jaswal & Neely, 2006; Koenig, Clément, & Harris, 2004; Pasquini, Corriveau, Koenig, & Harris, 2007; see also Harris, Koenig, Corriveau, & Jaswal, 2018; Tong, Wang, & Danovitch, 2020), there are many other ways in which a teacher can be unhelpful beyond being inaccurate. In fact, in real-world contexts, teachers may fail to be informative in more subtle but nonetheless misleading ways. Such variations in teacher quality raise questions about not just whether, but also how, learners might make fine-grained evaluations of others' teaching, and what factors might influence these evaluations. The current work has three key goals: We (1) present a computational account of teacher evaluation that considers the utility of demonstrated evidence given the teacher's prior knowledge; (2) provide empirical support for the model from experiments with adults; and (3) explore the development of teacher evaluations by studying four- to eight-year-old children.

Definitions of the term "pedagogy" vary broadly across fields of study (Kline, 2015). In this paper, "pedagogy" broadly refers to communicative acts motivated by the goal of helping a learner acquire some useful knowledge (Gweon, 2021; Shafto, Goodman, & Griffiths, 2014; see also Goodman & Frank, 2016). In such communicative contexts, a *teacher* selectively generates useful evidence for a *learner* who, in turn, draws inferences accordingly. This way of characterizing pedagogical communication is grounded in past work on models of pedagogical reasoning (Shafto et al., 2014; Shafto, Goodman, & Frank, 2012), particularly as they pertain to cognitive development (Bonawitz & Shafto, 2016). These computational models formalize reasoning in pedagogical contexts as a set of mutually constraining inferences: Given a space of possible hypotheses, the teacher selects evidence in a way that maximizes the learner's belief in the target (correct) hypothesis (i.e., they engage in *pedagogical sampling*), and the learner rationally updates their beliefs with the assumption that the information has been sampled pedagogically. This leads to an expectation that information provided by the teacher is not only true, but also sufficient for the learner to draw accurate inferences.

The consequences of such reasoning have often been investigated in informal, dyadic communicative contexts where the teacher's knowledge and intent is made explicit. For example, imagine someone who presents a complex-looking toy and declares, "I know all about this toy, let me show you how it works!" and then demonstrates that pulling out a tube on the toy plays music. Prior work has shown that given this kind of pedagogical demonstration about one function of a novel toy, preschool-aged children (Bonawitz et al., 2011; Yu, Landrum, Bonawitz, & Shafto, 2018) and even two-year-old toddlers (Shneidman, Gweon, Schulz, & Woodward, 2016) infer that the toy has no other relevant functions to be discovered: When given the chance to play with the toy themselves, children tend to focus on playing with the demonstrated function and explore the toy less broadly than children who observed the same function in non-pedagogical contexts (e.g., an adult who claims to be ignorant of the toy and accidentally discovers the function). Given that the teacher never explicitly denied the existence of other functions, why would children draw this inference? Intuitively, one might say that if the toy had additional functions, the teacher – assumed to be knowledgeable about the toy and with an intent to teach the learner – would have demonstrated those, too. Computational models of pedagogical reasoning provide a formal explanation for this intuition: When a teacher engages in pedagogical sampling, the learner infers that the most likely hypothesis given a space of possible hypotheses (say, about the existence of causal functions of a toy) is the one that is constrained to the demonstration provided by the teacher.[1]

Children in these past studies were, in fact, *misled* by the teacher, because the toy actually had multiple additional functions that were just as interesting as the demonstrated one. Thus, children who inferred that the toy had no other functions failed to discover all of its functions when they had an opportunity to explore it on their own. Importantly, children's inferences were reasonable given the pedagogical context; rather, it was the *teacher* who violated the learner's expectation by failing to demonstrate all of the functions, thereby committing a "sin of omission" (Gweon, Pelton, Konopka, & Schulz, 2014): By omitting demonstrations of other relevant functions, the teacher misled the learner to form a false belief about the toy (namely, that the toy had no other functions), even though the demonstration that *was* provided was still true of that toy. Evaluating such insufficient or under-informative pedagogy is not a trivial feat, because it involves more than a sensitivity to inaccurate information; instead, learners must consider *how* the teacher has selected evidence, and how that evidence would shape the learner's resulting beliefs.

More recent work has found that even preschool-aged children can detect and evaluate sins of omission (Gweon & Asaba, 2018; Gweon et al., 2014; see also Gweon, 2021). In these studies, given a teacher who demonstrated a single function of a device, children rated the teacher lower when they knew that the device had additional undemonstrated functions (i.e., the teacher committed a sin of omission) than when they knew that the demonstrated function was the device's only functional affordance. Furthermore, children adjusted their future learning from the teacher depending on that teacher's past history of omission: Given a demonstration on a novel toy that they had never seen before, children engaged in more compensatory exploration when the teacher had previously committed a sin of omission (Gweon et al., 2014). These results suggest that by the end of their preschool years, children already have an abstract understanding that "helpful teaching" is not just reflected in simple accuracy, but also requires providing the *best* set of evidence for the learner to be able to infer the correct hypothesis.

This past work, however, raises an important question: Should all omissions in pedagogical contexts be considered "sins"? Intuitively, not all omissions are equally blameworthy. In real-world contexts, teachers and caregivers may omit certain information from learners for various

[1] In this example and in the current study, we describe the hypothesis space in terms of the number of functions relevant to the learner (e.g., the learner infers that there are no additional functions to explore). However, the constraint in the model is agnostic to the hypothesis space, and pertains simply to the amount of information left to learn. Thus, the learner might be inferring that there is nothing else relevant to learn or worth exploring, nothing else that exists, or even nothing else that the child is permitted to act on (Bass, Shafto, & Bonawitz, 2018). The same reasoning applies to the empirical work cited here (Bonawitz et al., 2011; Shneidman et al., 2016; Yu et al., 2018), as all accounts are consistent with constraints on exploration given assumptions about informative utility. This is in line with characterizations of computational models of pedagogical reasoning (Shafto & Goodman, 2008).

reasons, and focusing on some information over others may even be considered beneficial in some cases.[2] For instance, exhaustively demonstrating all 20 buttons on a toy can be deemed over-informative and undesirable if demonstrating just a few buttons and omitting the rest is enough for the learner to learn how the toy works (Gweon, Shafto, & Schulz, 2018). Thus, it is possible that acts of omission may be evaluated differently depending on what and how much was omitted, and why. Yet, we still lack an integrated model that explains what factors influence these kinds of pedagogical evaluations. Although some prior proposals have characterized children's selective epistemic trust as a form of rational inference (Sobel & Kushnir, 2013; for formal models, see: Eaves & Shafto, 2017; Shafto, Eaves, Navarro, & Perfors, 2012), such work has focused primarily on the evaluation of informants who provide accurate or inaccurate evidence (e.g., a teacher who labels a cup a "cup" versus a teacher who labels it a "cow").

The current work aims to fill this theoretical gap by developing a cognitive model of teacher evaluation that formalizes the ways in which people assess how well a teacher sampled evidence for a learner, and to provide empirical support for this model. As we will describe below, our model assumes an ideal observer who can represent and reason about others' mental states, and draw inferences accordingly. Testing adult participants – whose reasoning in these domains is mature – will be a critical test of our model's predictions. Because these capacities may be still developing in young children, comparing their teacher evaluations against model predictions can provide additional insights about underlying cognitive mechanisms. As we discuss in the following three sections, prior developmental work collectively suggests that the ability to produce the sophisticated teacher evaluations predicted by our model may appear as early as the preschool years, and continue to develop throughout childhood. Thus, we start our developmental investigations with preschool-aged participants, and in subsequent experiments we extend our tested age range to seven- and eight-year-olds. Indeed, the preschool years represent a prime transitional age to begin assessing the ability to evaluate teacher quality, because preschoolers have not yet entered formal education (though many have had some initial exposure to structured learning environments through daycare settings), which may begin to shift children's expectations and assumptions coming into pedagogical interactions.

In what follows, we consider three factors that may influence people's evaluations of teachers who omit relevant evidence: the amount of omitted information, the value of omitted information, and the knowledge state of the teacher. In each of these sections, we also consider prior developmental work to motivate our predictions about each of these three factors.

1.1. Amount of information

First, the *amount* of omitted information may directly influence the degree to which we penalize sins of omission. This intuition can be illustrated by considering the following scenario: Imagine a novel device that has four different buttons on it; an adult demonstrates one of these buttons to a child. Contrast this with someone who instead shows the child three of the four buttons. Intuitively, while both of these teachers provided only partial information, the former teacher will likely be considered "less helpful" than the latter. Although providing partial information may generally be considered as undesirable, omitting more information would result in the learner inferring a hypothesis that is further from the truth, and ultimately acquiring less knowledge about the device.

As discussed above, past research has found that young children can evaluate sins of omission, distinguishing informants who provide complete evidence from those who provide partial evidence (Gweon & Asaba, 2018; Gweon et al., 2014). However, it remains unclear whether children are sensitive to the degree of omission, above and beyond simply detecting that an omission has taken place. Given two teachers, both of whom failed to show all the functions of a device, do children evaluate them differently depending on how much information they omitted? One recent study suggests that children understand that not all omissions are equally blameworthy: By 5 years of age, children consider "partial" demonstration of a toy (i.e., pressing 3 buttons on a 20-button toy) as sufficiently informative insofar as the learner has prior knowledge that supports accurate inferences from this partial evidence (Gweon et al., 2018). This provides indirect support for the possibility that five-year-old children are able to distinguish the relative informativeness of two teachers, even when neither presents complete evidence. It is also worth noting that outside of pedagogical contexts, children's ability to recognize effective questions and explanations develops throughout early childhood, possibly until age 9 or 10 (Jirout & Klahr, 2020; Mills, Sands, Rowles, & Campbell, 2019). Thus, meta-cognitive reasoning about what information gaps are, and how best to fill them (Loewenstein, 1994) – arguably a critical component of evaluating more subtle violations of pedagogical sampling assumptions – could still be developing in the preschool years.

1.2. Value of information

Second, people might also consider the *value* of the omitted information. An emerging body of work suggests that humans, even early in life, reason about others' actions in terms of their costs and rewards (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Liu, Ullman, Tenenbaum, & Spelke, 2017), and communicate in ways that maximize the listener's utilities (Goodman & Frank, 2016). Because not all knowledge is equally useful or valuable, omission can have varying consequences for the learner depending on the utility of what was omitted. Appealing to the same novel device example described above, we might be more likely to penalize a teacher who failed to demonstrate a highly useful function (e.g., a button that reports the weather) than a teacher who omitted a function that is relatively useless (e.g., a button that plays white noise). Further, what exactly should be considered "valuable" for the purposes of teaching could reasonably shift with development. We generally assume that if learners themselves consider information to be relatively more interesting or useful, they should also think that information will be relatively more valuable to teach.

Prior work suggests that even infants and toddlers expect agents to pursue rewarding goals (Liu & Spelke, 2017) and assign different subjective rewards to items depending on their preferences (Repacholi & Gopnik, 1997). However, an abstract understanding of teaching as an act that provides valuable information for others may not emerge until later in life. Consistent with this possibility, recent work has shown that 5- to 7-year-old children decide to teach information that maximizes others' utilities: Given a choice between two toys, they choose to teach what maximizes the learner's expected rewards and minimizes the learner's expected cost of learning (Bridgers et al., 2020). These findings are consistent with the prediction that children's evaluations of pedagogy also reflect the value of omitted information, but these abilities may emerge relatively late in the preschool years.

[2] Note that the very notion of pedagogy implies that a teacher is focused on teaching about a certain topic; this question of how teachers choose the topic of instruction (and how learners might infer the topic and represent the relevant hypothesis space) is not directly addressed here, but see Bridgers et al. (2020), as well as related work on pragmatic inference (e.g., Goodman & Frank, 2016 on how listeners infer the "question under discussion" in communicative contexts). We distinguish omission from the selective nature of teaching more broadly by studying a case in which there is already a set of clear hypotheses about what is being taught (e.g., four causal functions of a novel toy), all of which could be discovered and understood by the learner within the constraints of the learning context, such that a teacher's failure to demonstrate all of its functions clearly violate the pedagogical sampling assumption.

1.3. Teachers' knowledge

Third, even identical acts of omission (i.e., omitting the same number of functions of the same value) intuitively seem as though they are less blameworthy if the teacher did not actually *know* about the omitted function (and therefore could not demonstrate it) than when the teacher knowingly omitted a function. More formally, if the omitted information is not within the teacher's hypothesis space, it is impossible for the teacher to (intentionally) provide that information. Thus, while the teacher may be evaluated negatively for her ignorance, she may be (at least partially) forgiven for her omission. Such nuanced "exoneration" of sins of omission would require the learner to evaluate a teacher's pedagogical sampling with respect to the teacher's limited knowledge, which might involve a relatively mature Theory of Mind (Bass et al., 2019; Bonawitz, Shafto, Yu, Gonzalez, & Bridgers, 2020). Characterizing how this particular sensitivity develops may provide broader insight into social reasoning and its role in cognitive development.

Whether young children can consider teachers' knowledge states to exonerate "innocent" omissions is also an open question. Past work suggests that children's Theory of Mind reasoning is related to their understanding of intentionality in teaching (Ziv & Frye, 2004; Ziv, Solomon, Strauss, & Frye, 2016); and more recently, children's understanding of what constitutes "good" pedagogical evidence selections has been shown to be *causally* related to Theory of Mind reasoning (Bass et al., 2019). Theory of Mind undergoes significant developmental change during the preschool years (Wellman, Cross, & Watson, 2001). Furthermore, children under 9 years of age may have trouble understanding that others' lack of knowledge limits the testimony they're able to provide (Kominsky, Langthorne, & Keil, 2016; see also Kushnir & Koenig, 2017). Prior work on children's moral evaluation of unintended harm also provides some insights here: The ability to exonerate accidental harm caused by the agent's ignorance or false beliefs appears to be continuously developing through at least age eight (Cushman, Sheketoff, Wharton, & Carey, 2013; Nelson, 1980; Yuill & Perner, 1988), and may also rely on Theory of Mind development (Killen, Mulvey, Richardson, Jampol, & Woodward, 2011). Evaluating a teacher's omission based on whether the teacher *knowingly* omitted may present a similar challenge for young children. Thus, while preschoolers might be capable of evaluating omissions in general, doing so in light of what a teacher *knows* may be difficult until later in childhood.

It is important to note that these intuitions may reflect only a fraction of the various considerations that people make in real-world teacher evaluations. For example, as noted above, choosing to focus on some pieces of information over others is not an unreasonable tactic in real-world pedagogical interactions, and online inferences about information focus could be just one of many other ways in which learners may explain away omissions (Goodman & Frank, 2016). However, we focus on these three factors because the aforementioned intuitions about teacher evaluation may emerge from an integrated inferential process rather than a set of individual heuristics, and a model that evaluates teaching based on the utility of the information a teacher provides for a learner given her knowledge state should naturally produce these predictions. Thus, the goal of our modeling approach is to develop a cognitive model of teacher evaluation that can generate predictions that are consistent with all of these intuitions. Additionally, although we propose a single model, there may be a variety of component cognitive capacities and processing demands required to evaluate teachers in line with our model's predictions – which, as discussed above, could be elucidated by testing our model predictions against both adult and child participants.

Here, we present a Bayesian computational model that explains how evaluations of teachers may integrate considerations of number, value, and knowledge state, and test its predictions in a series of behavioral experiments with both adults and children. The model is an extension of the pedagogical model (first presented by Shafto & Goodman, 2008; see also Bonawitz & Shafto, 2016; Shafto et al., 2014; Shafto, Goodman, et al., 2012), and captures the idea that the more information a teacher shares relative to what they know, and the higher-value that information, the better they should be evaluated. In general, the behavioral experiments present participants with teaching scenarios in which a teacher finds a device with four functions (two high-value, two low-value), and discovers either all four or a subset of its functions. The teacher then shows a learner either all or some subset of the functions they initially discovered, and participants are asked to evaluate their quality. In Experiment 1, we test the model's predictions in detail with a large sample of adults. In Experiment 2, we test a subset of the model's predictions using an adapted method in a sample of preschool-aged children (2a) and adults (2b). Experiment 3 further investigates the role of information value in children's teacher evaluations. Experiment 4 attempts to replicate the effects of Experiment 2a with a second sample of preschoolers. Finally, Experiment 5 investigates the development of these evaluation abilities in a sample of seven- and eight-year-olds.

2. Computational model

Computational models have long been used to provide formal frameworks for understanding how learners might update their beliefs given some observed data in the world. By expressing this learning in the form of Bayesian inference, and appealing to existing models of pedagogical learning, a theory-based Bayesian approach (Tenenbaum, Griffiths, & Kemp, 2006) allows us to understand how learners can make inferences from noisy data provided by others, while simultaneously reasoning about the person presenting those data.

Prior computational models of pedagogical reasoning (Shafto & Goodman, 2008) provide an account of how recursive mental-state reasoning between a teacher and a learner results in the teacher's pedagogical sampling of evidence (i.e., selecting the set of evidence that would maximally increase the learner's belief in the target hypothesis):

$$p(d|h)_{teacher} \propto (p(h|d)_{learner})^{\alpha} \tag{1}$$

where α controls the degree to which a teacher will select useful examples. Learners, in turn, update their beliefs following Bayesian inference, under the assumption that the data have been pedagogically sampled by a teacher who knows the target hypothesis:

$$p(h|d)_{learner} \propto p(d|h)_{teacher}\ \ p(h) \tag{2}$$

Thus, pedagogy can be understood as a set of recursive, mutually dependent inferences: The teacher pedagogically samples evidence that maximizes the learner's belief in the target hypothesis; and the learner rationally updates their belief in that hypothesis with the assumption that the evidence has been sampled pedagogically (see also Shafto et al., 2014; Shafto, Goodman, et al., 2012).

This theory can be extended to explain not just how learners learn *from* teachers (as in Shafto & Goodman, 2008), but also in learning *about* the teachers themselves. In particular, a teacher's quality may be reflected in the importance they place on informing the learner, and in their ability to do so by choosing to present evidence that would be most useful to the learner. To formalize such a theory, we draw both from existing models of pedagogy, and from models of pragmatic inference (e.g., Goodman & Frank, 2016), which suggest that rational communicators should maximize the utility of their speech given their own prior knowledge and the true state of the world. Similarly, we describe how a teacher might choose what demonstrations to give in different situations with a standard softmax decision rule that chooses in proportion to the options' probabilities:

$$p(d|\ f_T, \alpha) \propto e^{\alpha\ \ U(d)-C(d)} \tag{3}$$

where f_T is the set of functions known to the teacher (with $f_{T_i} = 1$ if the teacher knows the i^{th} feature), and d is the set of features demonstrated to the learner (with $d_i = 1$ if the teacher presents the i^{th} feature). $C(d)$,

which is one of our model's two free parameters, may range from 0 to 1 and corresponds to the *communication cost*, or the cost of demonstrating a given set of functions. $U(d)$ is the utility a learner is expected to accrue from those demonstrations. The parameter α controls the degree to which the teacher chooses to maximize the pedagogical utility of their demonstration. This α parameter is analogous to that in the original pedagogical model and can be interpreted as the teacher's quality, with respect to their informativeness: An uninformative (and thus "low-quality") teacher ($\alpha = 0$) would engage in weak sampling, choosing demonstrations at random with a preference for less costly demonstrations; as α increases, a teacher would be more likely to select demonstrations that have high pedagogical utility, as informative ("high-quality") teachers should.

To formalize how a learner might evaluate the teacher, we must specify the utility of different sets of demonstrations: What qualifies as "good teaching"? The pedagogical model in Eqs. (1) and (2) suggests that a demonstration's utility should be proportional to the log-probability that the learner will infer the correct hypothesis. However, this model does not take into account the fact that some things in the world are more useful to know than others (i.e., higher in value). Indeed, recent work suggests that 5- to 7-year-old children are sensitive to the epistemic value of different causal mechanisms and prioritize teaching high-value mechanisms (Bridgers et al., 2020). To incorporate the idea that good teaching involves providing information that leads to not just accurate but also high-value knowledge, we define the pedagogical utility of a particular set of demonstrations (d) as:

$$U(d) = \sum_i V(f_{T_i}) ln p_{L0}(f_i = 1|d_i) \tag{4}$$

where d_i is 1 if the i^{th} function was presented and 0 otherwise; and $p_{L0}(f_i = 1|d_i)$ is the teacher's model of the naïve learner's (L0) beliefs about function i after demonstration d_i. $V(f_{T_i})$ is the value of the i^{th} function known to the teacher; for simplicity we assume that all functions are either high or low in value. The second free parameter in our model captures the difference in the utility of knowing a high- versus a low- value function. See Appendix A for additional details about this parameter.

The pedagogical utility function in Eq. (4) is a generalization of previous accounts, which have examined the special case in which all functions are equally valuable (effectively dropping $V(f_{T_i})$ from the equation). Assuming that a demonstration of each function provides an equal amount of information to the learner, and that *a priori* belief in each function is the same, Eq. (4) becomes:

$$U(d) = \sum_i k d_i V(f_{T_i}), \tag{5}$$

where k is a constant reflecting the change in the learner's belief that a function exists, resulting from a demonstration. (This constant will be absorbed into the overall calibration of utility below.)

The demonstrations that a teacher selects using Eq. (3), which includes the pedagogical utility function (Eq. (5)), depends on the precise balance between the communication cost ($C(d)$), the teacher's quality (α), and the value of each function ($V(f_{T_i})$). Specifically, the communication cost pulls against the tendency of high-quality teachers to teach all relevant functions, such that a good teacher with high communication costs may only teach high-value functions.

To use this model to *evaluate* (rather than learn from) a teacher, consider a learner (L1) who knows that a teacher knows functions f_T and observes the teacher's d demonstration(s). Critically, as discussed in the Introduction, evaluating the teacher requires prior knowledge; a naïve learner who does not know that a toy has four functions cannot detect the omission of a teacher who just shows one function. Thus, we denote this knowledgeable learner – observer and evaluator of a pedagogical demonstration – as L1, separately from the teacher's naïve target of demonstration (L0). Note that in our model (and in our behavioral experiments), we assume that the observer knows about all of the toy's functions ($f_{L1} = [1,1,1,1]$) and also has a true belief about the teacher's knowledge of the functions f_T.[3]

L1 can use Bayes' rule to invert their model of how teachers select demonstrations (Eq. (3)) to infer the teacher's quality:

$$p(\alpha|f_T, d) \propto p(d|f_T, \alpha)p(\alpha), \tag{6}$$

where $p(\alpha)$ represents L1's prior beliefs about the teacher's quality. (We assume $p(\alpha)$&Uniform$(0,1)$, but see Appendix A for notes on how this was implemented.) The resulting estimates of teacher quality are sensitive to the number of functions omitted, the value of those functions, and the teacher's knowledge state (or more accurately, L1's belief about the teacher's knowledge state). For example, the quality estimate of a teacher who knows all functions but only demonstrates two low-value functions would be lower than someone who showed the same functions but only knew the two low-value functions. Similarly, a teacher who knows both a high- and a low- value function would get the highest rating for showing both, a lower rating for omitting the low-value function, and an even lower rating for omitting the high-value function. Therefore, in general, the model predicts that the more information a teacher shares relative to what they know, and the higher-value that information, the better they will be evaluated. (See Fig. 1 for a causal graphical representation of the model. Additional implementation details can be found in Appendix A.)

3. Experiment 1: testing model predictions

So far, we have presented a formal Bayesian account of how a learner might evaluate a teacher. This model considers three factors – the number of demonstrated functions, their value, as well as the teacher's prior knowledge of the functions – to generate predictions. Thus, when a teacher fails to demonstrate all functions, the model predictions should reflect the degree of omission (i.e., how many functions were left undemonstrated?), their value (i.e., were the omitted functions high or low in value?), and the teacher's knowledge (i.e., did the teacher know about those omitted functions, or not?). In Experiment 1, we compared human judgments against our model predictions. To this end, we presented adults with a range of teaching scenarios across 15 between-subjects conditions in which we systematically varied these three factors, and asked them to rate the teacher's quality. High correspondence between the model and the behavioral data would suggest that human adults' evaluations of pedagogical demonstrations integrate the three intuitive factors captured by our model (i.e., higher ratings for teachers who demonstrate more functions that are higher in value relative to

[3] We acknowledge two limitations that come with these assumptions. First, there are cases in which L1's belief about what the teacher knows may be false; modeling and testing these "inaccurate" evaluations is not within the scope of the current work. Second, cases like our experimental scenarios – wherein the evaluator already possesses knowledge about the correct hypothesis (e.g., the toy's causal functions) as well as the teacher's epistemic state – may seem uncommon in practice, and how exactly learners might come to obtain this kind of knowledge in order to make informed evaluations is an open question for future work. We chose to study such cases because knowledge about the world (i.e., the toy) and the teacher's mental states (e.g., knowledge) are important prerequisites for recognizing violations in pedagogical sampling; an observer who does not know what the toy does or what the teacher knows would not be able to even recognize omission itself (Gweon, 2021). Furthermore, even young children make implicit real-time inferences both about teachers' knowledge and likely hypotheses for the learning problem at hand (Bonawitz et al., 2011). Thus, we restrict this work to consider the case in which L1 has perfect knowledge about the toy and the teacher, but our model could theoretically also be extended to capture cases in which L1 has uncertainty about these things. We intend to demonstrate that the integrative process described by our model can, in principle, be used by learners to discriminate between more and less helpful potential informants.

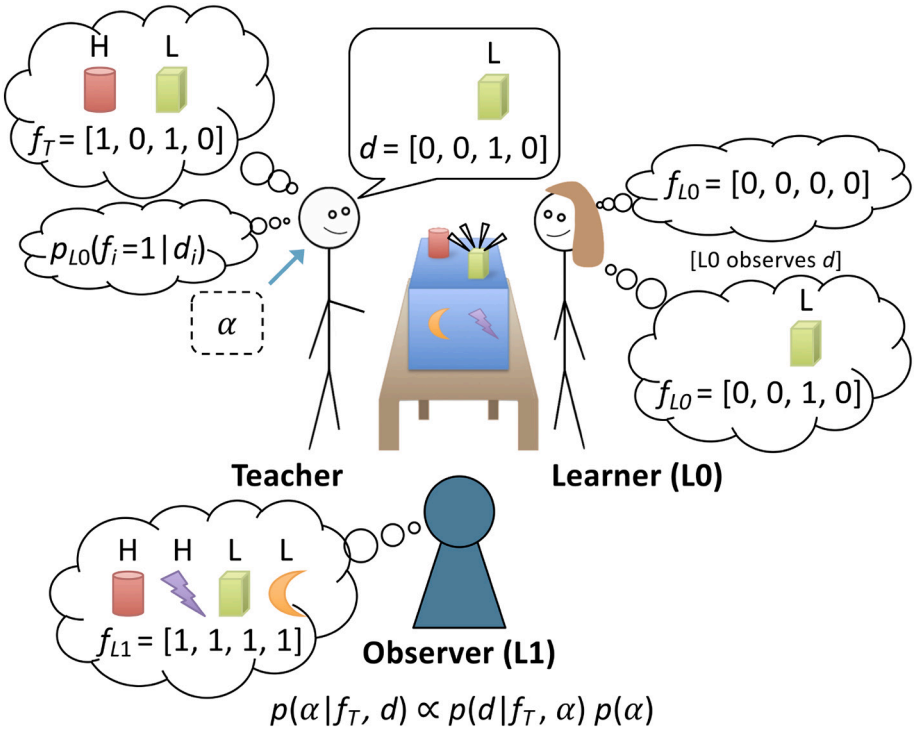

Fig. 1. A causal graphical representation of our computational model. An observer (L1) of a pedagogical interaction between a teacher (T) and a naïve learner (L0) may reverse-infer the the teacher's quality α from the quantity and value of their selected demonstrations d given the teacher's prior knowledge f_T. In our model, L0 is initially ignorant about the toy's functions, and learns deterministically about the functions demonstrated by the teacher (d). L1 observes the teacher's demonstration and evaluates its quality; L1 is fully knowledgeable about the toy's functions, and also has a true belief about the teacher's knowledge of the functions f_T.

what they know).

3.1. Method

This study was approved by the Oberlin College Institutional Review Board and the Wesleyan Psychology Ethics Committee.

3.1.1. Participants

All participants were recruited from Amazon Mechanical Turk, and were paid $0.50 for their participation. We had 15 between-subjects conditions built into our experiment, and a power analysis revealed that we would need roughly 70 participants in each condition to detect differences between conditions with 80% power (see Sections 3.1.3 and 3.1.4 for additional details). Therefore, our final sample consisted of 1168 participants ($M(SD)_{age} = 35.8(12.4)$ years, range = 18 − 72 years; $N = 661$ female), with between 69 and 92 participants in each condition. An additional 389 participants were dropped and replaced due to: failure to pass built-in check questions ($N = 318$; see Section 3.1.4 for details); or participating in the task more than once ($N = 71$). (This drop rate is not atypical for Mechanical Turk studies; see Zhou & Fishbach, 2016.)

3.1.2. Stimuli

We created cartoon scenarios in which a character named Paul encounters a box with four distinct buttons, each corresponding to a different function. Pressing a red circular button made the device tell the time (high-value); the purple lightning-shaped button reported the local weather (high-value); the orange crescent-shaped button made the device say "Hello!" (low-value); and the green square button generated a beep sound (low-value). The relative values of these functions were validated by a separate group of 52 mTurk participants, who rank-ordered the four functions by their perceived usefulness. Weather and Time ($M(SD) = 3.4(0.29)$) were ranked higher than Beep and Hello ($M(SD) = 1.6(0.29)$; $t(51) = 23.3$, $p < 0.001$), with no differences within the high-value pair or the low-value pair ($p \geq 0.298$; see Supplemental Materials for additional details).

In the cartoon scenarios, Paul (the teacher) enters a room to find this device sitting on a table, and begins to press the buttons to see what they do. After discovering either all four or a subset of its functions, another character (Laura; the naïve learner) enters the room, and asks how the device works.[4] Paul then shows Laura all or some subset of the functions he initially discovered. See Fig. 2 for an example.

3.1.3. Design

We varied the number of functions the teacher discovered (1, 2, or 4), the number of functions they demonstrated to the learner (1, 2, or 4), and the value of those functions (H for high, L for low). Crossing these variables while excluding impossible cases yielded 15 conditions (see Fig. 3 for a full list of conditions). For ease of reference, we abbreviate condition names by denoting which functions the teacher knew (K), and taught (T). For instance, in the KA_TA condition, the teacher knew all and taught all; in KHL_TL, they knew two functions (one high-value and one low-value) but taught only one low-value function. For conditions in which only one function of a given value was discovered or taught (e.g., KA_TH), we counterbalanced the exact function among the two equally valued functions (e.g., Weather vs. Time).

3.1.4. Procedure

After consenting to participate in the task on Mechanical Turk, participants were redirected to a Qualtrics survey, at which time they were randomly assigned to one of the 15 conditions. The first section of the experiment introduced participants to the device, to ensure they understood all of its functions. As an attention check, participants were then asked how many functions the device had; those who did not correctly answer this question were excluded from analysis. Next, participants were shown the cartoon scenario corresponding to their experimental condition, as described above. Participants then answered

[4] The primary purpose of having the teacher "discover" the device was to constrain the teacher's prior knowledge. To minimize the possibility that the teacher's failure to discover all functions is perceived as incompetence to explore (and affect participants' ratings), in all conditions, Paul's discovery was "interrupted" by Laura's entrance (see Fig. 2, panel 2).

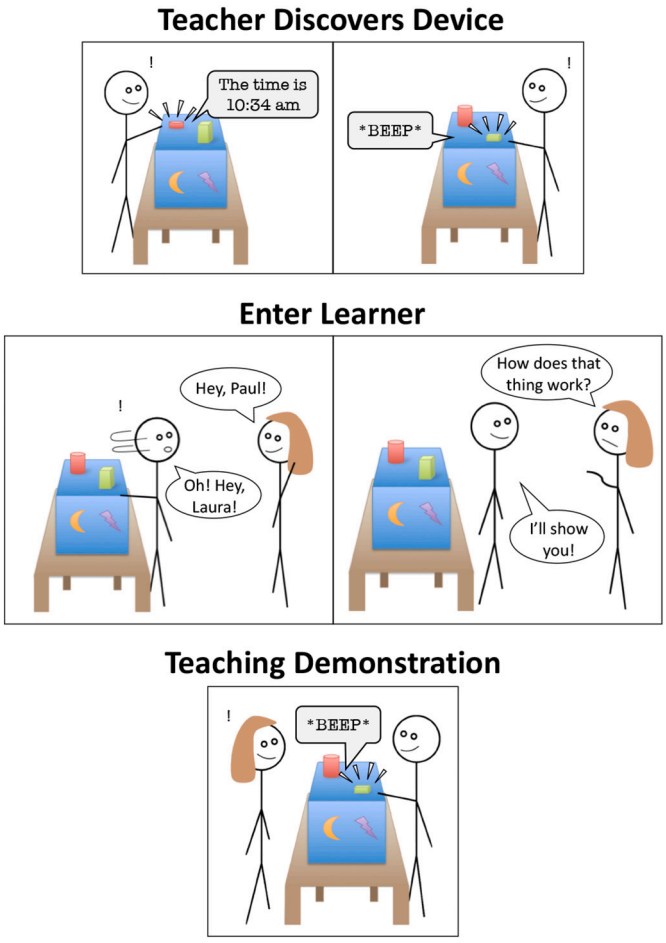

Fig. 2. A schematic of the cartoon scenarios presented to participants in Experiment 1, using the KHL_TL condition as an example. Here, the teacher (Paul) discovers a high-value function and a low-value function, and teaches the learner (Laura, L0) the low-value function.

the critical question: *Q1. Overall, how would you rate Paul's teaching abilities?* We also asked additional questions in a fixed order: *Q2. Which functions did Paul discover? Q3. Which functions did Paul teach? Q4. How well-intentioned do you think Paul was? Q5. How nice do you think Paul is? Q6. Given what he knew, how good a job did Paul do? Q7. How willing would you be to learn from Paul?* Finally, we asked the first question again (*Q8. Overall, how would you rate Paul's teaching abilities?*). We used a 1–7 Likert scale for Q1 and Q4–Q8. In addition to the number of functions on the toy, Q2 and Q3 were also used as attention checks; participants who could not correctly answer what the teacher had discovered and taught about the toy were excluded from analysis.

As an additional attention check, after answering all eight questions, participants were presented with a 4-sentence block of text in which the third sentence contained the following instructions: *If you are reading this question and have read all the other questions, please select the box marked 'other' and type 'Theory of Mind' in the box below.* Below this block of text, participants saw the final check question: *What was this study about?* There were four answer choices; participants who did not select "other" and type "theory of mind" into the box were excluded from analysis.

3.2. Results

Participants' ratings of the teacher's quality in the first and the last questions (Q1 and Q8) were highly correlated ($r(1165) = 0.832$, $p < 0.001$). We therefore used the average of these two ratings as the primary measure of quality (but see Results: Prompting intention and knowledge for analyses on the difference between Q1 and Q8). As the primary test of our hypothesis, we first compared the model predictions against this primary measure.

3.2.1. Model fit

The two free parameters in the model – the difference in the utility of knowing a high- versus a low-value function, and the cost the teacher incurs by communicating a function – were fit to the mean of participants' ratings of teacher quality across conditions. Prior to fitting the model we normalized participants' ratings to be on a scale from 0 to 1. The best-fit parameter values were 0.14 for the utility difference, and 0.73 for the communication cost. As seen in Fig. 3B, the fit between the model predictions and the participants' ratings of teacher quality was very high ($R^2 = 0.95$).[5] Supplemental figures can be found in Appendix B.

3.2.2. Behavioral results

We then ran confirmatory analyses to demonstrate that adult participants' evaluations indeed reflect all three factors in the model: number of functions, their value, and the teacher's prior knowledge. (Raw ratings were used in these analyses, across all experiments.) To this end, we coded these three factors as follows: The number of functions known and demonstrated could each take on values of 1, 2, or 4; and value was quantified as the proportion of demonstrated functions that were high-value, and could thus take on values of 0, 0.5, or 1. We included these three variables in a multiple linear regression model predicting participants' ratings. This regression was significant ($F(3, 1164) = 216.25$, $p < 0.001$, $R^2 = 0.358$), and all three factors were significant independent predictors. That is: The more the teacher showed, the higher they were rated ($\beta = 1.31$, $t(1164) = 25.13$, $p < 0.001$); the higher the value of their demonstrations, the higher the teacher was rated ($\beta = 0.358$, $t(1164) = 4.31$, $p < 0.001$); and the more functions the teacher knew, the lower they were rated ($\beta = -0.428$, $t(1164) = 11.70$, $p < 0.001$).[6] In particular, the effect of knowledge suggests that participants exonerated the teacher's omissions when their knowledge was limited, and therefore couldn't show what they omitted (rather than deliberately omitting). See Fig. 3A for mean ratings of teacher quality in all conditions.

To further investigate these main effects, we selected subsets of conditions that highlight the effects of number, value, and the teacher's knowledge state, and ran targeted analyses on these conditions. We present these results in the next two sections.

3.2.3. Number & value

First, we asked how the amount and value of the demonstrated information influenced evaluations of the teacher's quality. We ran a 2×2 ANOVA comparing conditions in which the teacher had discovered all four functions, and taught either 1 or 2 functions of high or low value (i.e., KA_THH, KA_TH, KA_TLL, & KA_TL). We found a significant main effect of number ($F(1, 280) = 36.3$, $p < 0.001$, $\eta_p^2 = 0.115$), such that teachers who showed one function ($M(SD) = 2.11(1.11)$) were rated significantly lower than those who showed two ($M(SD) = 3.02(1.45)$). Similarly, there was also a main effect of value ($F(1, 280) = 17.71$, $p < 0.001$, $\eta_p^2 = 0.059$), where demonstrating high-value functions

[5] Given the relatively high attrition rate in this experiment, we re-ran the model including those participants who had failed one or more the the check questions that served as exclusion criteria ($N = 1486$). Here, the best-fit parameter values were 0.13 for the utility difference (nearly identical), and 0.8 for the communication cost (slightly higher). Fit to the data was still very high: $R^2 = 0.96$. We take this as support for the robustness of our model to possible sources of variability.

[6] This pattern of results also holds when including only conditions in which the teacher omitted some amount of information relative to what they knew, suggesting the results are not driven by the KA_TA condition alone.

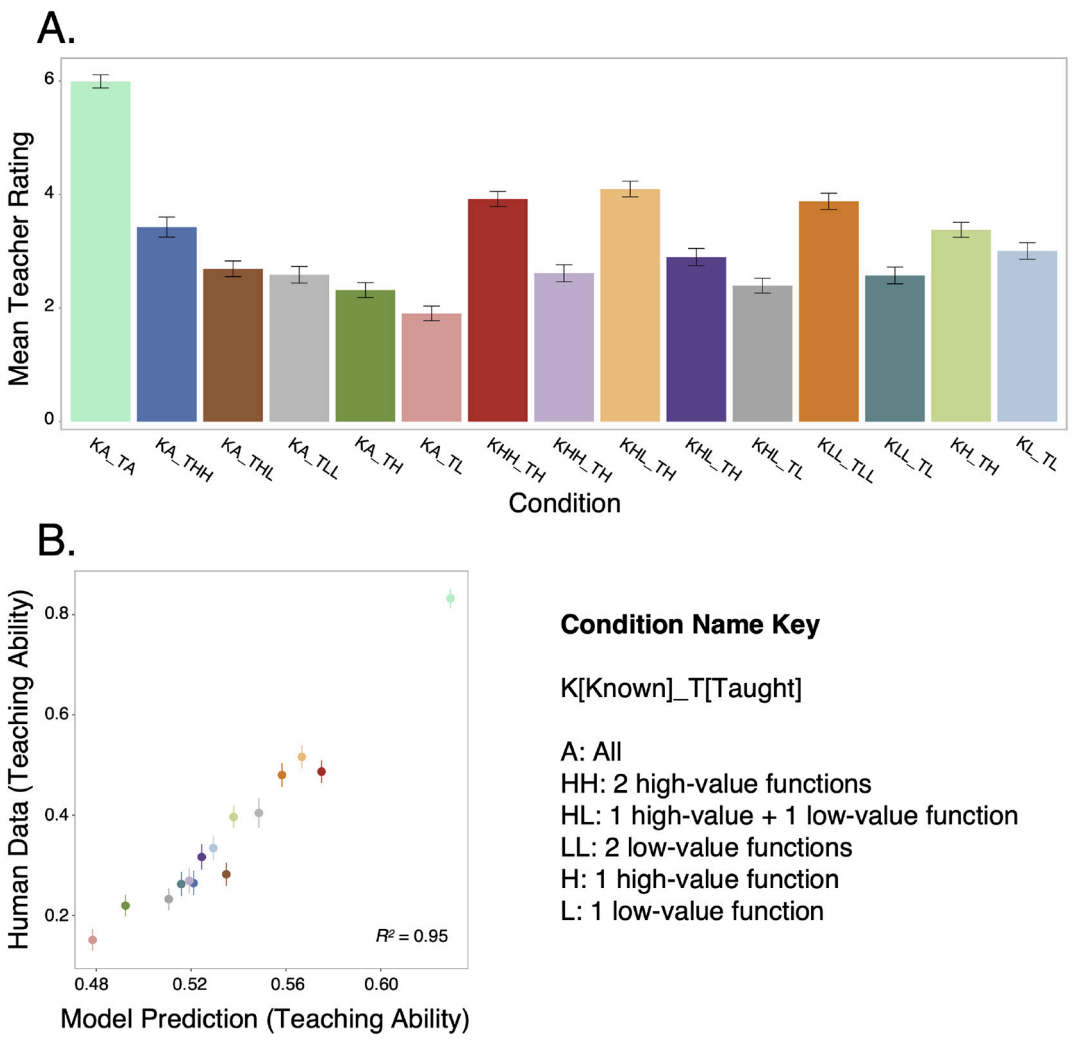

Fig. 3. A. Mean teacher ratings in all 15 conditions in Experiment 1. Ratings were significantly predicted by all three factors (number, value, and knowledge state). B. Comparison of behavioral data and the model predictions. Each colored point corresponds to a bar of the same color in A. The fit between model and human data is $R^2 = 0.95$.

($M(SD) = 2.89(1.45)$) was evaluated as better than demonstrating low-value functions ($M(SD) = 2.25(1.2)$). The interaction was non-significant (p=.150), such that the difference in ratings between teachers who showed high- versus low-value functions did not depend on the number of functions demonstrated.

3.2.4. Teacher's knowledge state

To understand how the teacher's knowledge state influenced participants' ratings, we performed two separate analyses. First, we looked at four conditions in which the teacher taught two functions; these two functions either constituted everything the teacher knew (i.e., KHH_THH, KLL_TLL), or only a subset of what they knew (i.e., KA_THH, KA_TLL). A 2 × 2 ANOVA revealed a significant main effect of knowledge state ($F(1, 315) = 34.85$, $p < 0.001$, $\eta_p^2 = 0.10$): Although the teachers in these conditions demonstrated the same number of functions, participants gave higher ratings to teachers who taught everything they knew ($M(SD) = 3.9(1.3)$) than they did to teachers who did not ($M(SD) = 3.02(1.45)$). We also saw a significant main effect of value ($F(1, 315) = 8.46$, $p = 0.004$, $\eta_p^2 = 0.026$), qualified by a significant interaction ($F(1,315) = 6.95$,$p = 0.009$, $\eta_p^2 = 0.022$). In particular, the value of the demonstrated functions influenced participants' ratings only when the teacher taught a subset of what they knew ($M_{dif} = 0.840$, $t(142) = 3.62$, $p < 0.001$); if they taught all they knew, information value did not have a significant effect on ratings of their quality ($M_{dif} = 0.041$, $p = 0.834$).

In order to assess the effect of knowledge state in a more fine-grained manner, we also performed a one-way ANOVA on conditions in which the teacher showed one low-value function, and had discovered either one low-value function, two low-value functions, or all four functions (i.e., KL_TL, KLL_TL, KA_TL). This ANOVA was significant ($F(2, 232) = 14.56$, $p < 0.001$, $\eta_p^2 = 0.112$). Post-hoc Bonferroni corrected pairwise comparisons revealed that while all three of these teachers provided identical demonstrations, the teacher who *knew* all 4 functions ($M(SD) = 1.91(1.08)$) was rated lower than the teachers who knew only 2 functions ($M(SD) = 2.58(1.32)$) or 1 function ($M(SD) = 3.01(1.34)$), $ps < 0.01$. Although ratings between the KLL_TL and KL_TL teachers did not significantly differ from each other, the overall pattern of results is consistent with our hypothesis that evaluations reflect the degree of omission relative to the teacher's knowledge state, where teachers who knew more and thus *could have* shown more are rated more harshly.

3.2.5. Prompting intention and knowledge

Our questionnaire had 8 questions, with the mean of the first and the last questions serving as our main dependent measure thus far. Between these two identical questions, participants were asked several questions that may have prompted them to think about the knowledge and intentions of the teacher. Thus, it is possible that participants' ratings

would be more likely to reflect the teacher's knowledge after having been explicitly prompted to think about it; if so, answering these questions would lead to amplified differences in ratings of omissions from full versus limited knowledge. Although such an effect is not directly predicted by our model, we tested this idea with an exploratory analysis.

We again compared the same three conditions in which the teacher showed one low-value function, and had discovered either one low-value function, two low-value functions, or all four functions (i.e., KL_TL, KLL_TL, KA_TL). However, here we used the difference between Q1 and Q8 as the dependent variable (as opposed to the mean) in a one-way ANOVA. This analysis revealed a significant main effect of prior knowledge ($F(2, 232) = 17.4$, $p < 0.001$, $\eta_p^2 = 0.130$). That is, the less the teacher knew, the more likely participants were to increase their ratings from Q1 to Q8. Post-hoc Bonferroni corrected pairwise comparisons revealed significant differences between the condition in which the teacher showed all they knew (KL_TL: $M(SD) = 0.663(1.02)$) and the two conditions in which they did not (KLL_TL: $M(SD) = 0.15(0.731)$; KA_TL: $M(SD) = -0.072(0.524)$; $ps < 0.001$). Further, a series of one-sample t-tests revealed that when the teacher taught all they knew, participants significantly increased their ratings from Q1 to Q8 ($t(85) = 6.00$, $p < 0.001$); this increase was marginal for the teacher who knew only two functions ($t(79) = 1.84$, $p = 0.070$), and non-significant when the teacher knew all four functions ($p = 0.254$). These results suggest that prompting participants to consider the teacher's intentions and knowledge may have led them to be more willing to pardon incomplete teaching when it could be explained by insufficient knowledge.

3.3. Discussion

Our model predictions were well-supported by adults' evaluations of teachers' quality across a variety of different pedagogical scenarios. Participants' ratings were graded with respect to the number of functions the teacher demonstrated, the value of those functions, and what the teacher knew, suggesting that adults make nuanced evaluations that integrate these factors. An additional exploratory analysis also revealed that prompting participants to consider the teacher's epistemic state led them to exonerate "innocent" omissions (i.e., omissions by teachers who could not show because they did not know) to a greater degree. While it is notable that participants were sensitive to the rather subtle manipulation of the teacher's epistemic state, we were able to see even clearer effects when we explicitly asked them to consider this factor.

Note, however, that naïve omission was not completely pardoned; as can be seen in Fig. 3A, teachers who knew less were still rated less favorably, even when they taught all they knew. For instance, although the teachers in KA_TA, KHH_THH, KLL_TLL, KH_TH, and KL_TL conditions all taught everything they knew, participants' ratings reflected teacher's degree of knowledge (f_T). While our model does not explicitly take into account how f_T itself might influence L1's evaluation of α, this pattern suggests that participants might bring their judgments of the teacher's knowledge to bear on their evaluations of overall quality, independent of what the teacher demonstrated to the learner.

The systematicity in adults' evaluations of teachers across various scenarios raise questions about how these abilities emerge in development. Given that by 4 years of age children appropriately prefer to learn from teachers who provide accurate (as opposed to inaccurate) information (e.g., Koenig et al., 2004) as well as those who provide complete (as opposed to incomplete) demonstrations (e.g., Gweon & Asaba, 2018), it is possible that children around this age are also capable of considering factors beyond accuracy, such as the amount and value of omitted information and the teacher's knowledge state. However, as outlined in the Introduction (Sections 1.1–1.3), there are also reasons to suspect that young children may have difficulty distinguishing between different instances of omission. The remaining experiments thus seek to understand whether the same kinds of nuanced teacher evaluations seen in adults in Experiment 1 may also be present in early childhood.

4. Experiment 2a: preschoolers' teacher evaluations

In order to assess children's teacher evaluation abilities, we developed a child-friendly method with reduced task demands. This involved two broad adaptations from Experiment 1. First, to address relatively smaller sample size and power compared with the adult experiment, we selected 5 key conditions among the 15 from Experiment 1, and implemented them in a within-subjects design. Second, we used prerecorded videos of teaching scenarios instead of cartoons, and provided memory cues for each (see Section 4.1.3 for additional details). To ensure that this modified task elicits ratings that are comparable to Experiment 1, we administered this new task with a group of preschoolers (Experiment 2a) as well as a group of adults (Experiment 2b).

4.1. Method

This study was approved by the Rutgers University – Newark Institutional Review Board, protocol 16–625Mc. Informed parental consent and child assent were obtained before the study was administered.

4.1.1. Participants

Participants were children recruited from and tested at local preschools and daycares. Because our method involved 24 possible counterbalanced orders of stimuli presentation (see Procedure), our final sample consisted of 24 preschoolers ($M(SD)_{age} = 60(5.30)$ months, range $= 49 - 72$ months; $N = 12$ female). An additional 12 children were dropped and replaced due to: *a priori* exclusion criteria ($N = 9$; see Procedure); wanting to terminate the study early ($N = 1$); or being outside of our target age range ($N = 2$).

4.1.2. Materials

See Fig. 4 for graphical depictions of many of the materials used in this task.

4.1.2.1. Rating scale. Participants used a 0 to 20 point rating scale to evaluate teachers. A small circular magnetic marker could be placed on the scale to indicate how good they thought a teacher was.

4.1.2.2. Teaching toy. The teaching toy was a square pyramid covered in blue felt with four colorful buttons, each corresponding to a different function. Two functions were low-value: The toy could beep, and it could produce white noise. The other two functions were high-value: The toy could play clips of two different children's songs. The relative value of these functions were validated in a separate group of 10 children ($M(SD)_{age} = 66.8(15.2)$ months, range $= 49 - 91$ months), who were asked to rate "how cool" each of the four functions were using the rating scale described above. The two songs ($M(SD) = 15.3(3.1)$) were rated significantly higher than the beep and the noise ($M(SD) = 8.8(3.7)$; $t(9) = 3.41$, $p = 0.008$), with no differences within the high-value pair or the low-value pair ($ps \geq 0.33$; see Supplemental Materials for additional details).

4.1.2.3. Teaching videos. Instead of using cartoon scenarios, we filmed live-action videos of adults discovering the toy and teaching its functions to another adult, as similar paradigms have been used successfully with young children in previous work see (Gweon & Asaba, 2018). Teaching videos were presented on a 15-inch MacBook Pro using Microsoft PowerPoint, and the format of these videos was similar to what occurred in the cartoons from Experiment 1: The teacher (equivalent to Paul in scenarios used in Experiment 1) enters a room to find the toy sitting on the table, proclaims aloud that she's never seen it before, and begins to press the buttons to see what they do. After discovering either all four or a subset of its functions, the learner (equivalent to Laura, the naïve learner (L0), in Experiment 1) enters the room (again startling the teacher out of her exploration), and asks how the device works. The

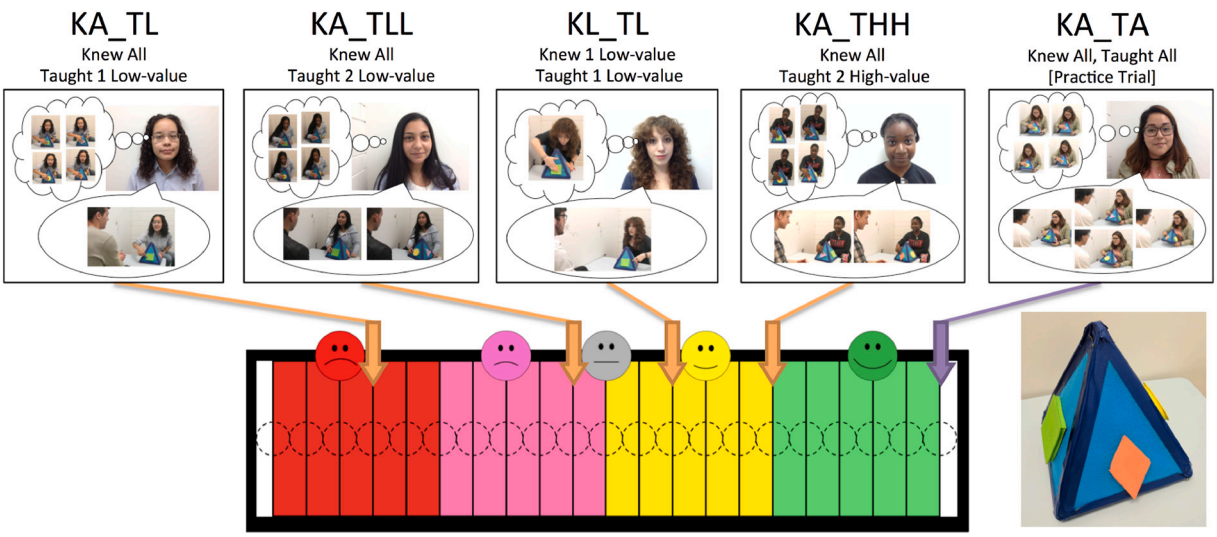

Fig. 4. The rating scale, memory cues, and toy used in Experiments 2–5. Memory cues were adhered to the rating scale as participants provided their ratings.

teacher then shows the learner all or some subset of the functions she initially discovered. At the conclusion of the video, the teacher says, "Pretty cool, right? That's how this toy works!" thereby clearly ending the teaching demonstration.

Five of the fifteen conditions from Experiment 1 were filmed for these teaching videos, specifically: KA_TA, KA_THH, KA_TLL, KA_TL, and KL_TL. We chose these particular conditions because planned pairwise comparisons would allow us to assess the effects of number, value, and knowledge state without needing to run all fifteen conditions. Specifically, comparing ratings of KA_TLL to KA_TL allows us to assess the effect of degree of omission, while holding value and knowledge state constant; comparing KA_TLL to KA_THH taps into the effect of value, while holding number and knowledge state constant; and comparing KA_TL to KL_TL will provide insight into the effects of knowledge state, while holding number and value constant.

In contrast with Experiment 1, we opted to use a within-subjects design for this study, such that all participants rated five teachers as opposed to just one. We fully counterbalanced the order in which the five teachers were seen, with one caveat: All participants saw the teacher who knew and taught all four of the toy's functions first, and were told that this was an example of excellent teaching. Our decision to anchor responses in this way was motivated by two prior findings: First, children reliably rate teachers highly when they provide complete and true information (e.g., Gweon et al., 2014; Koenig & Harris, 2005); second, four- and five-year-olds have trouble evaluating teachers who omit information without seeing a fully informative teacher first, possibly because they struggle to spontaneously consider relevant alternatives for reference points in this context (Gweon & Asaba, 2018). Presenting the KA_TA teacher first as an example of a fully informative teacher signaled to children that demonstrating all four functions was desirable, thereby aligning children's assumptions with those made in the model. Given that we were primarily interested in ratings of the four under-informative teachers relative to each other (and not to the fully informative teacher), we treated the KA_TA teacher as a practice trial, and anchored ratings of this "ideal" teacher at the top of the scale. Thus, the current study focuses on whether children have the competence to make nuanced evaluation of a teacher's omission given sufficient contextual support, which is separate from the question about the degree to which such competence manifests spontaneously.

4.1.2.4. Memory cues. To help participants recall what each teacher knew and taught, we created cards that depicted screenshots of the exploration and teaching phases from the teaching videos. Small arrows with adhesive backs were attached to each memory cue.

4.1.3. Procedure

4.1.3.1. Frame story & rating scale training. Participants were told that they would be meeting some people who were in teaching school learning how to be really good teachers; the experimenter needed help figuring out how good the different teachers were, because she had lost track of how much school each of the teachers still had left. The experimenter then introduced the rating scale, and participants were briefly trained on how to use it to indicate teacher quality. Those who were not able to pass this training were not included in subsequent analysis ($N = 1$).

4.1.3.2. Teaching toy exploration. Next, the experimenter introduced the teaching toy, and encouraged participants to figure out how it worked. After they had successfully pressed all four buttons, participants were then told that, "the other day, the teachers from the teaching school had found that same toy, and taught some new students about how it worked". It would be the participant's job to watch videos of the teachers discovering and teaching about the toy, in order to "figure out how good each teacher was at teaching, so we can know how much teaching school they all have left".

4.1.3.3. Teacher evaluations. All participants first saw the KA_TA condition as a practice trial. Before watching the video, the experimenter explained that this teacher had already graduated from the teaching school, and was therefore an example of a really good teacher. After watching this first video, participants were shown the memory cue for the KA_TA teacher, and were asked to provide a rating. Those who did not place the marker at or near the top of the rating scale were reminded that this teacher was already done with school. If after several prompts the participant still did not rate this practice teacher highly, the experiment was terminated (did not occur). After the participant had provided a rating, the experimenter adhered the memory cue's arrow to the rating scale in the same location, and then removed the marker from the scale.

Participants then viewed the remaining four teachers – who, they were reminded, were still in teaching school. These constituted the four test trials. The order in which these four teaching scenarios were presented was fully counterbalanced, yielding 24 different orders. While the actors in the videos and the test conditions were counterbalanced with respect to each other, the *order* of the actors was always the same (e.g., "Liz" was always the first test teacher, but the first test condition varied between participants). After watching each video, participants were shown the memory cue for the teacher they had just seen, and were

asked to provide a rating. The memory cue's arrow was adhered to the scale after each rating. Pilot testing revealed several patterns in participants' ratings that indicated they either did not understand the task or were not paying attention, and thus served as exclusion criteria: placing teachers in descending order on the rating scale as they saw them ($N = 7$); giving all teachers the same rating ($N = 1$); or rating any of the test teachers higher than the practice teacher (did not occur).

4.2. Results

As in Experiment 1, we first compare children's responses with the model predictions, and then provide results from additional confirmatory analyses.

4.2.1. Model fit

We fit our model to children's ratings in Experiment 2a. As we did in Experiment 1, we first standardized ratings such that values ranged from 0 to 1 (this was done in all experiments prior to model fitting). Because the KA_TA teacher was used as a practice trial to anchor children's responses – which was not the case for adults in Experiment 1 – it was not clear whether it was appropriate to include children's ratings of this teacher in model fitting. Therefore, we opted to fit the model to the data both ways (i.e., with KA_TA ratings included and excluded).

In both cases, the model fit was very high ($R^2 = 1.00$). The best-fit values for the utility difference parameter, however, was considerably smaller than they were for the adult data in Experiment 1 (KA_TA included: $= 0.02$; KA_TA excluded: $= 0.04$). For communication cost, the best-fit values were closer to what we found with adults (KA_TA included: $= 0.75$; KA_TA excluded: $= 0.86$). See Figs. 5B & C for modeling results. Supplemental figures can be found in Appendix B.

4.2.2. Behavioral results

We then asked whether children differentiated between the four teachers in the test trials. A repeated-measures ANOVA on raw ratings in these four trials revealed a significant main effect of condition ($F(3, 69) = 3.50, p = 0.020, \eta_p^2 = 0.132$; see Fig. 5A). We therefore conducted follow-up pairwise comparisons between conditions (described in the Teaching videos section above) to investigate children's sensitivity to number (i.e., number of demonstrated vs. undemonstrated functions, or the degree of omission), the value of these functions, and the teacher's knowledge state.

4.2.3. Number

To investigate the effects of number (i.e., the degree of omission), we compared ratings of the KA_TL teacher to the KA_TLL teacher. This comparison was significant: The teacher who omitted three functions ($M = 7.77$, $SD = 5.42$) was rated significantly lower than teacher who omitted only two ($M = 11.88$, $SD = 6.39$); $t(23) = 2.54, p = 0.019, d = 0.517$. We also found that a majority of participants (67%) rated the teacher who showed two functions as better than the teacher who showed just one; this proportion was marginally greater than 50% by binomial test ($p = 0.076$, one-tailed).

4.2.4. Value

Contrary to our predictions, we did not find a significant effect of value. Children's ratings of the KA_TLL teacher did not differ from the KA_THH teacher ($p = 0.874$). Only 58% of participants rated the high-value teacher as better than the low-value teacher, which did not differ significantly from chance ($p = 0.271$, one-tailed). This is consistent with the significantly lower parameter for the utility differences. We return to this result in Section 4.3 (as well as in Experiments 2b and 3).

4.2.5. Teacher's knowledge state

We explored the effects of the teacher's knowledge state on evaluations by comparing the teacher who *knowingly* omitted three of the toy's functions (KA_TL) to the teacher who did so "innocently" (KL_TL). We found significant differences between these two teachers: Although they provided identical demonstrations, the teacher who knowingly omitted was rated lower than the teacher who omitted naïvely ($M = 11.54$, $SD = 5.70$); $t(23) = 2.58, p = 0.017, d = 0.526$. A significant proportion of children (71%) rated the KL_TL teacher higher than the KA_TL teacher ($p = 0.032$, one-tailed).

4.3. Discussion

Here, we investigated whether preschool-aged children can evaluate various instances of omission in pedagogical contexts in ways that integrate the number of omitted functions, the value of these functions, and the knowledge state of the teacher. The children in our study made graded judgments in ways that are consistent with the model's predictions: They considered the degree of omission in the teacher's demonstration and the teacher's knowledge state.

This latter finding is particularly remarkable: Given that four-year-olds are just beginning to reason explicitly about others' minds (Wellman et al., 2001), and that we might not expect to see such sophisticated evidential reasoning abilities until later in childhood (Rhodes, Brickman, & Gelman, 2008; Rhodes, Gelman, & Brickman, 2008), it is impressive that preschoolers were able to evaluate a teacher's incomplete demonstrations in light of their limited knowledge. While this work lends preliminary support to the idea that children can make sophisticated inferences to guide their evaluation of teachers, given these rather striking results we found them worthy of replication; we return to this point in Experiment 4.

It is also encouraging that our model was able to capture children's ratings in this task so accurately. This lends support to the idea that even young children's cognitive models of teacher evaluations may be quite sophisticated. Importantly, however, the best-fit value for the utility difference between knowing a high- versus a low-value function was smaller as compared with Experiment 1. Indeed, while we did validate the perceived value of the teaching toy's functions with a separate group of children, the predicted effect of value was not reflected in children's ratings. This may have been due to a few potential reasons. First, although children did perceive the high-value functions as "cooler", unlike the high-value functions that were informational in Experiment 1 (e.g., reporting the weather), the value of these functions were arguably more perceptual. Thus, it is possible that children did not consider these values as relevant for evaluating *teachers*. Second, due to limitations in sample size and statistical power, we used a within-subjects design; however, it is possible that this made the task too demanding for young children. Indeed, instead of a binary choice task that contrasts just two teachers, here children had to track and contrast all relevant factors across five teaching scenarios. Even in adults (Experiment 1), the effect of value was quantitatively weaker than those of number and knowledge state. Thus, it could be that if asked to compare multiple teachers simultaneously, less salient factors fall by the wayside while others are prioritized. Finally, it is also possible that these null results reflect a genuine lack of competence in children's ability to consider the value of information in their teacher evaluations. We therefore found it necessary to run the same task with a group of adults to see whether this method can indeed elicit intuitions equivalent to those from the task used in Experiment 1.

5. Experiment 2b: adult performance on the child task

In Experiment 1, adults' teacher evaluations reflected small variations in the utility of the teacher's demonstrations relative to their knowledge state. An inability to reproduce these effects using the child-friendly task from Experiment 2a would suggest that this method may not be comparable to Experiment 1 in the kinds of ratings that are being elicited.

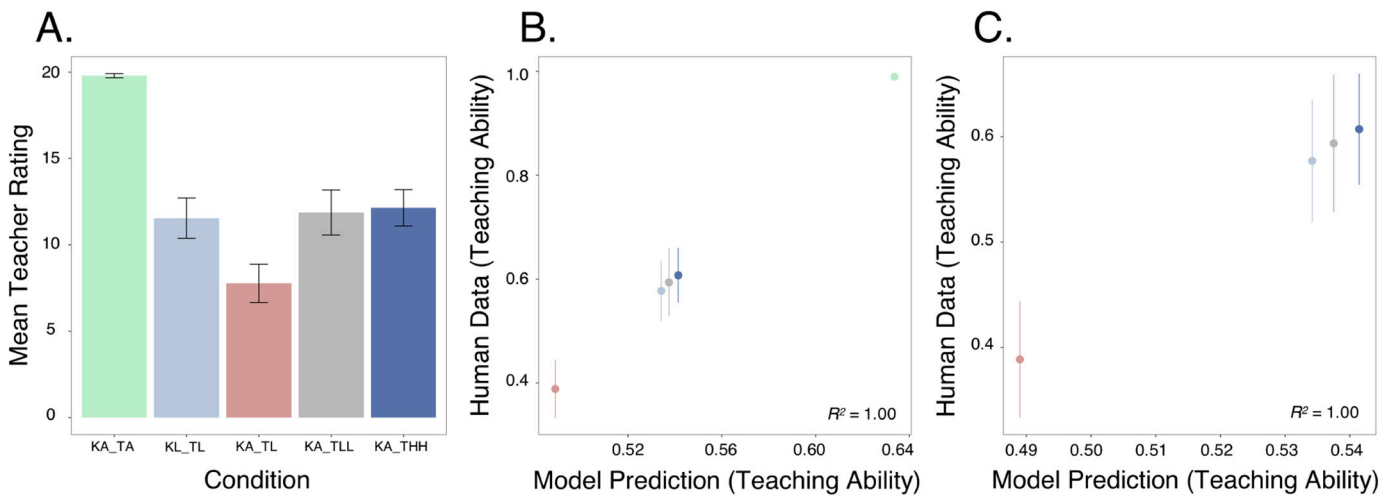

Fig. 5. A. Preschoolers' teacher ratings in Experiment 2a. Participants significantly differentiated teachers based on number and knowledge state, but not value. B & C. Comparison of behavioral data and model predictions when including and excluding the KA_TA condition respectively. A perfect fit was observed between the model and children's ratings in both cases, $R^2 = 1.00$.

5.1. Method

This study was approved by the Rutgers University – Newark Institutional Review Board, protocol 16–625Mc. Informed consent was obtained before the study was administered.

5.1.1. Participants

All participants were recruited from the Rutgers University – Newark psychology subject pool, which comprises Rutgers undergraduates enrolled in introductory psychology classes. They received course credit as compensation for participating. Our final sample consisted of 24 adults ($M(SD)_{age} = 23.6(5.21)$ years, range $= 18 - 40$ years; $N = 17$ female). An additional 3 participants were dropped and replaced due to: placing the teachers in descending order as they saw them ($N = 2$); or rating a test teacher higher than the practice teacher ($N = 1$).

5.1.2. Procedure

The materials and procedure were identical to Experiment 2a.

5.2. Results

5.2.1. Model fit

As with Experiment 2a, we fit the model to the data both with and without the KA_TA condition. Again, the fit was very high in both cases ($R^2 = 1.00$). The best-fit values for the utility difference (KA_TA included: $= 0.04$; KA_TA excluded: $= 0.02$) and communication cost (KA_TA included: $= 0.86$; KA_TA excluded: $= 0.75$) were also similar to Experiment 2a. See Figs. 6B & C for modeling results. Supplemental figures can be found in Appendix B.

5.2.2. Behavioral results

The initial repeated-measures ANOVA on ratings of the four test teachers was significant ($F(3, 69) = 10.38$, $p < 0.001$, $\eta_p^2 = 0.311$; see Fig. 6A). Therefore, as in Experiment 2a, we followed up with pairwise comparisons to investigate the effects of number, value, and knowledge state on adults' evaluations.

5.2.3. Number

Mirroring the effects from Experiments 1 and 2a, we found significant effects of number: The teacher who demonstrated two functions ($M = 10.42$, $SD = 2.90$) was rated significantly higher than the teacher who showed just one ($M = 6.04$, $SD = 3.88$); $t(23) = 6.45$, $p < 0.001$, $d = 1.32$. Further, 88% of participants rated the KA_TLL higher than KA_TL, which is significantly different from chance (50%) by binomial test ($p < 0.001$, one-tailed).

5.2.4. Value

The two teachers who demonstrated the same number of functions but of differing values (KA_TLL, KA_THH) were not significantly differentiated in participants' ratings ($p = 0.220$). This is in contrast to adults' ratings in Experiment 1, but similar to children's in Experiment 2a. Interestingly, only 21% of participants rated the KA_THH teacher higher than the KA_TLL teacher, which is significantly *below* chance ($p < 0.001$, one-tailed). We discuss possible interpretations of this result in Section 5.3.

5.2.5. Teacher's knowledge state

Finally, we compared ratings of the KA_TL teacher to the KL_TL teacher to examine whether adults' ratings were sensitive to knowledge state in this paradigm. Again reflecting the results from Experiments 1 and 2a, we found that this former teacher was rated lower than the latter ($M = 8.50$, $SD = 6.01$); $t(23) = 2.06$, $p = 0.051$, $d = 0.42$. Although a majority of participants rated the KL_TL teacher higher than the KA_TL teacher (58%), this proportion did not differ from chance ($p = 0.271$, one-tailed).

5.3. Discussion

The results from this experiment directly mirror those from Experiment 2a: Both adults' and children's ratings in this adapted paradigm reflected sensitivity to number and knowledge state, but not to value. Unfortunately, adults' current failure to differentiate teachers based on information value does not actually allow us to rule out any of the aforementioned possibilities for the reasons behind children's failures; in fact, it is possible that adults and children failed for entirely different reasons. On the one hand, in Experiment 1 adults did consider information value in their evaluations of teachers; but they may have failed to do so here because they attributed value to knowing (and teaching) all of these functions (thus not discriminating between teachers who provide low- versus high-value functions), or because they did not consider this more perceptual operationalization of "value" as being relevant for evaluation of teachers. In fact, given that we did not validate the perceived value of these functions with adults, it is not even clear whether they considered the high- and low-value functions to be perceptually distinct in this task. On the other hand, pilot testing revealed that children did find the high-value functions to be "cooler"

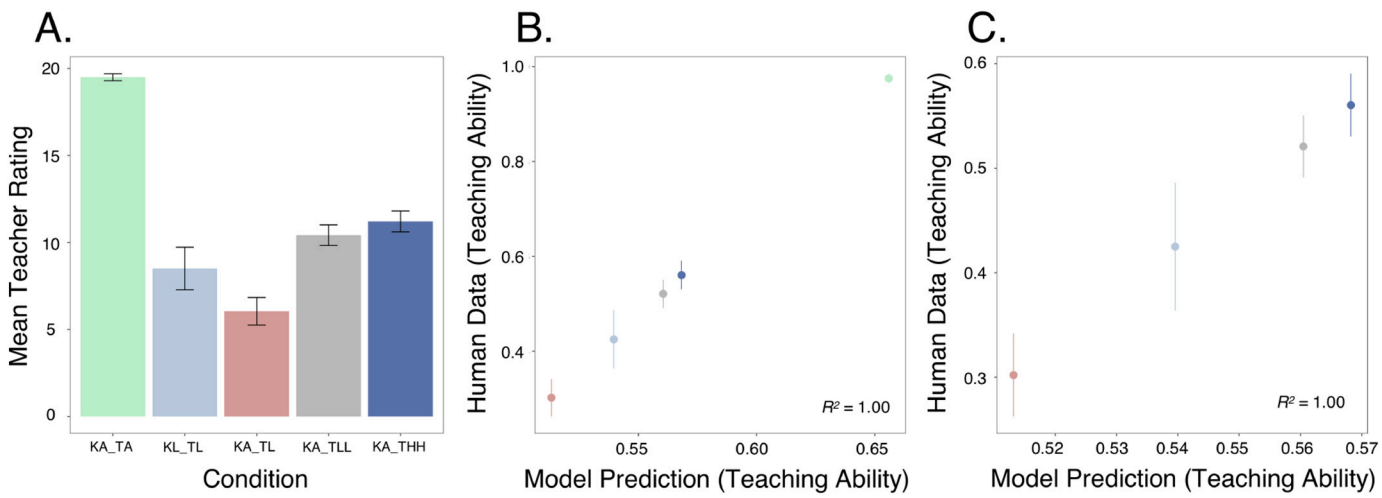

Fig. 6. A. Adults' teacher ratings in Experiment 2b. Participants significantly differentiated teachers based on number and knowledge state, but not value. B & C. Comparison of behavioral data and model predictions when including and excluding the KA_TA condition respectively. As in Experiment 2a, the fit between the model predictions and participants' ratings were very high in both cases $R^2 = 1.00$.

than the low-value functions on this toy; however, children might not have interpreted these functions as providing a basis for differential evaluations of *teachers*, or they could have failed due to the various demands of the paradigm, or even because they genuinely lack the ability to integrate information value into their teacher evaluations. We try to further tease apart these possibilities in Experiment 3.

6. Experiment 3: targeting the effect of value

In Experiment 3, we revisit the question of whether preschoolers can consider value in their teacher evaluations. We use a modified between-subjects design, but with the same toy and functions. If children's ratings reflect the value of taught information in this simpler task, this would suggest that young children do possess the ability to attribute value in evaluations of pedagogy, but that the paradigm used in Experiment 2 might have been too demanding for this competence to be elicited.

6.1. Method

This study was approved by the Rutgers University – Newark Institutional Review Board, protocol 16–625Mc. Informed parental consent and child assent were obtained before the study was administered.

6.1.1. Participants

Participants were children recruited from and tested at local preschools, daycares, and zoos. A power analysis revealed that we would need roughly 20 participants in each condition to detect differences with 80% power. Therefore, our final sample consisted of 40 preschoolers ($M(SD)_{age} = 59.4(6.04)$ months, range $= 47 - 71$ months; $N = 18$ female). An additional 16 children were dropped and replaced due to: giving all teachers the same rating ($N = 7$); failure to pass the rating scale training ($N = 6$); failure to pass the practice teacher rating ($N = 2$); or noncompliance ($N = 1$).

6.1.2. Procedure

The procedure was nearly identical to Experiment 2, with the following exception. Instead of rating four test teachers, children saw the practice teacher and then just one test teacher: Either the KA_THH teacher, or the KA_TLL teacher. Children were randomly assigned to one of these two between-subjects conditions ($N = 20$ in each condition).

6.2. Results

6.2.1. Model fit

Given the small number of conditions, we opted to model the three conditions together, averaging the KA_TA results across the two between-subjects conditions. The fit between the model and behavioral data were again very high ($R^2 = 1.00$). The best-fit value for the utility difference was higher than it was for all previous experiments at 0.23. For communication cost, the best-fit value was similar to previous experiments, 0.86.[7] See Fig. 7B for modeling results.

6.2.2. Behavioral results

Children who saw the teacher who demonstrated two high-value functions rated their test teacher significantly higher ($M = 13.5$, $SD = 2.72$) than children who saw a teacher present two low-value functions ($M = 9.63$, $SD = 5.21$); $t(38) = 2.95, p = 0.006, d = 0.93$. See Fig. 7A.

6.3. Discussion

In addition to the high model fit, behavioral analyses showed that children were able to discriminate between teachers who provided demonstrations of differing values in this simpler, between-subjects design. While these results cannot speak to why we found discrepancies in adults' ratings between Experiments 1 and 2b, they do align with the possibility that the task in Experiment 2a was too demanding for preschoolers to be able to track every component of the sampling process, and thus the effect of value on children's ratings was washed out under these circumstances.

7. Experiment 4: replication

In Experiments 2a and 3, we found rather surprising success in preschool-aged children's evaluation of teachers who commit sins of omission: Despite the fact that all but one teacher omitted relevant function(s) of a toy, children evaluated them differently depending on the number and value of omissions as well as the teacher's epistemic state. Given the importance of these findings, we wanted to assess the

[7] There were three pairs of parameter values that resulted in the model perfectly fitting the data. However, we believe this was a result of fitting our model to so few conditions, and so we only report the pair that most closely mirrors what we found in previous experiments. See Appendix B for additional details.

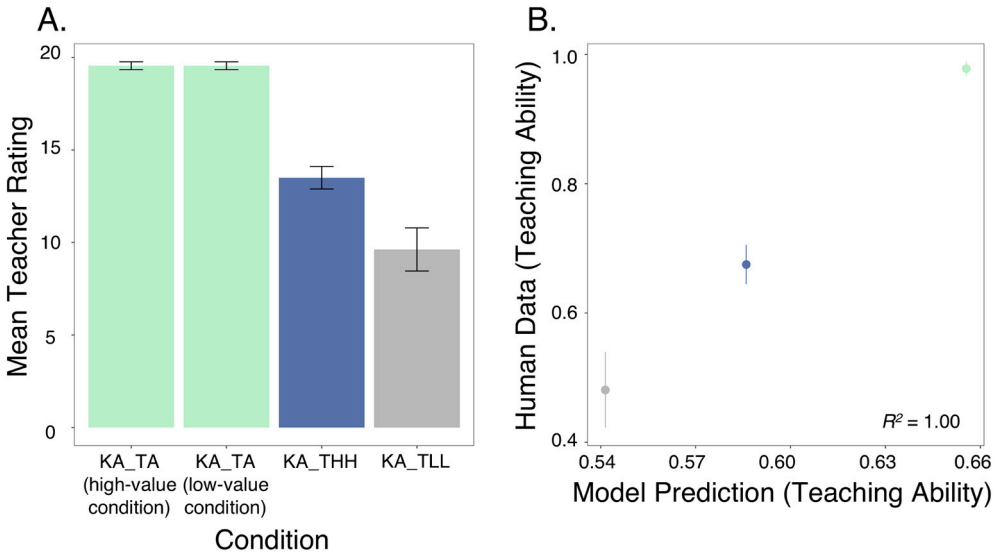

Fig. 7. A. Preschoolers' teacher ratings in Experiment 3. In a between-subjects design, a teacher showing high-value functions was rated as significantly better than a teacher showing low-value functions. B. Comparison of behavioral data and model predictions. The fit between the model and children's ratings was high ($R^2 = 1.00$).

robustness of these results. Thus, in Experiment 4, we ran a direct replication of Experiment 2a. This experiment was pre-registered through AsPredicted.org: https://aspredicted.org/blind.php?x=rs2xg9.

7.1. Method

This study was approved by the Rutgers University – Newark Institutional Review Board, protocol 16–625Mc. Informed parental consent and child assent were obtained before the study was administered.

7.1.1. Participants

Participants were children recruited from and tested at local preschools, daycares, and museums. A power analysis using the effect sizes from Experiment 2a revealed that 24 participants would be sufficient for detecting differences between teachers with 80% power. Therefore, our final sample consisted of 24 preschoolers ($M(SD)_{age} = 58.9(3.07)$ months, range $= 51 - 63$ months; $N = 10$ female). An additional 19 children were dropped and replaced due to: failure to pass the rating scale training ($N = 10$); placing the teachers in descending order as they saw them ($N = 6$); rating the test teachers higher than the practice teacher ($N = 1$); giving all teachers the same rating ($N = 1$); or terminating the study early ($N = 1$).

7.1.2. Procedure

The materials and procedure were identical to Experiment 2a.

7.2. Results

7.2.1. Model fit

Fitting the model to the data in this experiment yielded relatively worse fit compared with previous experiments. When KA_TA was included, the fit was reasonably high ($R^2 = 0.838$) but without KA_TA, the correlation was in fact negative (Pearson $R = -0.96$; $R^2 = 0.914$). The best-fitting parameter for the utility difference was 0.01 when the KA_TA condition was included, and 0.11 when it was excluded; for the communication cost parameter, the best-fit values were 0.95 when including the KA_TA condition, and 0.01 when excluding it. Thus, when

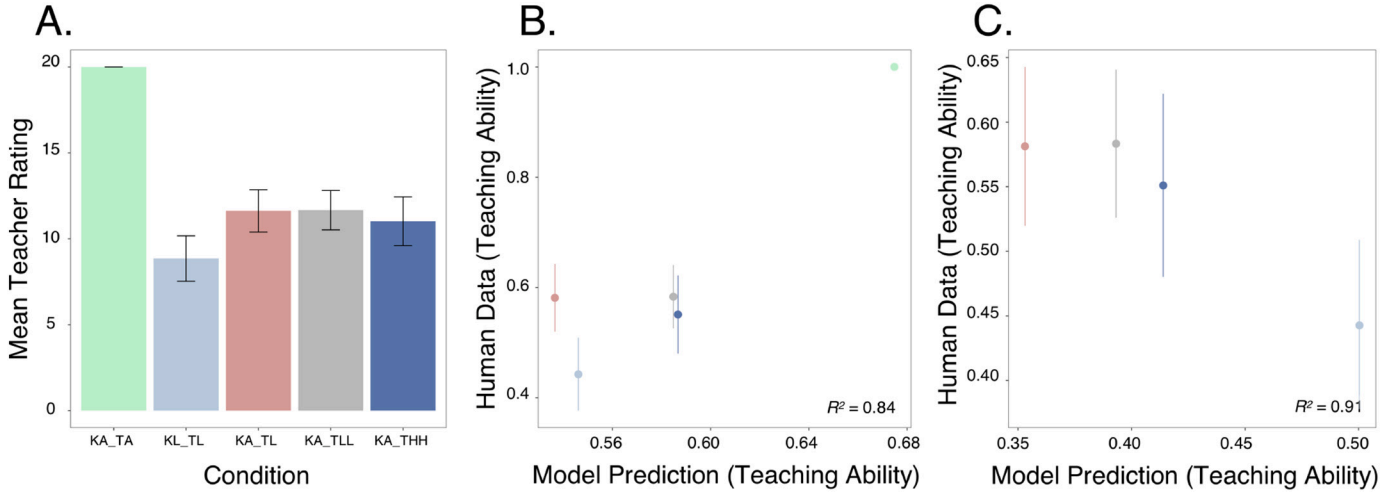

Fig. 8. A. Preschoolers' teacher ratings in Experiment 4 (replication of Experiment 2a). Children's ratings were not significantly different across teachers in this sample. B & C. Comparison of behavioral data and model predictions when including and excluding the KA_TA condition respectively. Model fit was relatively good when KA_TA was included (B), but poor when it was excluded (C).

modeling the KA_TA condition, we find very small utility differences between knowing a high- versus a low-value function, but high communication costs; when modeling only the test conditions, the opposite pattern emerges. See Figs. 8B & C for modeling results. Supplemental figures can be found in Appendix B.

7.2.2. Behavioral results

As suggested by the poor model fit, a repeated-measures ANOVA on ratings in the four test trials did not yield significant results ($p = 0.23$). Therefore, we did not perform follow-up pairwise comparisons between conditions. See Fig. 8A for a summary of these results.

7.3. Discussion

In sum, we were unable to successfully replicate the behavioral effects initially found in Experiment 2a. It is possible that the failure is due to potential differences across experiments. Although we maintained most aspects of the task from Experiment 2a, small differences – such as changes in the experimenter administering the task, or the environment in which testing took place – might have contributed to inconsistent findings across samples. However, the failure may also indicate that the effect sizes we observed in Experiment 2a were actually larger than the underlying true effect sizes, or even a false positive; after all, we found these initial results rather striking given that preschool-aged children still experience difficulty in basic Theory of Mind tasks, and chose to replicate these effects using an independent sample. Thus, we take seriously the possibility that the original findings are fragile and preschool-aged children may not yet exhibit robust sensitivity to these factors. As discussed in the Introduction (Sections 1.1–1.3), concurrent development of the cognitive capacities that likely support these judgments may make it difficult for young learners to integrate these factors into their informant evaluations until later in childhood. And indeed, in piloting this task after our failed replication, we anecdotally did not start to see evidence of more consistent success until age 7. Therefore, in Experiment 5, we use the same method to run the study once again, but with a group of slightly older children.[8]

8. Experiment 5: second-graders' teacher evaluations

In Experiment 5, we administered the same task used in Experiments 2 and 4, this time with a group of seven- and eight-year-old children. While our child-friendly method was unsuccessful in eliciting reliable teacher evaluations from preschool-aged children, we suspect that we might have more success with older children for several reasons. First, children in this age range may be more attuned to how information utility can shape decisions about what to teach (Bridgers et al., 2020). Further, children's ability to use intentions to evaluate others' behaviors – in particular, forgive accidental transgressions – develops at least through age seven (Fu, Xiao, Killen, & Lee, 2014; Killen et al., 2011). Thus, second-graders may also be more likely to be able to detect and pardon omissions resulting from limited knowledge. This experiment was pre-registered through AsPredicted.org: https://aspredicted.org/blind.php?x=fs3dn3.

[8] As we discuss later in the General Discussion (Section 9.1), directly tying teacher evaluations to individual differences in capacities like Theory of Mind reasoning and utility judgments, above and beyond effects of age, will be a critical direction for future work. Individual difference analyses were beyond the scope of the pre-registered experiments run in this paper for practical reasons (although see Experiment 5 for a foray into this approach), and would not directly test the predictions of our model per se, which was the main goal of this paper.

8.1. Method

This study was approved by the Rutgers University – Newark Institutional Review Board, protocol 16-625Mc. Informed parental consent and child assent were obtained before the study was administered.

8.1.1. Participants

Participants were children recruited from and tested at local museums and community events. Our final sample consisted of 24 second-graders ($M(SD)_{age} = 94.0(6.83)$ months, range = 85 – 106 months; $N =$ 10 female). An additional 11 children were dropped and replaced due to: placing the teachers in descending order as they saw them ($N = 6$); failure to pass the practice teacher rating ($N = 2$); rating the test teachers higher than the practice teacher ($N = 1$); terminating the study early ($N = 1$); or being outside of our target age range ($N = 1$).

8.1.2. Procedure

The materials and procedure were identical to Experiment 2a.

8.2. Results

As in previous experiments, we provide the model fit results first and then provide results from confirmatory analyses.

8.2.1. Model fit

We fit the model to children's ratings in the same way that we did for the previous experiments. We were able to achieve excellent fit to the data, both when ratings of the KA_TA teacher were included ($R^2 =$ 0.996), and when they were excluded ($R^2 = 1.00$). Here, best-fit parameters for the utility difference were 0.06 (KA_TA included) and 0.09 (KA_TA excluded), and the value for the communication cost was 0.85 in both cases. See Figs. 9B & C for modeling results. Supplemental figures can be found in Appendix B.

8.2.2. Behavioral results

As we did in the previous experiments, we first ran a repeated-measures ANOVA on all four test conditions. This analysis was significant: $F(3, 69) = 5.89$, $p = 0.001$, $\eta_p^2 = 0.204$. The follow-up pairwise comparisons were therefore performed, to examine the specific effects of number, value, and knowledge state. See Fig. 9A for a summary of these results.

8.2.3. Number

Children's ratings in this age group did reflect the degree of the teacher's omission: The teacher who showed just one function (KA_TL: $M = 7.5$, $SD = 5.1$) was rated significantly lower than the teacher who showed two (KA_TLL: $M = 10.69$, $SD = 3.78$); $t(23) = 4.81$, $p < 0.001$, $d = 0.98$. Additionally, 79% of participants gave higher ratings to the teacher who demonstrated two functions, which was significantly greater than chance ($p = 0.003$, one-tailed).

8.2.4. Value

In this older group of children, we also found a significant effect of value. The KA_TLL teacher, who taught the low-value functions, was rated significantly lower than the KA_THH teacher, who taught the high-value functions ($M = 11.85, SD = 4.20$); $t(23) = 2.04$, $p = 0.053$, $d =$ 0.42. Only 54% of children rated the high-value teacher more favorably than the low-value teacher – which did not differ from chance ($p = 0.42$, one-tailed).

8.2.5. Teacher's knowledge state

Although children did rate the KL_TL teacher ($M = 9.5$, $SD = 6.28$) higher than the KA_TL teacher (which is the direction we would expect if children were sensitive to the teacher's knowledge state in the manner predicted by the model), this difference was not statistically significant

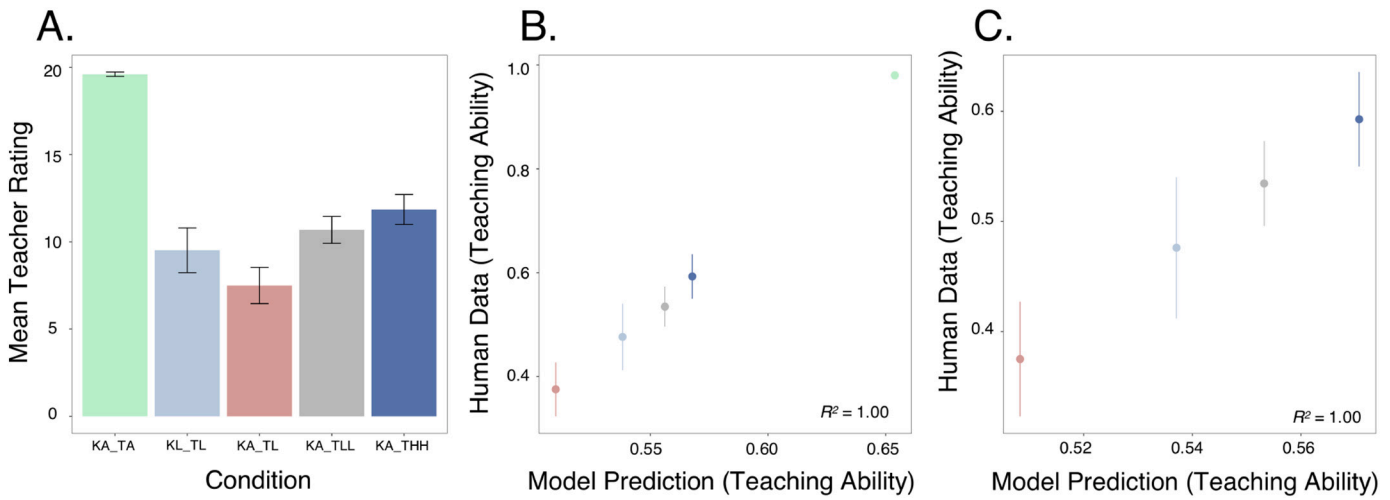

Fig. 9. A. Second-graders' teacher ratings in Experiment 5. Participants significantly differentiated teachers based on number and value, but not knowledge state. B & C. Comparison of behavioral data and model predictions when including and excluding the KA_TA condition respectively. In both cases, the model fit between participants' ratings and the model predictions were very high (B: $R^2 = 0.996$, C: $R^2 = 1.00$).

($p = 0.17$). However, we explore this analysis further below.

8.2.6. Exploratory analysis of explanations

Upon coding the data for analysis, we noticed that although they were not prompted to, many of the children in this sample provided spontaneous explanations as they were making their ratings. This was not the case for the preschoolers we tested in the previous experiments. In past work, coding and analysis of children's verbal explanations for events has been used as a way of potentially tapping into their rich causal explanatory cognitive models (Amsterlaw & Wellman, 2006; Goodman et al., 2006; Hickling & Wellman, 2001). Therefore, as part of a series of post-hoc analyses, we coded children's explanations for mentions of the variables made explicit in our computational model. In addition to our three factors of interest (*f*, what the teacher knew; e.g., "She didn't know any of the other buttons."; *d*, what the teacher showed or didn't show; e.g., "She only showed one thing!" or "She left out two of them."; *V*, the value of the teacher's demonstration; e.g., "She showed the ones I like."), we also coded for mentions of cost ($C(d)$; e.g., "Maybe she didn't have time to show them all?"), the teacher's perceived quality (α; e.g., "I think she was really good."), and miscellaneous variables (explanations that did not fall into any of the above categories).

First, we asked whether children were more likely to provide spontaneous explanations that appealed to knowledge state when rating the teacher whose knowledge state was limited (KL_TL). Of the 12 children who provided explanations for this teacher, 8 of them invoked knowledge state (only 2 of 12, 1 of 12, and 2 of 10 children did so in the KA_TL, KA_TLL, and KA_THH conditions, respectively). As an exploratory analysis, we compared these proportions in the KL_TL condition to all of the other conditions collapsed (i.e., 8/12 vs. 5/34) by Fisher's exact test, which yielded significant results ($p = 0.001$, two-tailed), suggesting that children were more likely to spontaneously reference the teacher's knowledge state when she had partial knowledge than when the teacher had full knowledge.[9]

Next, we asked whether differences in ratings between the two teachers who provided the same demonstration, but possessed disparate knowledge (KL_TL and KA_TL), were related to children's propensity to appeal to knowledge state in their spontaneous explanations. We found that children who did mention knowledge state ($N = 8$) in reference to the KL_TL teacher successfully differentiated between these two teachers in their ratings: The KL_TL teacher ($M = 12.38$, $SD = 7.80$) was rated significantly more favorably than the KA_TL teacher ($M = 4.13$, $SD = 4.13$); $t(7) = 2.58$, $p = 0.036$, $d = 0.913$. This was not the case for children who did not invoke knowledge state in their explanations for their ratings of the KL_TL teacher ($N = 16$; $p = 0.146$). More specifically, the difference in ratings between these two teachers (KL_TL - KA_TL) was significantly greater for knowledge state explainers ($M(SD)_{dif} = 8.25(9.04)$) than it was for non-knowledge state explainers ($M(SD)_{dif} = -1.09(2.86)$); $t(22) = 3.84$, $p < 0.001$, $d = 1.39$. See Fig. 10 for a summary of these results.

It could be the case that these children who spontaneously mentioned knowledge state were simply older, or more competent at detecting differences in teachers in general. However, knowledge state explainers ($M(SD)_{age} = 93.27(5.38)$) were not significantly older than non-knowledge state explainers ($M(SD)_{age} = 94.34(7.59)$; $p = 0.73$). We also compared the degree to which value and number were detected and evaluated by these two groups. There was no effect of value: Knowledge state explainers did not discriminate between the KA_THH

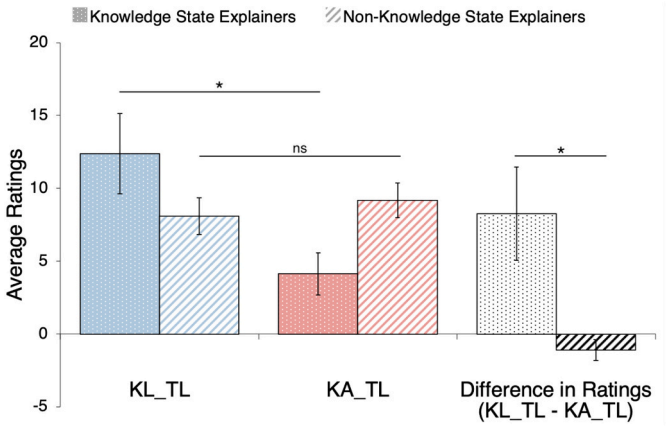

Fig. 10. Second-graders' evaluations of the KL_TL and KA_TL teachers in Experiment 5. Children who invoked knowledge state as a causal explanatory variable for the KL_TL teacher rated her better than the KA_TL teacher, whereas those who didn't did not differ in their ratings for these two teachers. Difference scores between the KL_TL and KA_TL teachers were significantly greater for children who spontaneously appealed to knowledge state.

[9] We did not find this kind of pattern for explanations invoking the other coded variables. For example, the number of children who appealed to the amount of information demonstrated across conditions was as follows: KL_TL = 10 of 12; KA_TL = 9 of 12; KA_TLL = 9 of 12; KA_THH = 7 of 10.

and KA_TLL teachers any more than the non-knowledge state explainers did ($p = 0.31$). We did find a marginal effect of number, such that the difference in ratings between the KA_TLL teacher and the KA_TL teacher was slightly, but nonsignificantly, higher for knowledge state explainers ($M(SD)_{dif} = 4.88(1.73)$) than it was for non-knowledge state explainers ($M(SD)_{dif} = 2.34(3.53)$); $t(22) = 1.9$, $p = 0.071$, $d = 0.914$. It is worth noting, however, that both groups of children rated the teacher who showed two functions as significantly better than the teacher who showed just one ($ps \leq 0.018$). Qualitatively, this is different from the analysis described above, where *only* the knowledge state explainers rated the KL_TL teacher significantly higher than the KA_TL teacher. Thus, one way to interpret our overall findings in these exploratory analyses is that children's explanations – while limited – are providing some insight into the variables that they are factoring into their teacher evaluation. Additional interpretations can be found in Section 8.3.

8.3. Discussion

Given the failure to replicate our original findings in Experiment 4, in Experiment 5 we sought to replicate the results with a sample of older children. First, we found clear effects of both number and value. It is noteworthy that information value was reflected in second graders' teacher evaluations, even in this within-subjects design; perhaps the ability to integrate multiple components into evaluations of pedagogy becomes less cognitively demanding with age. Indeed, older children's success here is consistent with the idea that adults failed to distinguish teachers based on value not because they were incapable of doing so, but because they did not believe this toy's functions provided a basis for differential evaluations of the teachers demonstrating them. This raises interesting open questions about differences in perceptions of pedagogical utility across development.

Our findings with respect to knowledge state were mixed: Only children who spontaneously appealed to knowledge state when explaining their ratings of the teacher who naïvely omitted information evaluated this teacher as better than one who knowingly omitted. This points to some interesting tentative conclusions. For instance, it could be the case that the ability to consider the teacher's knowledge state as a relevant variable for evaluating pedagogy comes relatively late in development, and with high variability across the preschool and early elementary school years. There are vast individual differences in the ability to reason about others' mental states (e.g., Cutting & Dunn, 1999), and past work in moral reasoning has found direct links between Theory of Mind reasoning skills and the ability to pardon accidental transgressors (Killen et al., 2011). Furthermore, understanding the relevance of knowledge state in these evaluations and generating appropriate alternative behaviors (i.e., what the teacher could have shown given their knowledge) might also develop throughout preschool and early school years (Gweon & Asaba, 2018). Might individual differences in underlying cognitive capacities have influenced the degree to which children's ratings reflected the factors in our model? Our data cannot directly speak to this question; nevertheless, the current findings lead to interesting hypotheses that could be tested in future work – some of which are discussed at greater length in Section 9.

9. General discussion

Effective social learning requires sophisticated inferences both about the meaning of pedagogically sampled evidence, and about the person who is doing the pedagogical sampling. Here, we formalize these inferences with a Bayesian computational model of teacher evaluation that integrates information about the degree of omission, the value of demonstration, and the teacher's knowledge state. We tested our model predictions in several independent samples of adults and children across five experiments. Intuitions about the factors that influence pedagogical evaluations were supported both by our model and adults' ratings of teacher quality (Experiments 1 & 2b). These abilities, however, may be tenuous, and perhaps difficult to capture in preschoolers (Experiments 2a, 3, & 4); by second grade, while children in our study had the ability to successfully distinguish teachers based on the amount and value of what was demonstrated, their ability to evaluate omissions relative to the teacher's knowledge state was related to their tendency to spontaneously appeal to this variable in their own explanations (Experiment 5).

Broadly speaking, this work represents two novel contributions to the literature. First, we extend past work on learners' sensitivity to "sins of omission" (Gweon & Asaba, 2018; Gweon et al., 2014) to new cases where the key variables (that were held constant in previous work) were varied systematically. This suggests that people's evaluations of pedagogical demonstrations are far more than simple enumerations of the amount of information omitted; rather, people consider a number of factors to make nuanced and graded judgments about the quality of teaching. Second, by presenting a single Bayesian computational model of teacher evaluation that considers the utility of communicated information given the teacher's prior knowledge, our findings extend a broader class of utility-theoretic accounts of social cognition, such as pragmatic understanding in linguistic communication (Goodman & Frank, 2016), interpretation of goal-directed actions (Jara-Ettinger et al., 2016), as well as pedagogical decision-making (Bridgers et al., 2020). Together, these contributions begin to offer insight into our understanding of the cognitive processes that give rise to these kinds of social evaluative abilities. In particular, this work highlights the ability of children to apply social judgments flexibly: Omitting any information is unhelpful, but omitting less is not quite as blameworthy; sharing uninteresting information when one could have demonstrated something engaging at the same communication cost is noticed and penalized; and identical demonstrations are evaluated relative to what *could* have been shown, given the teacher's own prior knowledge. By comparing the model predictions with children's responses, our work also represents a particular theoretical perspective on the development of these abilities. Rather than appealing to a set of heuristics or cue-sensitivities, our approach aims to explain children's evaluations with respect to an abstract, causal model of social evaluation that assesses the utility of information a teacher provides to the learner (at least within contexts in which more, and more interesting, information is considered to constitute higher-utility teaching).

9.1. Developmental findings & implications

Although our findings from older children are broadly consistent with our hypotheses, the overall pattern of developmental findings were rather mixed, particularly in preschool-aged children. Across experiments, there were limitations both in our design and in our results that are worth discussing, and may point to promising directions for future research. For one, Experiment 1 conceptualized differences in information value as the degree to which the functions conveyed useful information, whereas all other experiments used a value manipulation that relied on children's perceptions of which functions were "cooler". Intuitively, this former operationalization seems as though it should be more relevant for evaluations of pedagogy than the latter – and indeed, discrepancies in adults' sensitivity to value across experiments (i.e., Experiment 1 vs. Experiment 2b) reflect this intuition. In contrast, children's teacher evaluations did incorporate perceptual value – but preschoolers only did so in a simple, between-subjects design

(Experiment 3), whereas older children's evaluations reflected value even when contrasting multiple teachers.

While our inconsistent manipulation of information value across experiments limits the claims we can make about how this factor is generally integrated into evaluations of pedagogy, it also leads to interesting questions about the relevant dimensions for teacher evaluation across development. It could be that something adults regard as an important element of teacher quality is not something that children necessarily consider – and vice versa. Do children also consider the informativeness of the functions (i.e., epistemic value) as a relevant dimension for teacher evaluation? How might these differences shape the kinds of teachers that are sought out across development? These are open questions that would be important directions for future work.

Additionally, we were unable to replicate our initial findings with preschoolers (Experiment 2a) in a second sample (Experiment 4). Thus, although it is unclear from the current results, it still remains possible that the sophisticated teacher evaluation abilities seen in adults are within the capabilities of preschool-aged children, albeit in a fragile way. In particular, while preschoolers may possess the underlying competencies required for performing flexible, integrated evaluations of teachers' quality, whether this manifests in their ratings may be contingent on a variety of factors. Perhaps children were sensitive to subtle pragmatic variations across experimenters that led them to detect differences between teachers in Experiment 2a, but not in Experiment 4. Maybe there were other underlying differences between the children in these two samples that resulted in differential sensitivity to differences between teachers – for instance, different kinds of experience with pedagogy in home environments, leading to disparities in conceptualizations of what constitutes "high-utility" teaching (Yu et al., 2018; Yu, Shafto, & Bonawitz, 2020). Of course, this is merely speculation. Whatever the case may be, our results suggest that the effect sizes in preschool-aged children are likely small, and with high variability. This raises an important general point: In order to make any strong claims about preschool-aged children's competencies in these kinds of tasks more broadly, testing large, diverse samples will be crucial in accounting for this variability. Indeed, understanding whether preschool-aged children's evaluations of pedagogy reflect the same integrated process observed in adults is a key direction for future work. The preschool years represent a critical time to investigate pedagogical reasoning, because children of this age have not yet been exposed to much structured classroom learning typical of formal schooling. As such, preschoolers' teacher evaluation abilities may represent a more intuitive understanding of informal pedagogy (e.g., Csibra & Gergely, 2009; Gerstenberg & Tenenbaum, 2017; Gopnik & Meltzoff, 1997; Shafto, Goodman, et al., 2012; Wellman & Gelman, 1992) than we may be tapping with older children and adults.

The post-hoc analyses from Experiment 5 also raise the importance of considering individual differences in how we interpret developmental findings. In particular, although the current work cannot directly address this hypothesis, our results are consistent with the idea that children's tendency to explicitly appeal to the variables in our model related to and may have even influenced their ratings. Future work could directly test these open questions by actually collecting individual difference data on relevant measures (e.g., Theory of Mind, executive function, utility judgments), or explicitly asking children to provide causal explanations for their ratings. Along with the larger sample sizes and additional test measures that would be required of this kind of analysis, these approaches might elucidate the extent to which different children are invoking different causal variables in their own models of teacher evaluation. Understanding how particular cognitive capacities support an integrated process for evaluating the quality of teachers may provide additional insight into the development of these abilities, above and beyond simple success or failure on this task.

9.2. Computational findings & implications

In addition to considering the ways in which our behavioral results may open directions for future work, it is also important to examine the ways in which our modeling work may be extended. By and large, the model was able to capture our behavioral data exceptionally well. In many of our experiments, the model was fit to only a few conditions, which might partially explain how we were able to obtain perfect fits. But importantly, we believe this also speaks to our experimental manipulations being set up in such a way that a clean linear pattern in ratings across conditions *could* emerge. Further, the best-fit values for our model's two free parameters (the utility difference, and the communication cost) were similar – albeit not identical – across several of the experiments. How and why these parameter weights shift across development and under different circumstances are yet open questions. It might be an interesting endeavor to examine how accurately we could model all of our data jointly, using only one set of parameters values for both adults and children. On the other hand, it would also be informative to instead use empirical data for these parameter values, collected from adults and children separately. These kinds of ventures could shed additional light on how exactly communication costs and perceived utility factor into teacher evaluations across development, and why particular methodologies might elicit different kinds of ratings.

We acknowledge that our model by no means captures all of the sophistication and complexities that likely factor into gauging others' quality as teachers. Extensions to our model could lead to important future work. For one, we have assumed that the learner and the teacher have identical utility functions (i.e., what is valuable for the learner to know is what is valuable for the teacher to show). Indeed, this is often not true in the real world, particularly in cases where young children or students learn from adult teachers: There is often a discrepancy between what young learners *want* to learn versus what adult teachers *believe* is useful for the learners to know (Oscarsson, Jidesjö, Strömdahl, & Karlsson, 2009), and because children often do not get to choose what they are taught, they are taught what adults think is valuable for children to know, rather than what children they themselves think is valuable. Additionally, sometimes good teachers *repeat* demonstrations, especially if they think it is difficult or particularly valuable for learners to know, or if the learner seems to have missed it. Although our model assumes that learning given the teacher's demonstration is "perfect" (i.e., L0 deterministically infers the functions that the teacher demonstrated), one might imagine an extension of the model that incorporates whether a teacher takes into account the fidelity of the learner's updating process. Future computational and empirical work may be useful in formalizing what constitutes "good teaching" when individuals' utility functions do not align, or when the learner's updating process is compromised.

Additional relevant variables that were not included in our model, such as prior experience with particular teachers (Corriveau & Harris, 2009) and expectations about the style with which information is expected to be imparted (Bass et al., 2018), may be worth trying to incorporate in future formalizations. Further, an important next step may lie in directly comparing our model to alternatives, in order to understand whether our formalization *better* captures teacher evaluation abilities than other possibilities. In the same vein, it may be critical to compare our full model to cases where particular variables are "lesioned", and examining how this changes its efficacy in predicting human response.

9.3. Additional theoretical contributions & future work

This research connects to a growing literature charting the development of children's epistemic reasoning in pedagogical contexts. In

past work, children have shown competence in reasoning about how different sets of evidence may lead naïve learners to different conclusions (Rhodes, Bonawitz, Shafto, Chen, & Caglar, 2015). Further, young children understand that learners with different incorrect beliefs may need distinct evidence in order to be led to the correct conclusion, and the ability to select such belief-appropriate evidence is directly related to false-belief reasoning abilities (Bass et al., 2019). Finally, children, as teachers, are also capable of prioritizing information that maximizes the learner's utilities (Bridgers et al., 2020). Here, we extend this past work and examine not only how children are able to reason about pedagogically sampled evidence, but how they may use this observed evidence to make inferences about an unobservable quality of the person doing the sampling – namely, how effective they are at selecting samples for the purposes of teaching others. While prior research has investigated some aspects of these abilities (Gweon & Asaba, 2018; Gweon et al., 2014), we build on these findings by testing the effects of nuanced yet important differences in several aspects of under-informative pedagogy across a wide age range. Overall, our results suggest that children are quite adept at reasoning about evidence from an early age, and are able to make a slew of rich social inferences from remarkably little data.

Finally, while the current work focuses on learning from others in informal teaching contexts, it is worth considering how this may be extended to classroom learning. We note that although our behavioral experiments allowed a precise manipulation of what was taught and what the teacher knew, there remains an open question about whether students in real-world pedagogical settings are also sensitive to these factors. Our participants were third-party observers of a teacher-student interaction, rather than the students themselves; and the observer (L1) in our model had true beliefs about the teacher's knowledge state f_T and the functionality of the toy, while the teacher's target of demonstration (L0) was fully naïve to both. This raises the question of how learners might evaluate teachers when they themselves are privy neither to the target hypothesis being taught, nor to the teacher's prior knowledge. Ongoing work suggests that as learners, both young children and adults may be sensitive to other factors that influence decisions about pedagogical sampling. For example, the degree to which a teacher is expected to be fully informative rationally shapes learners' inferences both about the amount of information there is to learn, as well as the importance of demonstrations, when that teacher provides new information in a pedagogical context (Bass et al., 2018). Related lines of research have suggested that drawing learners' attention to relevant features of a learning problem using *pedagogical questions* (i.e., questions that a knowledgeable teacher asks a learner as a means of teaching them) is as effective as direct instruction at transmitting target knowledge, but better than direct instruction at fostering further exploration (Yu et al., 2018; see also Yu, Bonawitz, & Shafto, 2019). This brings up important questions about the kinds of inferences learners themselves may make from teaching that occurs within the context of the classroom – and indeed, research in education has found that some teaching styles may better support students' learning than others (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011; Fisher, Hirsh-Pasek, Newcombe, & Golinkoff, 2013; Honomichl & Chen, 2012). In light of our current findings, this past work is consistent with the idea that both as observers and as learners, children integrate a variety of factors into their evaluations of the reasons behind an teacher's pedagogical selections. Future work may directly ask whether the particular factors investigated here (number, value, and knowledge state) are also considered by learners evaluating teachers in more realistic, live classroom interactions.

10. Conclusions

As learners, we often find ourselves dealing with uncertainty both about the world, as well as the degree to which we can trust others to teach us about it. Indeed, violations of pedagogical sampling occur frequently in real-world learning contexts: A well-intentioned, helpful teacher might provide insufficient information because she did not possess all the relevant knowledge; a fully knowledgeable teacher might omit information because she thought it was not worth teaching. How real-world learners exploit diverse sources of information to simultaneously learn about other people and about the world is an incredibly rich and exciting area for future computational and empirical research. Here, we take an important step towards delineating the factors that we consider in deciding from whom it will be best to learn. Together, this work adds to the growing body of literature on children's developing ability to draw rich inferences from others' pedagogical demonstrations, and offers insight into how we may rationally evaluate others' quality as evidence selectors for the purposes of future learning.

CRediT author statement

All authors contributed in many capacities to this work, such that we do not feel it is appropriate to delineate CRediT contributions at this time.

Declarations of interest

None.

Acknowledgements

We thank Shirley Abbelard, Brianna Ali, Grace Bennett-Pierre, Haneen Daas, Lauren Leotti, and Gabrielle Vicente for their assistance with data collection and coding. This work was supported by: the Varieties of Understanding grant from the John Templeton Foundation to HG; a Jacobs Foundation Early Career Research Fellowship to HG; NSF SMA SL-CN (#1640816) to EB; a Jacobs Foundation Early Career Research Fellowship to EB; the Jerome Davis Research Fund at Oberlin College to IB; the Henry Rutgers Presidential Fellowship to IB; and an American Fellowship from AAUW to IB. Portions of this work were previously published in the Proceedings of the 37th and 39th Annual Conferences of the Cognitive Science Society. Supplemental materials, data, and stimuli can be found at https://osf.io/mpnr9/?view_only=10ae4642a647435c83e6ae18cfaeb905.

Appendix A

Full modeling materials can be found at the following links:

- https://osf.io/mpnr9/files/?view_only=10ae4642a647435c83e6ae18cfaeb905, in the "Modeling materials" subfolder; see README.md for an overview of all of the materials in this subfolder.

- https://github.com/forestdb/forestdb.org/blob/gh-pages/models/inferring-teachers-skill.md, which contains a description and webchurch implementation of the model from the portion of this work that was previously published in the Proceedings of the 37th Annual Conferences of the Cognitive Science Society.

A.1. Parameter fitting

Our model has two free parameters: the difference in the utility of knowing a high- versus a low-value function, and the cost the teacher incurs by communicating a function. Each of these parameters could range from 0 to 1. To find the parameter values that led to the best fit between the model and our data, we performed a grid search over values from 0 to 1, by increments of.01, for each parameter. The best-fit values were those that produced the highest value for R^2, which was simply calculated by squaring the Pearson correlation between the behavioral data and model predictions by experimental condition.

A.2. Additional notes

A.2.1. Prior on α

We assume $p(\alpha)$&Uniform$(0,1)$. However, for implementation purposes, we discretized $p(\alpha)$ into 11 bins ranging from 0 to 1 in increments of.1, as opposed to using a continuous uniform distribution.

A.2.2. Utility difference

The utility difference parameter reflects the difference in the between high-value functions and low-value functions. To quantify this, the utility difference was symmetrical around a value of.5; specifically, high-value functions were assigned a utility of $0.5 + (\textit{utility\_d ifference}/2)$, and low-value functions had utility $0.5 - (\textit{utility\_d ifference}/2)$. Thus, a utility difference of 0 would mean that both high- and low-value functions have equal utility of.5; in contrast, a utility difference of 1 suggests that low-value functions have 0 utility, whereas high-value functions are maximally useful.

Appendix B

In what follows, we provide supplemental plots from the modeling results for each experiment. In particular, for each experiment we include heat-map plots for the best-fit free parameter values. Encouragingly, these plots reveal generally unimodal clusters for the best-fit parameter values. An exception to this pattern lies in Experiment 3 – see below for details.

We also include the means plot, heat-map plot, and comparison of behavioral data to model predictions for the dataset including those participants who failed one or more the the check questions that served as exclusion criteria in Experiment 1.

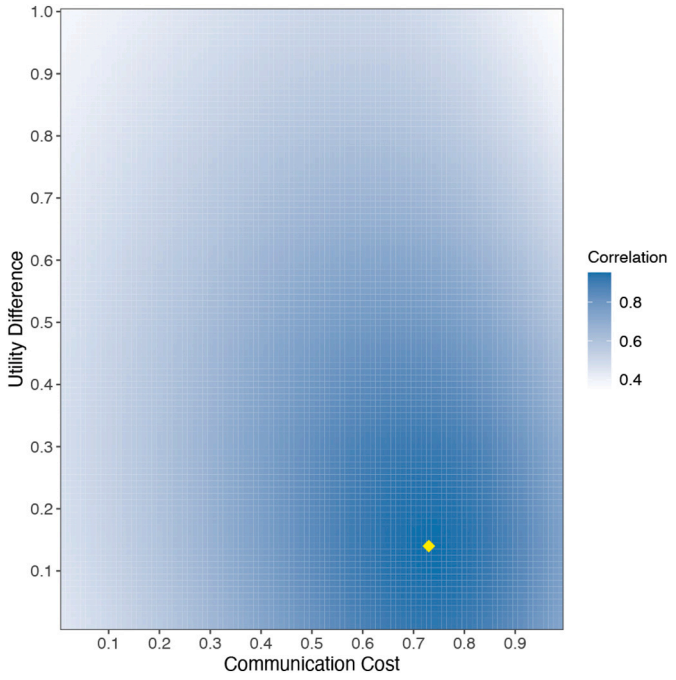

Fig. 11. Free parameter heat-map plot for modeling results from Experiment 1 (adults, $N = 1168$).

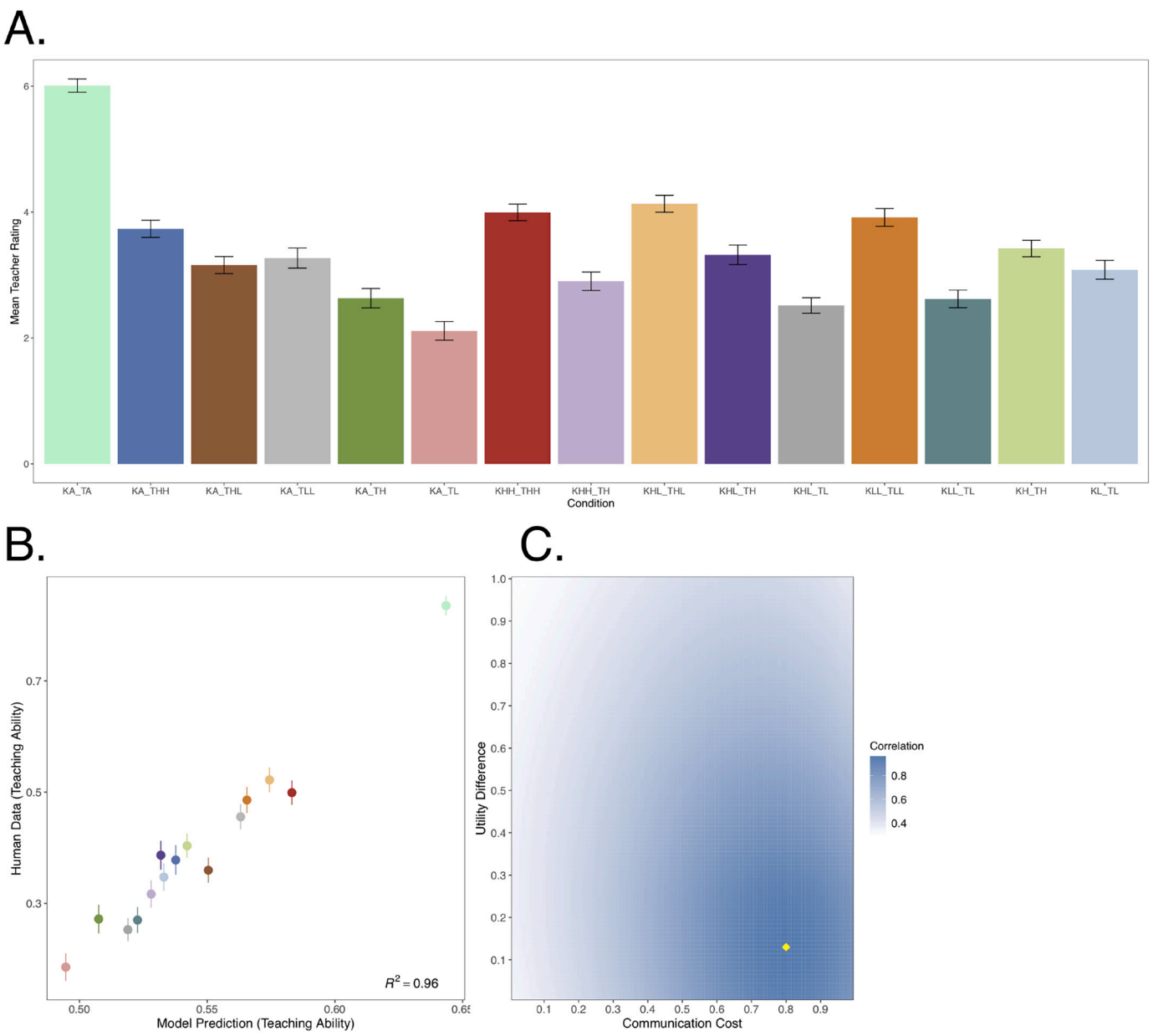

Fig. 12. Means plot (A), model fit (B), and heat-map plot (C) for participants who failed one or more the the check questions that served as exclusion criteria in Experiment 1 (adults, $N = 1486$). Patterns of results are nearly identical to analyses performed using only participants who passed all of the check questions ($N = 1168$).

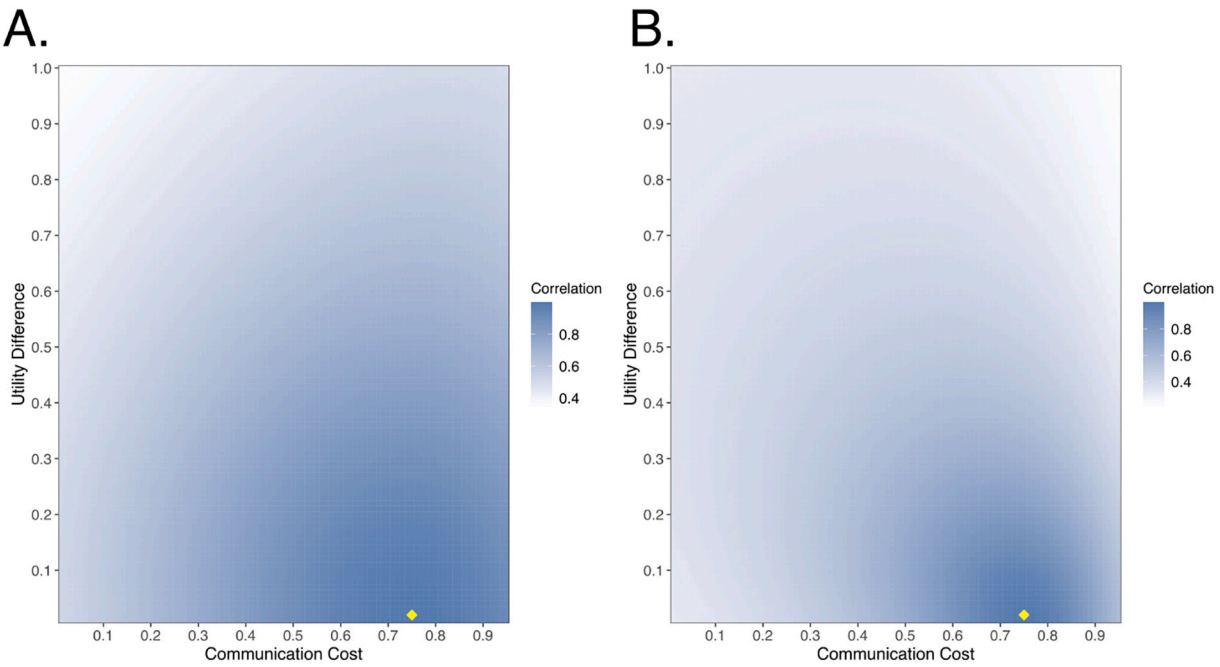

Fig. 13. Free parameter heat-map plot for modeling results from Experiment 2a, both including (A) and excluding (B) the KA_TA condition (preschoolers, $N = 24$).

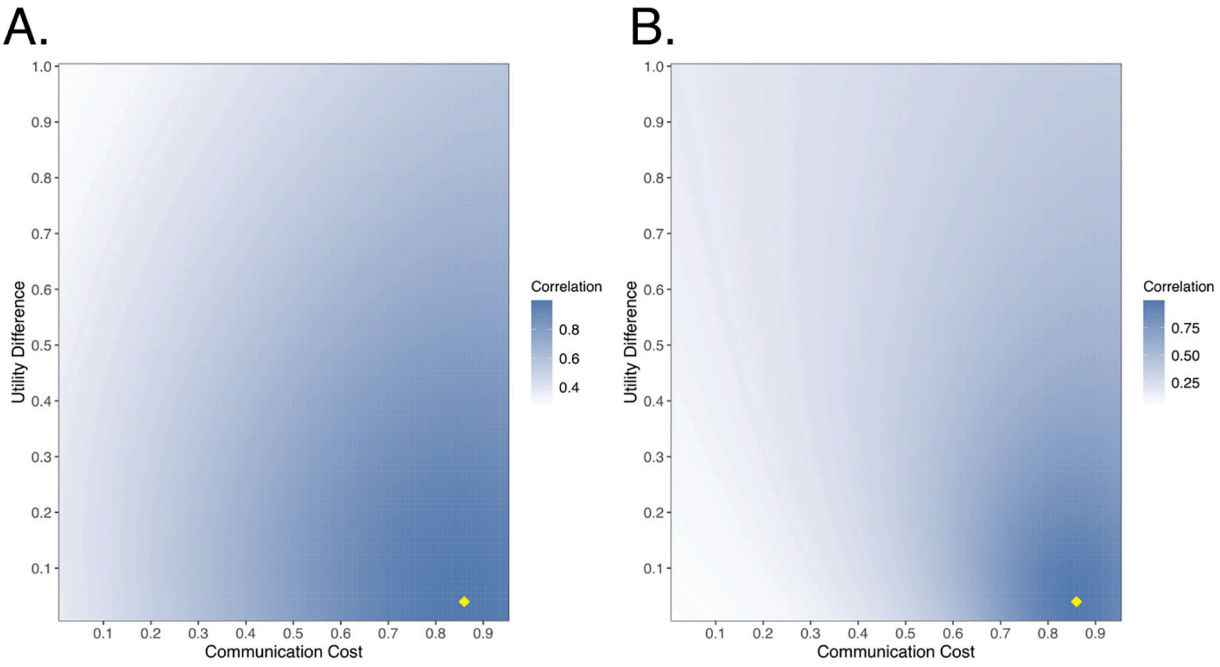

Fig. 14. Free parameter heat-map plot for modeling results from Experiment 2b, both including (A) and excluding (B) the KA_TA condition (adults, $N = 24$).

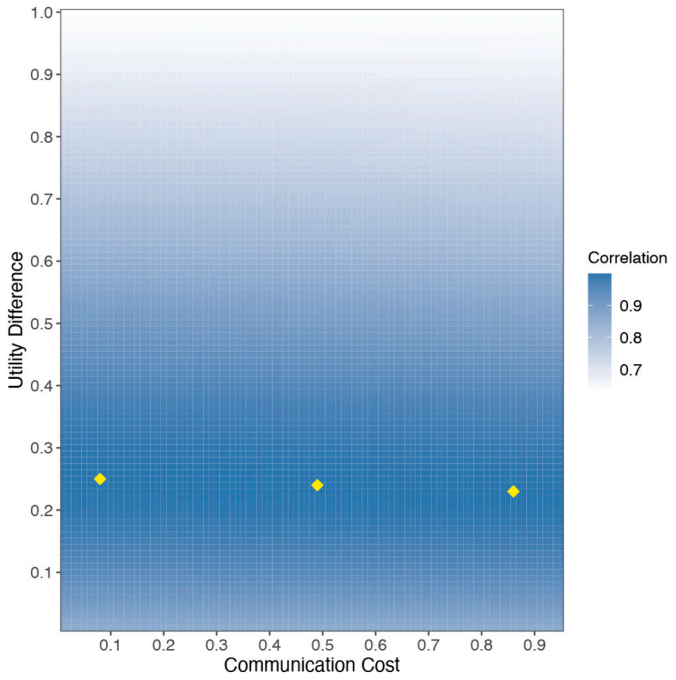

Fig. 15. Free parameter heat-map plot for modeling results from Experiment 3 (preschoolers, $N = 40$). Here, we found three pairs of parameter values that all resulted in the model fitting the data perfectly ($R^2 = 1.00$): (1) communication cost = 0.08, utility difference = 0.25; (2) communication cost = 0.49, utility difference = 0.24; (3) communication cost = 0.86, utility difference = 0.23. The heat-map plot clarifies this result: We find good fit across several values of communication cost, but with the utility difference hovering closely around.24. Intuitively this makes sense, given that the two test teachers rated in this experiment presented the same number of functions, making it difficult to infer what exactly the cost might be of demonstrating any particular function.

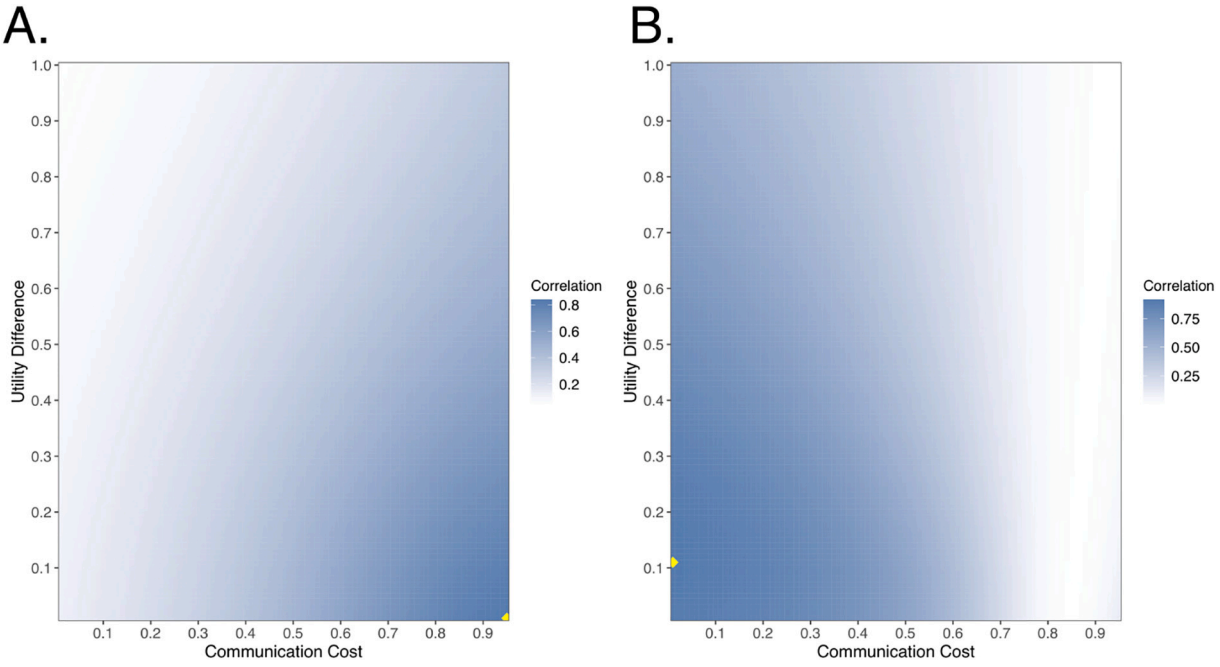

Fig. 16. Free parameter heat-map plot for modeling results from Experiment 4, both including (A) and excluding (B) the KA_TA condition (preschoolers, $N = 24$).

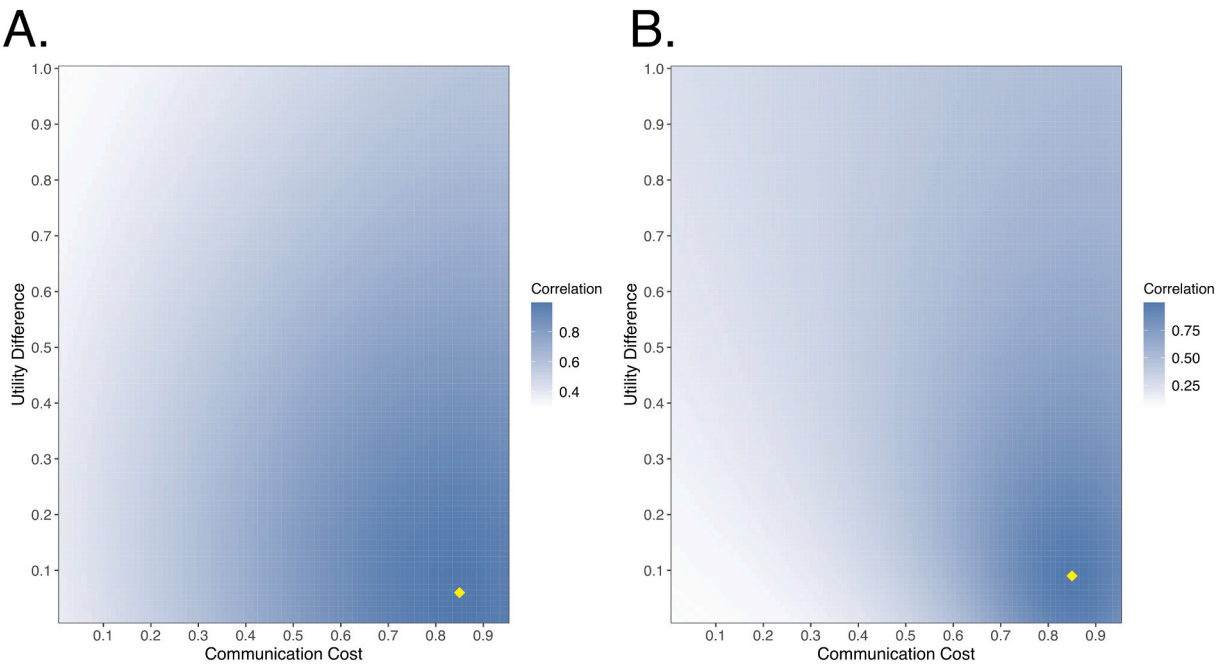

Fig. 17. Free parameter heat-map plot for modeling results from Experiment 5, both including (A) and excluding (B) the KA_TA condition (second-graders, $N = 24$).

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104999.

References

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*(1), 1–18.

Amsterlaw, J., & Wellman, H. M. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development, 7*(2), 139–172.

Bass, I., Gopnik, A., Hanson, M., Ramarajan, D., Shafto, P., Wellman, H., & Bonawitz, E. (2019). Children's developing theory of mind and pedagogical evidence selection. *Developmental Psychology, 55*(2), 286–302.

Bass, I., Shafto, P., & Bonawitz, E. (2018). That'll teach 'em: How expectations about teaching styles may constrain inferences. In *Proceedings of the 40th annual conference of the cognitive science society*.

Birch, S. A. J., Vauthier, S. A., & Bloom, P. (2008). Three- and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition, 107*(3), 1018–1034.

Bonawitz, E., & Shafto, P. (2016). Computational models of development, social influences. *Current Opinion in Behavioral Sciences, 7*, 95–100.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: instruction limits spontaneous exploration and discovery. *Cognition, 120*(3), 322–330.

Bonawitz, E., Shafto, P., Yu, Y., Gonzalez, A., & Bridgers, S. (2020). Children change their answers in response to neutral follow-up questions by a knowledgeable asker. *Cognitive Science, 44*(1), e12811.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2020). Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour, 4*(2), 144–152.

Corriveau, K., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science, 12*(3), 426–437.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13*(4), 148–153.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition, 127*(1), 6–21.

Cutting, A. L., & Dunn, J. (1999). Theory of mind, emotion understanding, language, and family background: Individual differences and interrelations. *Child Development, 70*(4), 853–865.

Eaves, B. S., & Shafto, P. (2017). Parameterizing developmental changes in epistemic trust. *Psychonomic Bulletin & Review, 24*(2), 277–306.

Fisher, K. R., Hirsh-Pasek, K., Newcombe, N., & Golinkoff, R. M. (2013). Taking shape: Supporting preschoolers' acquisition of geometric knowledge through guided play. *Child Development, 84*(6), 1872–1878.

Fu, G., Xiao, W. S., Killen, M., & Lee, K. (2014). Moral judgment and its relation to second-order theory of mind. *Developmental Psychology, 50*(8), 2085–2092.

Geraghty, K., Waxman, S. R., & Gelman, S. A. (2014). Learning words from pictures: 15- and 17-month-old infants appreciate the referential and symbolic links among words, pictures, and objects. *Cognitive Development, 32*, 1–11.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., Schulz, L., & Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the 28th annual conference of the cognitive science society*.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*(11), 818–829.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

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

Gweon, H., & Asaba, M. (2018). Order matters: Children's evaluation of under-informative teachers depends on context. *Child Development, 89*(3), e278–e292.

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. (2018). Development of children's sensitivity to overinformativeness in learning and teaching. *Developmental Psychology, 54*(11), 2113–2125.

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology, 69*, 251–273.

Hewlett, B. S., Fouts, H. N., Boyette, A. H., & Hewlett, B. L. (2011). Social learning among Congo Basin hunter-gatherers. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1567), 1168–1178.

Hickling, A. K., & Wellman, H. M. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology, 37*(5), 668–683.

Honomichl, R. D., & Chen, Z. (2012). The role of guidance in children's discovery learning. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(6), 615–622.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences, 20*(8), 589–604.

Jaswal, V. K., & Neely, L. A. (2006). Adults don't always know best: Preschoolers use past reliability over age when learning new words. *Psychological Science, 17*(9), 757–758.

Jirout, J., & Klahr, D. (2020). Questions - and some answers - about young children's questions. *Journal of Cognition and Development, 21*(5), 729–753.

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition, 119*(2), 197–215.
Kline, M. A. (2015). How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals. *Behavioral and Brain Sciences, 38*, e31.
Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science, 15*(10), 694–698.
Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development, 76*(6), 1261–1277.
Kominsky, J. F., Langthorne, P., & Keil, F. C. (2016). The better part of not knowing: Virtuous ignorance. *Developmental Psychology, 52*(1), 31–45.
Kushnir, T., & Koenig, M. (2017). What i don't know won't hurt you: The relation between professed ignorance and later knowledge claims. *Developmental Psychology, 53*(5), 826–835.
Liu, S., & Spelke, E. S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition, 160*, 35–42.
Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science, 358*(6366), 1038–1041.
Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin, 116*(1), 75–98.
Mills, C. M., Sands, K. R., Rowles, S. P., & Campbell, I. L. (2019). i want to know more!": Children are sensitive to explanation quality when exploring new information. *Cognitive Science, 43*(1), e12706.
Nelson, S. A. (1980). Factors influencing young children's use of motives and outcomes as moral criteria. *Child Dev, 51*(3), 823–829.
Oscarsson, M., Jidesjö, A., Strömdahl, H., & Karlsson, K. G. (2009). Science in society or science in school: Swedish secondary school science teachers' beliefs about science and science lessons in comparison with what their students want to learn. *Nordic Studies in Science Education, 5*(1), 18–34.
Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology, 43*(5), 1216–1226.
Repacholi, B., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14-and 18-month-olds. *Developmental Psychology, 33*(1), 12–20.
Rhodes, M., Bonawitz, E., Shafto, P., Chen, A., & Caglar, L. (2015). Controlling the message: Preschoolers' use of information to teach and deceive others. *Frontiers in Psychology, 6*, 867.
Rhodes, M., Brickman, D., & Gelman, S. A. (2008a). Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. *Cognition, 108*(2), 543–556.
Rhodes, M., Gelman, S. A., & Brickman, D. (2008b). Developmental changes in the consideration of sample diversity in inductive reasoning. *Journal of Cognition and Development, 9*(1), 112–143.
Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012a). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science, 15*(3), 436–447. https://doi.org/10.1111/j.1467-7687.2012.01135.x
Shafto, P., Goodman, N., & Griffiths, T. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology, 71*, 55–89.
Shafto, P., & Goodman, N. D. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the 30th annual conference of the cognitive science society*.
Shafto, P., Goodman, N. D., & Frank, M. C. (2012b). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science, 7*(4), 341–351.
Shneidman, L., Gweon, H., Schulz, L. E., & Woodward, A. L. (2016). Learning from others and spontaneous exploration: A cross-cultural investigation. *Child Development, 87*(3), 723–735.
Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review, 120*(4), 779–797.
Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*(7), 309–318.
Tong, Y., Wang, F., & Danovitch, J. (2020). The role of epistemic and social characteristics in children's selective trust: three meta-analyses. *Developmental Science, 23*(2), e12895.
Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Dev, 72*(3), 655–684.
Wellman, H. M., & Gelman, S. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43*(1), 337–375.
Yu, Y., Bonawitz, E., & Shafto, P. (2019). Pedagogical questions in parent-child conversations. *Child Development, 90*(1), 147–161.
Yu, Y., Landrum, A. R., Bonawitz, E., & Shafto, P. (2018). Questioning supports effective transmission of knowledge and increased exploratory learning in pre-kindergarten children. *Developmental Science, 21*(6), e12696.
Yu, Y., Shafto, P., & Bonawitz, E. (2020). Inconvenient samples: modeling biases related to parental consent by coupling observational and experimental results. *Open Mind, 4*, 13–24.
Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Developmental Psychology, 24*(3), 358–365.
Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology, 111*(4), 493–504.
Ziv, M., & Frye, D. (2004). Children's understanding of teaching: The role of knowledge and belief. *Cognitive Development, 19*(4), 457–477.
Ziv, M., Solomon, A., Strauss, S., & Frye, D. (2016). Relations between the development of teaching and theory of mind in early childhood. *Journal of Cognition and Development, 17*(2), 264–284.