Cognitive Psychology 71 (2014) 55–89

Contents lists available at ScienceDirect

Cognitive Psychology

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

A rational account of pedagogical reasoning: Teaching by, and learning from, examples ☆

Patrick Shafto [a,*], Noah D. Goodman [b], Thomas L. Griffiths [c]

[a] University of Louisville, United States
[b] Stanford University, United States
[c] University of California, Berkeley, United States

ARTICLE INFO

Article history:
Accepted 31 December 2013

Keywords:
Learning
Teaching
Bayesian model

ABSTRACT

Much of learning and reasoning occurs in pedagogical situations—situations in which a person who knows a concept chooses examples for the purpose of helping a learner acquire the concept. We introduce a model of teaching and learning in pedagogical settings that predicts which examples teachers should choose and what learners should infer given a teacher's examples. We present three experiments testing the model predictions for rule-based, prototype, and causally structured concepts. The model shows good quantitative and qualitative fits to the data across all three experiments, predicting novel qualitative phenomena in each case. We conclude by discussing implications for understanding concept learning and implications for theoretical claims about the role of pedagogy in human learning.

1. Introduction

One of the most basic questions in cognitive science is how people are able to learn the knowledge they need in order to function in the world. Traditionally, formal approaches to learning have focused on different kinds of knowledge representation and inductive biases that facilitate learning about the

☆ This research was supported in part by NSF award DRL-1149116 to PS. A preliminary version of this paper appeared in the Proceedings of the 30th annual meeting of the Cognitive Science Society. We gratefully acknowledge Kristen Barnes for assistance with data collection.

* Corresponding author. Address: 317 Life Sciences Building, Louisville, KY 40205, United States.
E-mail address: p.shafto@louisville.edu (P. Shafto).

http://dx.doi.org/10.1016/j.cogpsych.2013.12.004
0010-0285/© 2014 Elsevier Inc. All rights reserved.

world (e.g. Bruner, Goodnow, & Austin, 1956; Medin & Schaffer, 1978; Murphy & Medin, 1985; Nosofsky, Gluck, Palemeri, & McKinley, 1994; Pothos & Chater, 2002; Rogers & McClelland, 2004; Rosch & Mervis, 1975; Shepard, Hovland, & Jenkins, 1961; Tenenbaum, Griffiths, & Kemp, 2006). These approaches emphasize individual learners and the explanations of how knowledge is obtained focus entirely on each learner's direct experience with the world and the consequent effects on beliefs. While these capacities are certainly an important part of the explanation of how people come to knowledge about the world, the focus on representation, inductive biases, and individual experience has overlooked another potentially important mechanism that can facilitate learning: other people.

Of the many ways other people may influence an individual's learning, pedagogical situations stand out as having the greatest potential impact on learning. Pedagogical situations are settings in which one agent is choosing information to transmit to another agent for the purpose of teaching a concept (Csibra & Gergely, 2009). Societies have gone to great lengths to facilitate pedagogical situations. In schools, teachers impart their knowledge to students about mathematics, science, and literature through examples and problems. From early in life, parents teach children words for objects and actions by providing them with examples, and establish cultural and personal preferences through subtle glances and outright admonitions.

In addition to providing a means by which individuals can rapidly learn about the world, researchers have argued that pedagogy plays a critical role in cultural evolution. One of the central questions in research on cultural evolution is why humans seem to accumulate knowledge over generations at a much more rapid pace than other animals. Or, in the words of Tomasello (1999), what forms the cultural ratchet that allows knowledge to accumulate? Csibra (2007) has argued that teaching is the explanation—that only humans have a natural ability to engage in and take advantage of explicit teaching situations. However, no one has provided a formal description of what pedagogical reasoning is and how having a teacher who chooses information to present differs from situations in which information is sampled by a relatively uninformative random process.

The key characteristic of pedagogical situations that differentiates learning from the typically assumed model is the presence of a teacher who samples (or chooses) data to help the learner infer the correct answer. Standard approaches to learning assume that data are sampled by some relatively uninformative random process, either implicitly (e.g. Nosofsky, 1986; Pothos & Chater, 2002; Rogers & McClelland, 2004) or explicitly (e.g. Anderson, 1991; Fried & Holyoak, 1984; Tenenbaum, 1999; Tenenbaum & Griffiths, 2001a); however, intuitively it seems that random selection of data does not capture teaching. Instead, it seems more natural to think about teachers as choosing data purposefully, to achieve the goal of teaching. Understanding pedagogical reasoning requires formalizing this process of pedagogical sampling and describing how it affects learning.

In principle, the helpful sampling of data seems likely to allow learning to proceed much more rapidly than if no instruction were provided. If the learner were aware of the teacher's intention to help, they could use this knowledge to make even stronger inferences. Indeed, recent research has argued that from a very young age children understand implications of pedagogical situations, and use this knowledge to guide inferences (Topal, Gergely, Miklosi, Erdohegyi, & Csibra, 2008).

In this paper, we examine pedagogical contexts from the perspective of a rational reasoner, asking how a concept can be optimally taught to a learner by a teacher. We formalize pedagogical reasoning in terms of two problems, one from the perspective of the teacher and one from the perspective of the learner. For the teacher, the problem is to choose the examples that will most help the learner infer the correct concept. For the learner, the problem is to infer the correct concept, given the knowledge that the teacher is choosing helpful examples. The solution to these two interlinked problems is a rational account of pedagogical reasoning.

Our approach contrasts with previous work investigating social influences on learning. For instance, researchers have investigated the effects of cooperation among learners (e.g. Gureckis & Goldstone, 2006), the effects of the responsibility to teach on motivation to learn (e.g. Chase, Chin, Oppezzo, & Schwartz, 2009), and the effects of communication on category structure (e.g. Markman & Makin, 1998) to name just a few. Critically, these approaches focus on how social context influences learning via basic learning processes or how social context influences the kinds of concepts we are inclined to adopt. Our focus is on how the purposeful, goal-directed behavior of others can be leveraged to expedite learning.

We test the predictions of our formal framework using a novel experimental paradigm we call *teaching games* (see also Avrahami et al., 1997). Teaching games are two-part experiments in which participants play the role of teacher, in which they know the answer and are asked to choose the data that will be most helpful, and of learner, in which they are provided examples chosen by a teacher and are asked to infer the true concept. These games are used to investigate pedagogical reasoning in situations where prior knowledge is well controlled and allow systematic tests of the model predictions.

We begin by discussing evidence that people are sensitive to sampling processes in general. We then introduce the idea that we can model learning as Bayesian inference and formalize pedagogical reasoning within this framework. Next, we present a series of experiments testing the predictions of the pedagogical model with different kinds of conceptual structures: rule-based, prototype, and causal concepts. We conclude by discussing implications for concept learning and implications for claims about the role of pedagogy in human learning.

2. Sensitivity to how data are sampled

For people to use pedagogically sampled data to guide learning two minimal conditions must be met. First, teachers must choose helpful examples. Second, learners must be able to use knowledge about the sampling process to guide inferences and understand the implications of data chosen by a teacher. There has been relatively little research on which examples people choose to teach in concept learning tasks (but see Avrahami et al., 1997). However, recent research has provided ample evidence that adults, even very young children, understand that different sampling processes warrant different inferences and they make different inferences in pedagogical situations than in other situations.

Xu and Denison (2009) investigated infants' expectations about balls drawn from a bin either intentionally or randomly. In the intentional conditions, the person drawing had visual access to the bin and could therefore choose a preferred ball. The unintentional conditions consisted of a blindfolded person sampling from the bin. The results showed that children were relatively unsurprised when people in the intentional condition drew low probability balls from the box, while children in the unintentional condition were surprised when low probability balls were drawn from the box, suggesting that even infants are sensitive to the how intentional sampling affects outcomes.

Similarly, Xu and Tenenbaum (2007a, 2007b) have shown that children and adults use knowledge about how language constrains sampling to guide word learning. In these experiments, participants were asked to generalize a novel word from either one or three examples. The results showed that people generalized labels more broadly from one example than from three examples, consistent with an understanding that named examples are constrained to be *positive* examples of a concept; when sampling is restricted to only positive examples, the absence of labeled examples provides evidence against larger concepts. Together these results provide strong evidence that people use their knowledge of how data are sampled to guide inferences.

Additional evidence suggests that even young children draw different inferences in pedagogical and non-pedagogical settings (Csibra & Gergely, 2006, 2009). In one particularly compelling demonstration, Topal et al., 2008 showed that the behavior of 10-month-old infants was highly dependent on the pedagogical nature of experimental settings. In their experiment, children were given the standard *A*-not-*B* task (Piaget, 1955). In this task, a toy is repeatedly hidden under one of two containers (container *A*) and the child is encouraged to search for the toy. These trials are accompanied by persistent and repeated ostensive-pedagogical cues—calling the child's name and alternating eye-gaze between the child and container *A*. Then, on the critical trial, the toy is hidden under the other container, container *B*, in view of the child. The standard finding is that the child continues to search under container A, even though they observed the toy being placed under container *B*. Topal et al. (2008) asked whether children's perseverative behavior was a result of the pedagogical nature of the situation. They tested this by contrasting children's behavior with and without ostensive cuing. The results showed significantly less perseveration in non-pedagogical situations, consistent with they claim that infants draw different inferences in pedagogical and non-pedagogical situations.

The results of these studies illustrate that human learners are sensitive to the way in which the data they observe are sampled and are able to use the information that they are in a pedagogical setting. Intuitively, these results seem connected—being in a pedagogical setting changes the way in which data are sampled. However, we are still left with the questions of *why* reasoning is affected by pedagogical situations and what inferences are warranted in these situations. Answering these questions will require a deeper understanding of how the way in which samples are generated should influence the conclusions drawn from those samples—the problem that we turn to in the next section.

3. Using Bayesian models to capture the effects of sampling

Bayesian models of cognition provide a framework within which the relationship between sampling and learning can be investigated. Bayesian models formalize rational belief updating based on observed data (Jaynes, 2003), and therefore provide a useful reference point for understanding human learning. By indicating how a rational agent should reason, Bayesian models provide us with models of human cognition expressed at Marr's (Marr, 1982) computational level, and consistent with Anderson's (Anderson, 1990) program of rational analysis (see for a review Chater & Oaksford, 1999). Updated *posterior* beliefs depend on two components: *prior* knowledge, and knowledge about how the data are *sampled*. A growing body of evidence suggests that human reasoning is consistent with the predictions of rational models for a wide range of tasks (Griffiths & Tenenbaum, 2006) including learning conceptual structures (Anderson, 1991; Fried & Holyoak, 1984; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Gopnik, Glymour, Sobel, Schulz, & Danks, 2004; Gosselin & Schyns, 2001; Griffiths & Tenenbaum, 2005). However, people's behavior need not be perfectly consistent with rationality for these models to be useful. Computational-level models of cognition help to explain why people do the things they do, even if people only approximate optimal performance. In these cases, the models provide a useful staging point from which to investigate process-level effects, such as imperfect memory (Anderson, 1990; Marr, 1982; Sanborn, Griffiths, & Navarro, 2010).

Formally, the problem of learning is a problem of updating one's belief in a hypothesis, *h*, based on some observed data, *d*. If degrees of belief are expressed in terms of probabilities, the solution to this problem is provided by Bayes' theorem,

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}, \tag{1}$$

where $P(d|h)$ is the probability of observing (or *sampling*) the data assuming the hypothesis is true, $P(h)$ represents the learner's prior beliefs in the hypothesis, and $\mathcal{H}$ is the set of all hypotheses under consideration. Eq. (1) makes it clear that learning will depend on two distinct factors: a factor based on what we believed before we saw the data, expressed via the prior $P(h)$ and the hypothesis space $\mathcal{H}$, and a contribution based on how we think data are sampled given a hypothesis, expressed in the likelihood $P(d|h)$.

One strength of this approach is that it allows straightforward integration of different kinds of sampling assumptions through the use of generative models (Tenenbaum et al., 2006). Generative models formalize the link between hypotheses and possible sets of observed data, by specifying which data are likely given a particular hypothesis. Differences between generative processes such as random sampling and sampling by a teacher can be formalized in this framework as different methods of generating data.

Traditionally, models of learning have tended to overlook or downplay the role of sampling in learning. In most models, there is no discussion of how data are generated (e.g. Medin & Schaffer, 1978; Nosofsky et al., 1994; Pothos & Chater, 2002; Rogers & McClelland, 2004). For instance, connectionist models specify the structure of models and the rules for updating weights but generally no mention is made of where the data come from or whether that even matters. Implicitly, it seems the models assume that examples are some kind of relatively uninformative random sample, otherwise it is hard to imagine how they would be successful in learning the structure of the world. Even Bayesian approaches, though required by the formalism to specify how the data are generated, have almost exclusively considered data to be the consequence of some random process (Anderson, 1991; Fried & Holyoak, 1984; Goodman et al., 2008; Gopnik et al., 2004; Griffiths & Tenenbaum, 2005). In summary, though the generative process by which data are sampled plays a key role in learning, this issue has been largely overlooked in previous research.

3.1. Generative models as sampling processes

As summarized above, recent research has shown that even simple sampling assumptions can have a powerful effect on learning. One distinction that has been introduced in the literature on concept learning is the difference between *weak sampling* and *strong sampling* (Hsu & Griffiths, 2009; Tenenbaum, 1999; Tenenbaum & Griffiths, 2001a). Weak and strong sampling correspond to two different ways in which data can be generated, and correspond to different likelihood functions assumed by the learner.

In weak sampling, examples are selected at random from the set of all possible objects and are then labeled as to whether or not they are instances of the target hypothesis. The key idea is that in weak sampling the process by which examples are selected is independent of the hypothesis; the hypothesis is merely used to provide labels. Thus, the examples provide little information about the target hypothesis—only whether this example is in or out.

In strong sampling, examples are generated at random from the set of examples that are true of the hypothesis. Unlike in weak sampling, in strong sampling the process by which examples are selected depends on the target hypothesis. Thus, when hypotheses tend to be small, positive examples are rare and consequently strong sampling provides somewhat more information about the target hypothesis.

More concretely, consider the case where the hypothesis, h, is a specific concept, say “cat”, and the data, d, is one of the n objects in the room that could be labeled as an instance of this concept. In weak sampling, the example is chosen uniformly at random from the whole set then labeled correctly, so $P(d|h) = \frac{1}{n}$. Because any object is equally likely to be considered, the evidence is equally probable for all hypotheses for which the data are consistent. This corresponds to a situation in which a speaker is thinking of a “cat”, chooses an object at random, such as a desk, then states whether the example is a member of the concept (here, “no”). As the learner, this example rules out some concepts such as “desk”, “furniture”, and “object”, but is consistent with a vast set of other possibilities. In language learning, a similar sampling process would apply if the learner heard a set of randomly generated sentences and was told which sentences were grammatical and which were not (Hsu & Griffiths, 2009).

In contrast, in strong sampling the example is chosen at random from the set of examples consistent with the hypothesis. Consequently, $P(d|h) = \frac{1}{|h|}$ for any example consistent with h, where $|h|$ is the number of positive examples of the concept h in the set of n objects. Strong sampling induces a bias toward smaller concepts because small concepts have fewer positive examples (Xu & Tenenbaum, 2007a, 2007b). In word learning, this corresponds to a situation in which a speaker chooses an example at random from the things that happen to be cats and labels it as a “cat”. Likewise, in language learning, strong sampling corresponds to the assumption that the sentences one hears are generated at random from the set of grammatical sentences in the language. Because strong sampling depends on the hypothesis, it takes different forms for different kinds of concepts. For prototype concepts, the underlying model may be a normal distribution, where data differ in the degree to which they are typical of the concept. In this case, strong sampling suggests that the data are drawn from the true model; they are samples from the correct normal distribution. Regardless of the type of concept, a learner who correctly assumes strong sampling will learn faster than one who assumes weak sampling.

Despite their differences, both strong and weak sampling assume that learners are provided with data that are generated by a relatively uninformative random process. The idea of relatively uninformative, randomly-sampled data makes sense for some situations, but does not capture the informative, intentional sampling process that underlies teaching. Rather than choosing uninformatively, a person teaching a concept could be expected to choose data that tend to be good—data that help the learner to infer the correct hypothesis—over other equally true examples. For rule-based concepts, all examples consistent with the rule are equally in the concept, but not all are equally helpful. The best examples are those that tend to eliminate alternative hypotheses. Thus, in teaching situations, the generative process should depend on the true concept, as well as the other concepts that it might be confused with.

It seems that learning from pedagogically sampled data could be more powerful than either weak or strong sampling. Indeed, this is why pedagogy has been implicated as an explanation of how people and cultures accumulate knowledge. However, to understand the contribution of pedagogy to human learning, it is important to formalize what the problem of pedagogical reasoning is, and how pedagogical situations could affect inferences.

4. A Bayesian model of pedagogical reasoning

We begin by sketching the problem of pedagogical sampling, then develop a computational model of teaching, and learning from teaching, based on the idea that data are helpfully sampled. In pedagogical settings, there are two parties, a teacher and a learner. The teacher is assumed to know the correct answer and their goal is to choose data that facilitate learning (see Fig. 1). The learner is assumed to know that the teacher is being helpful, and their goal is to infer which concept the teacher is trying to teach them. Pedagogical reasoning thus involves two interrelated problems: the teacher needs to choose data, and the learner needs to make an inference from those data. We formalize this process via two assumptions: the first describes which data the teacher should choose to help a learner; the second, how the learner should update their beliefs, given the data that the teacher has chosen data to help them learn. We then describe how these two assumptions combine together to prescribe reasoning by teachers and learners in pedagogical settings.

The teacher's problem is deciding which data to present to the learner. In contrast with strong and weak sampling, pedagogical sampling is purposeful, and depends on what the teacher thinks the learner will infer given the data. To best achieve the pedagogical goal, the teacher should choose data d that will maximize the learner's belief in the correct hypothesis h. In the ideal case, the distribution from which the teacher generates data, $P_{\text{teacher}}(d|h)$, will thus spread probability evenly among those d that maximize the posterior belief of the learner in $h, P_{\text{learner}}(h|d)$ (Tenenbaum & Griffiths, 2001b, cf. who used a similar approach to formally define the extent to which data d are representative of a hypothesis h).

Choosing only data that maximize $P_{\text{learner}}(h|d)$ may be ideal when the teacher has perfect knowledge of the hypothesis space, prior, and likelihood used by the learner, but is not robust to differences in inference that may arise as a consequence of differences between assumptions of the teacher and the actual state of the learner. A more robust strategy is to select data that give the target hypothesis high posterior probability, even if they are not the most representative examples. This more robust strategy can be formalized through a "soft" maximization of the posterior probability:

$$P_{\text{teacher}}(d|h) \propto (P_{\text{learner}}(h|d))^{\alpha}, \tag{2}$$

where α captures how strongly the teacher tends toward maximizing the posterior. As $\alpha \to \infty$, this reduces to choosing the data that maximize the posterior. As $\alpha \to 0$, the teacher chooses uniformly among data consistent with the hypothesis (i.e. those for which $P(d|h) > 0$), meaning that pedagogical

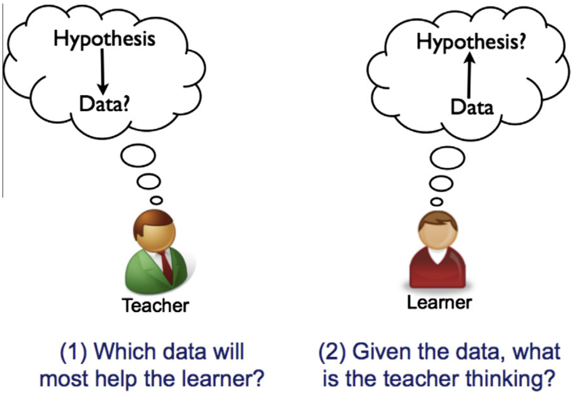

Hypothesis
Data?
Teacher
Hypothesis?
Data
Learner
(1) Which data will most help the learner?
(2) Given the data, what is the teacher thinking?

Fig. 1. A schematic depiction of pedagogical reasoning. On the left, the teacher knows the correct hypothesis, which they intend to teach to the learner by choosing helpful data. On the right, the learner observes the data, and knowing the teacher's intention, infers which hypothesis is being taught.

sampling contains random sampling as a special case—the case where the teacher does not choose data helpfully. When $\alpha = 1$, data are generated in direct proportion to the posterior probability they give to the target hypothesis, implementing a kind of "probability matching." For all of the results in the paper, we fixed the value of α to 1. Response selection rules similar to Eq. (2) are common in cognitive modeling, being an instance of the Luce choice rule (Luce, 1959).

Now consider the learner's problem, updating their beliefs given the teacher's data. In such a situation, the learner knows that data will not be generated by a relatively uninformative process, but instead are sampled by a helpful teacher. The learner can therefore assume that the data are sampled according to Eq. (2). For a rational agent, belief updating will then be given, by Bayes' theorem, by the product of the prior probability and the likelihood consistent with this sampling assumption:

$$P_{\text{learner}}(h|d) = \frac{P_{\text{teacher}}(d|h)P(h)}{\sum_{h'} P_{\text{teacher}}(d|h')P(h')}. \tag{3}$$

The learner's posterior beliefs depend on their prior bias and the degree to which the data are likely to be chosen by a helpful teacher given that hypothesis.

Because the behavior of the teacher and learner each depend on (their assumptions about) the behavior of the other, Eqs. (2) and (3) form a mutually dependent *system* of equations. The distribution from which a teacher should generate data, and which a learner should use as a likelihood, is the solution to the system of equations defined by substituting Eq. (3) into Eq. (2), with

$$P_{\text{teacher}}(d|h) \propto \left(\frac{P_{\text{teacher}}(d|h)P(h)}{\sum_{h'} P_{\text{teacher}}(d|h')P(h')} \right)^{\alpha}. \tag{4}$$

To understand this account of pedagogical reasoning it may help to consider one way of solving the system of equations: fixed-point iteration. Imagine that you are the learner, and wish to update your beliefs. To do so you will need an estimate of the likelihood $P_{\text{teacher}}(d|h)$ of seeing the examples you are given. You can estimate this likelihood by assuming the teacher is rational—Eq. (2)—but to do this you need an estimate of the $P_{\text{learner}}(h|d)$ used by the teacher. If you assume the teacher assumes that you are rational, you can use Eq. (3) as such an estimate; this requires an estimate of $P_{\text{teacher}}(d|h)$, and so on. This recursive reasoning will eventually no longer change—we then say that the process has iterated to a fixed point and this fixed point will necessarily be a solution to the system of equations defining rational pedagogical reasoning. Thus we can understand the model as capturing the outcome of a recursive mental reasoning process. However, we emphasize that rational pedagogical reasoning describes the *outcome* of this process (or rather the solution to the system of equations) and it is entirely possible that this reasoning may be implemented by a psychological process that does not require any explicit recursion.

In general, there are multiple solutions to the system of equations. Thus, in principle there are multiple possible ways in which teachers and learners may satisfy rational pedagogical reasoning. A sufficient condition for identifying a single solution is specification of an initial distribution for the teacher's selection of data. Iteration to a fixed point, in the way outlined in the previous paragraph, transforms this initial distribution into a solution satisfying our assumptions for pedagogical reasoning. An intuitive choice for initial distribution often comes from the generative model which would be used to describe the situation in a non-pedagogical context. In the cases we consider in the remainder of the paper, we take the initial distribution to be that obtained from unbiased random sampling from the generative model—weak sampling in the cases where negative evidence is possible and strong sampling in cases where it is not.

One of the strengths of this pedagogical model is that it allows predictions for any concept for which one can specify a reasonable set of hypotheses. Applying the model requires only specification of the hypothesis space, the space of examples, a common prior belief distribution, and an initial distribution for the teacher. Importantly, because sets of hypotheses are interrelated in different ways for different concept learning problems, and the pedagogical model is sensitive to the hypothesis being taught and the alternative hypotheses, it will generate different qualitative predictions for different domains. In the remainder of the paper, we describe and test the predictions of the pedagogical model for teaching and learning for three kinds of concepts: rule-based, prototype, and causal.

5. Pedagogical reasoning about rule-based concepts

Consider a simple example which we call the rectangle game: a game where the teacher thinks of a rectangle on a board, and tries to teach that concept to a learner by choosing to label points inside and/or outside the rectangle (cf. Tenenbaum, 1999). In the rectangle game, the learner's job is to try to infer, given the labeled examples chosen by the teacher, which concept the teacher is thinking of (i.e. which rectangle). The concepts are rule-based: a point is either inside or outside the rectangle, and all points inside (or outside) the rectangle are a priori equal.

Fig. 2 presents potential teacher and learner scenarios. In each case, there seem to be choices which are obviously better than others. As a person trying to teach someone the rectangle in blue (top left), the examples in the middle panel seem better than those on the right. Similarly, as a learner, given the examples in the bottom left, the rectangle in the middle panel seems like a better guess than that on the right. Notice that in both Figs. 2a and b, the middle and right panels are possible; however, our intuition tells us that the middle panels are better guesses than the right panels.

For these rule-based concepts, data include both the location of a point and a label indicating whether that point is inside or outside the concept. To define a generative model for rule-based concepts, specifying the likelihood $P(d|h)$, we must include a method for generating both the locations of points, y, and their labels, l. For simplicity, we assume a generative model in which points are chosen at random, then labeled correctly. This corresponds to weak sampling.[1] If the locations are chosen uniformly at random from the entire board, then any collection of m points that is labeled in a way that is consistent with a hypothesis h has probability $1/B^m$, where B is the area of the board. The hypothesis space, $\mathcal{H}$, includes all possible rectangles on the board, varying in size (both horizontal and vertical) and location. Importantly, the addition of labels provides information about the correct hypothesis—for each randomly-selected point, the label tells whether that point is inside or outside the true hypothesis. Pedagogical sampling is then an issue of inferring which points will be most helpful for the learner.

Applying the model to the examples in the rectangle game, we can ask which examples are best for teaching. In the case of two positive examples, the prediction is that the teacher will generally place examples in opposite corners of the rectangle, and the learner will infer a rectangle such that the examples are near opposite corners. To understand why this is a solution to Eq. (4), consider the recursive reasoning described above (idealized to avoid complications of uncertainty): if the learner assumes that the teacher will choose examples at opposite corners, then Eq. (3) implies that the unique inference made by the learner is the tightest rectangle around two examples; if the teacher assumes that this is what the learner is doing, then according to Eq. (2) the teacher will usually choose examples in opposite corners of the true rectangle. In the case of one positive example and one negative example, the reasoning is similar: the learner assumes that the teacher will choose a negative example close to the boundary of the rectangle, enabling the learner to rule out larger rectangles; in turn the teacher will chose such examples under the assumption that this is how the learner will reason. The model predictions are shown in Fig. 3 (see Appendix for full details about model implementation, as well as a worked example of how the model generates predictions). We test these predictions in the following experiment.

6. Experiment 1: Teaching and learning rule-based concepts

In this experiment, people played the rectangle game. People played the roles of teacher, choosing the examples given a rectangle, and learner, guessing a rectangle given examples. In each case, participants did not see a partner, but were informed of the pedagogical nature of the situation—that they were choosing examples for teaching or that a teacher had chosen the examples they saw. We contrast these situations with learning from non-pedagogically sampled evidence. In no case was a partner (teacher or learner) present.

[1] Alternatively, one could assume that a set of labels are chosen randomly, then the locations of the points are chosen conditioned on the labels. This would correspond to strong sampling (modified to handle negative data). Both processes lead to qualitatively similar answers for this case, and we are not committed to the psychological plausibility of this particular account.

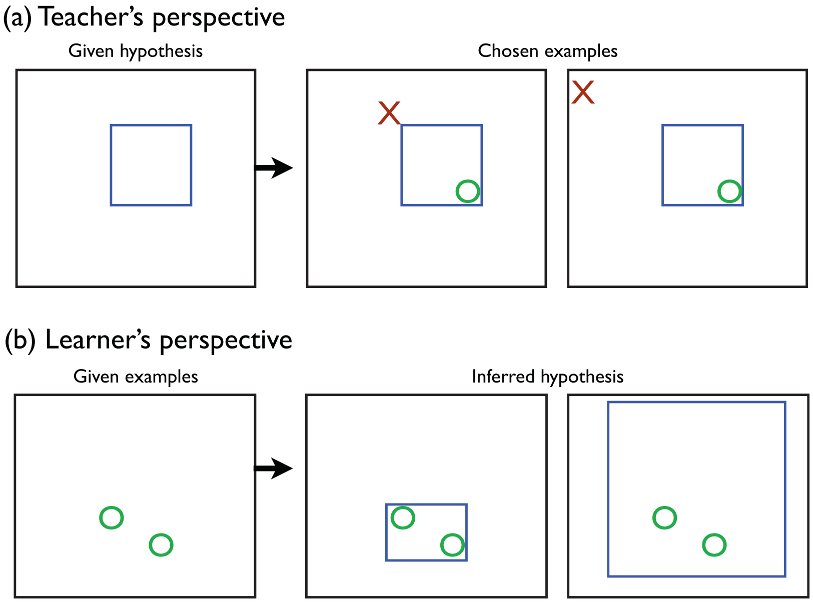


Fig. 2. Two rectangle game scenarios. (a) Teachers are given a hypothesis (shown on the left) and choose which examples to provide. The two panels on the right show two possible pairs of labeled points, which we refer to as examples. The first set of examples are intuitively better than the second. (b) Learners are given labeled examples provided by a teacher (shown on the left) and infer which hypothesis is correct. The two panels on the right show two possible hypotheses. The first hypothesis is intuitively better than the second.

We included three conditions: *Teaching-Pedagogical Learning*, *Pedagogical Learning*, and *Non-Pedagogical Learning*. In the *Teaching-Pedagogical Learning* condition, participants first played the role of teacher, then played the role of learning from a teacher. In the *Pedagogical Learning* condition, participants only learned from examples selected by a (not present) teacher. Finally, In the *Non-Pedagogical Learning* condition, participants only played the role of learner, where they knew that examples were not selected by a teacher. The three conditions allow us to explore which examples people choose to teach, whether learning from ostensibly pedagogically-selected examples differs from non-pedagogically selected examples, and whether there are effects of teaching first on later learning.

In our analyses, we will investigate how the pedagogical model predicts the qualitative and quantitative patterns observed in the human data. For the teaching task, we will ask whether the pattern of responses is random, as predicted by weak and strong sampling. We will also characterize people's preferred pattern of responses and ask whether the pedagogical sampling model correctly predicts the pattern of examples generated by people. For the learning task, we will ask comparable questions: are learner's inferences consistent with the assumption that examples are randomly-sampled? If not, do learner's inferences reflect knowledge about the pedagogical nature of the data?

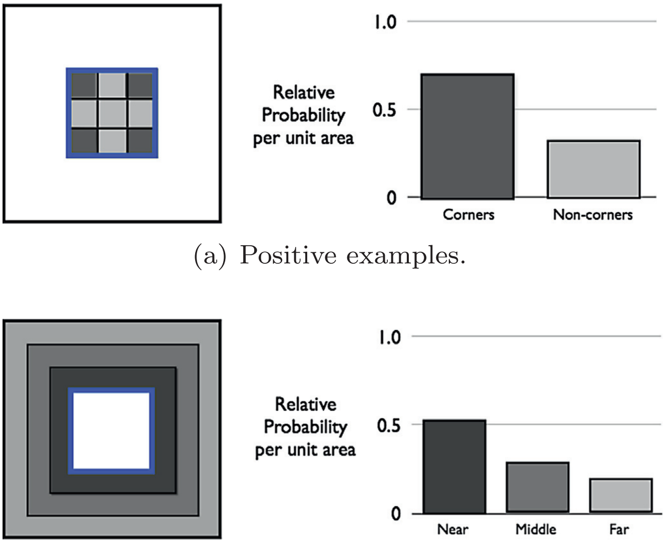
6.1. Method

6.1.1. Participants

Seventy-three University of Louisville undergraduates participated in exchange for course credit.

6.1.2. Design

Participants were randomly assigned to either the *Teaching-Pedagogical Learning* condition ($n = 18$), the *Pedagogical Learning* condition ($n = 26$), or a *Non-Pedagogical Learning* condition



(a) Positive examples.

(b) Negative examples.

Fig. 3. Predictions of the pedagogical model for Experiment 1: teaching with (a) positive examples and (b) negative examples. The blue square represents the rectangle. (a) For positive examples, we collapse the data into two categories: corners and non-corners, as indicated by the shaded areas. The pedagogical model predicts that examples should be in the corners of the rectangle. (b) For negative examples, we collapse the data into three categories based on distance from the edge of the rectangle: near, middle, and far, as indicated by the shaded areas. The pedagogical model predicts that examples should be near the boundary of the rectangle. Note that because the rectangle was positioned randomly in the experiment, the thickness of the near, middle, and far is relative to the distance between the boundary of the rectangle and the edge of the board. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

($n = 29$). For the *Teaching-Pedagogical Learning* condition, the experiment consisted of two parts, teaching and learning from a teacher, presented in that order. For the *Pedagogical Learning* condition, the experiment consisted of two parts, observing the hypotheses (without choosing examples to teach) and learning from a teacher, presented in that order. For the *Non-Pedagogical Learning* condition, the experiment consisted of two parts, observing the hypotheses (without choosing examples to teach) and learning without a teacher, presented in that order. Across all three conditions, learners saw trials with one, two, and three examples. Similarly, on different trials, teachers chose one, two, or three examples. We focus our analyses on the trials with two examples, as these are where the model predictions are the most interesting.

The rectangles used in the experiment ranged in both width and height. The game board was a square. For the purposes of placing the rectangles, the board was divided into a 6 × 6 grid. The rectangles used in the experiment ranged from a minimum of $\frac{2}{6}$ the size of the board to $\frac{5}{6}$ the size of the board. The location of the rectangle was randomized.

6.1.3. Procedure

Experiments were run on Apple Mac Pro desktop computers using MATLAB. Participants were seated at a computer and told that they were going to learn about a game called the rectangle game. In the *Pedagogical Learning* conditions, participants were told that in the rectangle game there is a teacher and a learner. It is the teacher's job to help the learner infer the correct rectangle by choosing helpful examples, points that can be inside or outside the true rectangle. In the *Non-Pedagogical Learning* condition, participants were told that in the rectangle game the goal is to guess the position of the rectangle.

In the *Teaching-Pedagogical Learning* condition, participants first participated in a teaching task. For the teaching task, participants were shown a rectangle and asked to choose examples to help the learner infer the location of the rectangle. The examples were chosen by clicking on the screen. A green circle automatically appeared if the click was inside the rectangle, and a red X if the click was outside the rectangle. There were 30 rectangles for which participants chose two examples—3 of each size. The positions and ordering of the rectangles were randomized.

In the *Pedagogical Learning* and *Non-Pedagogical Learning* conditions, participants were shown the same rectangles as in the teaching condition and were asked to click anywhere on the screen to advance to the next rectangle. A pause was inserted to ensure that each rectangle appeared on the screen for at least .5 s.

In the learning task, participants in the *Pedagogical Learning* conditions were shown labeled examples. Learners were told that the examples were chosen by a teacher who was trying to help them to infer the correct rectangle. Participants in the *Non-Pedagogical Learning* condition were told that they could click twice on the board to ask a teacher whether those points were inside the rectangle.

After observing the labeled examples, participants in all conditions were asked to infer the correct rectangle by clicking on the screen and dragging with the mouse. In the *Pedagogical Learning* conditions, examples were generated based on a small set of patterns to reflect the possibilities given the hypotheses, i.e. positive pairs are more likely to be close together than negative pairs, and mixed pairs are relatively unconstrained. For mixed pairs, examples were chosen to differ by between $\frac{0}{6}$ and $\frac{4}{6}$ the size of the board on the *x* and *y* dimensions. (Examples were constrained such that they could not differ by $\frac{0}{6}$ on both dimensions.) For positive pairs, examples were chosen to differ by between $\frac{0}{6}$ and $\frac{3}{6}$ the size of the board on both the *x* and *y* dimensions. For negative pairs, examples were chosen to differ by between $\frac{0}{6}$ and $\frac{4}{6}$ the size of the board on one dimension and $\frac{2}{6}$ and $\frac{4}{6}$ on the other. The basic patterns were positioned randomly on the board. In the *Non-Pedagogical Learning* condition, participants' choices appeared as black dots and after they picked both points, the black dots changed to green O's or red X's. There were 36 pairs of examples in total (24 mixed, 6 positive, 6 negative). When participants completed the task, they were debriefed and thanked.

6.2. Results & discussion

We will consider performance on the teaching task first, then the learning task. For the teaching task, we will consider whether people's data conform to the predictions of weak and strong sampling or pedagogical sampling by separately analyzing the distributions of the positive and negative examples.

Pedagogical sampling predicts that for positive examples, examples in the corners are more informative than examples in the middle. Both weak sampling and strong sampling predict that positive examples are distributed uniformly at random. To make responses for different sizes of rectangles comparable, we divided each rectangle into a 3×3 grid. Grids were normalized based on the size of the rectangle, so that the grid was finer for smaller rectangles than for larger ones. This allowed us to ignore the size of the rectangle and focus on the relative position of the examples. Frequencies of examples in each area of the grid were tallied, and collapsed into two groups of examples—corners and non-corners—and the frequency per unit area was calculated. To account for the fact that some people produced more positive examples than others, we computed the proportion of corner examples per unit area. Fig. 4 shows that people in the teaching task were more likely to choose positive examples in the corners than in non-corners, $M_{corner} = .87,\ M_{non-corner} = .13,\ t(17) = 6.55,\ p < .0001$ by one-sample *t*-test, consistent with the predictions of the pedagogical model.

For negative examples, pedagogical sampling predicts strong effects of distance—the most helpful negative examples are those that are near the boundaries of the rectangle. Weak sampling predicts that examples should be distributed uniformly at random, while strong sampling makes no prediction. We analyzed people's choices by classifying examples based on the relative distance from the boundary of the rectangle to the outside of the board. We divided the area from the boundary of the rectangle to the edge of the board into three bins (see Fig. 3). Because the rectangles were randomly positioned, the width of these bins depends on the distance between the boundary of the rectangle and the edge of the board. To test whether people were more likely to choose examples near the boundaries, for each person, we computed the frequency of examples that were near the rectangle (shaded with the darkest gray; see Fig. 3), middle (the next ring), and far (the outer ring shaded with

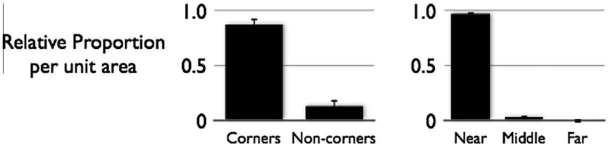


Fig. 4. Results from the Experiment 1 teaching task divided into positive examples (left) and negative examples (right). Error bars represent two standard errors here and throughout the paper. When choosing positive examples, people overwhelmingly chose examples in the corners, as predicted by the pedagogical model. When choosing negative examples, people overwhelmingly chose examples near the boundaries, as predicted by the pedagogical model.

the lightest gray) per unit area, and converted to proportions per unit area to control for different numbers of negative examples produced by different participants. The results are shown in Fig. 4, right panel. The results show that people were more likely to choose negative examples that were near the boundaries of the rectangle, $M_{near} = .97,\ M_{middle} = .03,\ M_{far} = 0,\ F(1,15) = 355.8,\ p < .0001$ by one-way, repeated measures ANOVA.[2] These results are consistent with the predictions of the pedagogical model.

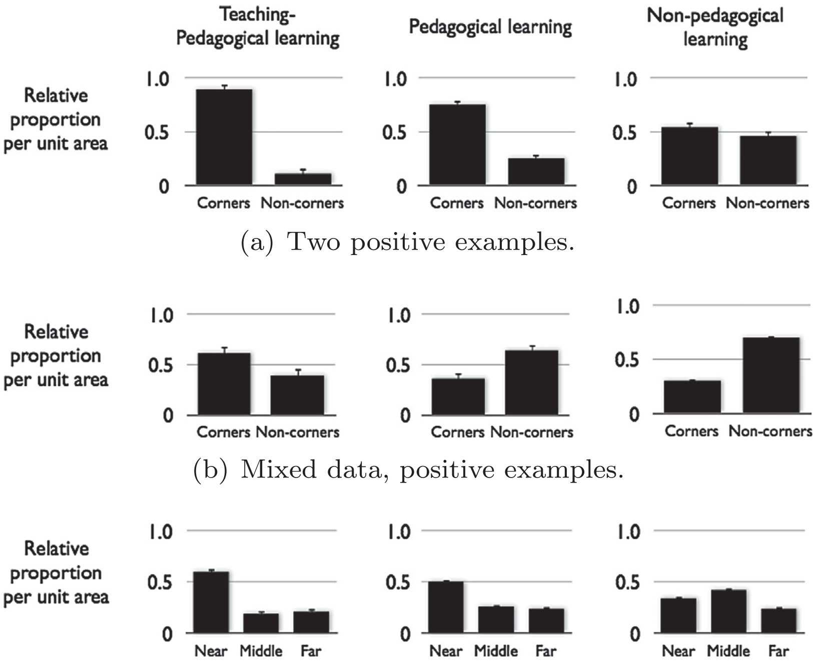
Pedagogical sampling predicts that learners should know the strategies that the teacher will use to teach different concepts. Therefore, if learners use pedagogical knowledge to guide inferences, we expect learners should draw rectangles that recover the patterns observed in the teaching data. Weak sampling predicts that positive examples should be randomly distributed with respect to the rectangles inferred by learners. Fig. 5a shows the proportion of examples per unit area in the corners of learners' inferred rectangles for cases with two positive examples, for each of the three conditions. The *Teaching-Pedagogical Learning* and the *Pedagogical Learning* conditions show significantly more corner examples per unit area, $t(17) = 7.96,\ p < .001$ and $t(25) = 5.86,\ p < .001$, as predicted by the pedagogical model, while there was no significant difference in the *Non-Pedagogical Learning* condition, $t(27) = 1.02,\ p = .32$. These results are consistent with the interpretation that learners understand the difference between pedagogical and non-pedagogical situations, and use this knowledge to guide inferences.

Fig. 5b shows the results for positive examples where one example was positive and one example was negative. The results show a divergence between the *Teaching-Pedagogical Learning* condition and the *Pedagogical Learning* condition. The *Teaching-Pedagogical Learning* condition shows higher proportion per unit area in the corner, $t(17) = 1.76,\ p < .05$ by one-tailed test, but the *Pedagogical Learning* condition shows the opposite pattern, $t(25) = -2.21,\ p < .05$, like the *Non-Pedagogical Learning* condition, $t(28) = -7.39,\ p < .001$. This suggests that going through the teaching trials first may have had an effect on pedagogical learning, an issue that we return to in the General Discussion.

Turning to the negative examples, we collapsed the pairs of negative examples with the mixed negative examples because there were no differences in the pattern of results. The pedagogical model predicts that negative examples are more likely to be near the boundary of the rectangle than they are to be middle or far, and that middle and far should be approximately similar (see Fig. 3c). Fig. 5c shows the results. The *Teaching-Pedagogical Learning* and the *Pedagogical Learning* conditions both show the predicted pattern of results, $F(1,36) = 81.21,\ p < .0001$ and $F(1,52) = 36.08,\ p < .0001$ by planned contrast with weights 1, $-.5$, and $-.5$, and the Non-Pedagogical Learning condition does not, $F(1,58) = 3.22,\ p = .16$, as predicted.

Together these results show that the pedagogical model provides good fits to human data in rule-based concept learning. The non-pedagogical scenario shows that the qualitative effects predicted by the model and observed in the data are not the result of simple perceptual preferences; rather, people appear to be sensitive to when pedagogical sampling applies.

[2] Two participants produced no negative examples and were omitted from the analysis, dropping the number of participants from 18 to 16 for this analysis.



(a) Two positive examples.

(b) Mixed data, positive examples.

(c) Negative examples.

Fig. 5. Learning results for Experiment 1: (a) two positive examples, (b) mixed data, positive examples, and (c) all negative examples. In each case, the results from the *Teaching-Pedagogical Learning* (left), *Pedagogical Learning* (middle), and *Non-Pedagogical Learning* (right) conditions are shown. The results for the *Teaching-Pedagogical Learning* and *Pedagogical Learning* conditions match with the model predictions for positive and negative pairs, but the *Pedagogical Learning* condition deviates somewhat in the case of the mixed data.

7. Pedagogical reasoning about prototype concepts

The question of whether or to what degree concepts are rule-based is a contentious one and has incited long-running debates in the literature. Many authors have argued that the classical view of concepts as rule-based is untenable in the face of evidence that category membership appears to be graded and category boundaries do not appear to be crisply defined (Posner & Keele, 1968; Rosch & Mervis, 1975). As a consequence, many existing models of category learning are based on probability density estimation, rather than inferring rules (e.g. Anderson, 1991; Ashby & Alphonso-Reese, 1995; Fried & Holyoak, 1984; Griffiths, Sanborn, Canini, & Navarro, 2008; Nosofsky, 1986, 1991).

A strength of the pedagogical model is that it can by applied to different kinds of concept learning tasks by implementing different spaces of hypotheses, $\mathcal{H}$. For one-dimensional prototype concepts, we can capture graded category membership with hypotheses that determine category membership based on a probabilistic distribution. A simple choice is a normal distribution, where different hypotheses, $h \in \mathcal{H}$, vary in their means, μ, and a variances, σ. An individual hypothesis is thus a pairing of a mean and a variance, $h = \{\mu, \sigma\}$. The learner infers the mean and variance based on the data that they observe.

Consider the prototype concept represented by the examples in Fig. 6. The examples are instantiations of a concept based on line length, where the lines have a mean length and vary to some degree around the mean. These examples have been sampled at random from the true concept and examples near the mean are more likely than shorter (middle row, left column) and longer (middle row, middle column) examples. This is the space of possible examples and observed data are selected from this set. A generative model based on random selection from among these examples is then equivalent to

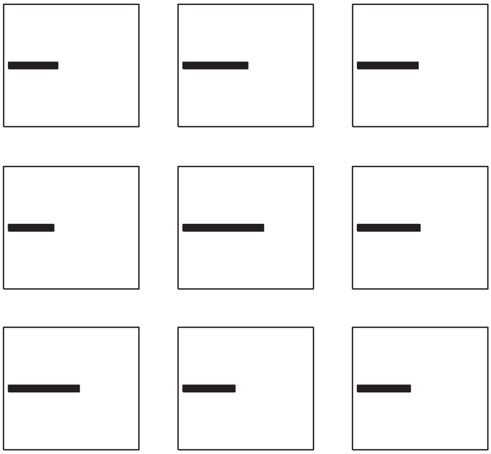


Fig. 6. Examples from a prototype concept based on line length. The frequency of each length is proportional to the probability given the true mean and variance. In the teaching experiment, people received 27 examples of this kind, which represented examples of a concept. Participants were asked to choose three examples to teach someone else the concept.

sampling from the true hypothesis and the probability of generating data is proportional to the probability of the data under the hypothesis, $P(d|\mu,\sigma)$.

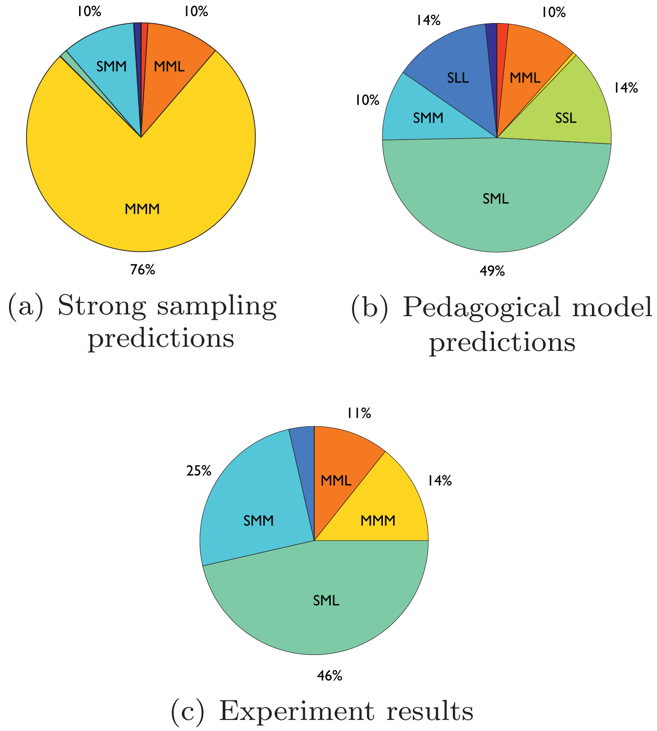
Now, imagine a teacher who has a catalog of examples of the concept in Fig. 6 to choose from. Which three examples will be most helpful for the learner? Because the possible examples are drawn from a distribution, the optimal strategy for the teacher is to provide information about what a typical example is, as well as the breadth of examples that they have observed. The teacher, therefore, would choose examples that marked the center of the distribution and two examples that marked the extremes.

In contrast, consider the examples that would be chosen under strong sampling. Strong sampling suggest that examples should be drawn randomly from the true concept. For prototype concepts, this means drawing examples randomly from a normal distribution with the correct mean and variance. These examples will be representative of the true distribution and would therefore tend to consist of examples near the mean. If we were to draw many random sets of three and tracked which examples were chosen, then by the law of large numbers, the distribution of sampled examples would converge to the true hypothesis.

From the perspective of the learner, pedagogically chosen examples should allow more confident inferences about the center and tighter inferences about the extent of the concept. Inferences about the center of the distribution should be strong because the teacher would explicitly mark the mean with an example, and chosen triads are predicted to be symmetric about the mean. In contrast, strong sampling results in examples that are randomly sampled, and though good inferences about the mean are guaranteed given enough examples, triads are more likely to be asymmetric resulting in greater variance about the estimated mean. Because inferences are stronger from pedagogically sampled data, learners' subsequent generalizations should extend less broadly. In contrast, learners observing randomly sampled data should generalize more broadly because they have greater uncertainty about the true hypothesis.

8. Experiment 2: Teaching and learning prototype concepts

For the teaching task, participants were presented with a set of examples which vary on a single dimension (see Fig. 6). The examples were concentrated around a mean value, with the number of examples at each value varying based on proximity to the prototype. Participants were asked to choose three examples with which to teach another person the concept. The predictions that result from strong sampling and pedagogical model are shown in Fig. 7a and 7b (see Appendix for full details about model implementation). The pedagogical model predicts that examples are most likely to



(a) Strong sampling predictions

(b) Pedagogical model predictions

(c) Experiment results

Fig. 7. Model predictions and empirical results for the prototype teaching task in Experiment 2. The figures show the probability of choosing different triplets of examples. Examples have been binned into three groups: small, medium, and large. (a) Strong sampling predicts that people would choose three medium-sized examples. (b) The pedagogical model predicts that people would choose one small, one medium, and one large. (c) Approximately half of the participants chose one small, one medium, and one large, as predicted by the pedagogical model.

include one small, one medium, and one large example. Strong sampling predicts that examples are most likely to include three medium examples.

For the learning task, the sampling of the examples (chosen by a teacher or randomly sampled) was crossed with a cover story about whether the examples were pedagogical or not, resulting in four conditions. Fig. 8a shows the model predictions. We expect people in the teacher cover story conditions (cyan[3] and green lines) to generalize less broadly than those in the random sampling cover story conditions (blue and red lines). The residual differences between the curves reflect the fact that random samples tend to have a smaller range than those chosen by the teachers, resulting in more narrow generalizations.

8.1. Method

8.1.1. Participants

Twenty-eight University of California, Berkeley, undergraduates participated in the teaching task exchange for course credit. Eighty-four members of the community at the University of California, Berkeley participated in the learning task, in exchange for either course credit or compensation of $10/hr.

8.1.2. Stimuli

Stimuli were heavy black lines shown inside a thin black rectangle, on a white background. Each line was 1.5 mm thick, and the rectangle was 31 mm wide and 29 mm high, with the line centered

[3] For interpretation of color in Fig. 8, the reader is referred to the web version of this article.

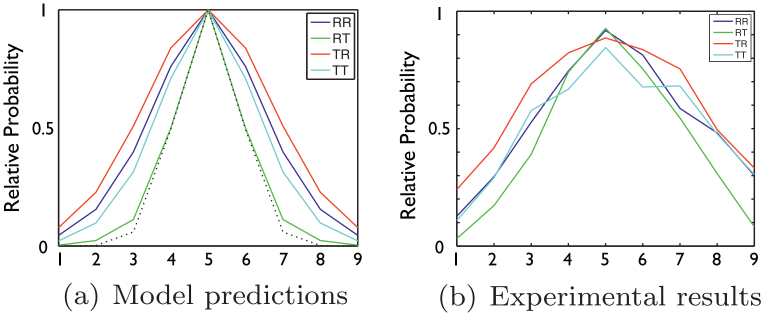


Fig. 8. Model predictions and empirical results from the prototype learning task in Experiment 2. (a) The model predictions indicate that generalizations in the *Random/Teacher* (RT) condition should be most restricted, followed by the *Random/Random* (RR) and *Teacher/Teacher* (TT) conditions. The *Teacher/Random* (TR) condition should be the broadest. (b) The results from the experiment show that inferences from the *Random/Teacher* are most restricted, followed by *Random/Random* and *Teacher/Teacher*, and *Teacher/Random* conditions.

vertically and extending horizontally, starting 1 mm from the left hand side of the rectangle. The length of each line was generated from a normal distribution with a mean of 14.5 mm and standard deviation of 3.5 mm. A total of 27 such stimuli were used. The resulting lines ranged in length from 8 mm to 22 mm, with a mean of 14.5 mm and standard deviation of 3.4 mm. Sample stimuli are shown in Fig. 6. These stimuli were attached to cards 90 mm wide and 50 mm high in order to make them easier to handle and to show to participants.

8.1.3. Design

For the teaching task, participants were shown the 27 stimuli and were allowed to choose three examples to teach someone else about the distribution of line lengths.

For the learning task, participants were shown three examples of a novel category, and then asked to evaluate whether a set of other items belonged to the category. Two factors were manipulated in a between-subjects design: the sampling scheme used to generate the examples and the instructions provided to participants about the way in which those examples were generated. Both factors had two levels, corresponding to *Random* generation from the category and generation by an informed *Teacher*. To manipulate the sampling scheme, random examples were generated by randomly choosing 12 sets of three examples (from the total 27 stimuli) and the teaching examples were 11 sets of three examples chosen by participants in the teaching condition. The sampling schemes and instructions were crossed to create four conditions: participants in the *Teacher/Teacher* condition saw three examples generated by a teacher and were told that a teacher had generated them, the *Teacher/Random* condition saw three examples generated by a teacher but were told that they were randomly generated, the *Random/Teacher* condition saw three examples generated randomly but were told that they were generated by a teacher, and the *Random/Random* condition saw randomly generated stimuli and were told that they were randomly generated.

8.1.4. Procedure

For the teaching task, the set of 27 stimuli were shuffled and laid out in front of the participant in an array with three rows and nine columns. The participant then received the following instructions:

> In this experiment, you will see a random assortment of "widgets"—objects consisting of a rectangle with a line inside it. The rectangle is always the same size and the line always starts at the same point, but the line varies in length. If you look closely, you can probably see that widgets are more likely to have lines of some lengths than others. Imagine that you had to teach somebody about the distribution of line lengths that one sees on widgets, but could only do so by showing them three of the widgets in front of you. Which three widgets would you choose?

Participants then selected three widgets from the array, and their choices were noted.

For the learning task, each participant was provided with basic instructions about the task, similar to those used in teaching task:

> In this experiment, you will see some examples of "widgets"—objects consisting of a rectangle with a line inside it. The rectangle is always the same size and the line always starts at the same point, but the line varies in length. If you look closely, you can probably see that widgets are more likely to have lines of some lengths than others.

The participant was then shown three examples of widgets, generated via one of the schemes outlined above. The 27 stimuli themselves were identical to those used in the teaching task. Participants in the *Teacher* instruction condition received the following instructions:

> The widgets in front of you were *specially selected* by a participant in a previous experiment of ours. This participant saw all the objects that were widgets, and was asked to choose three widgets specifically to teach somebody about the distribution of line lengths that one sees on widgets. These examples should thus give you a sense of what makes an object a widget.

Participants in the *Random* instruction condition saw the following paragraph instead:

> The widgets in front of you were sampled at random. The lines inside them are random samples from the distribution of line lengths that one sees on widgets. These examples should thus give you a sense of what makes an object a widget.

Finally, participants in both conditions received the following instructions about the task they would perform:

> We are going to show you some more objects, and we want you to tell us how likely it is that they are widgets (that is, that the line inside the object is the right length for it to be a widget). For each object, indicate on a scale from 0 to 10 how likely you think it is to be a widget, where 0 indicates DEFINITELY NOT a widget, 10 indicates DEFINITELY a widget, and 5 indicates that it is equally likely to be a widget or not a widget.

Each participant was then shown a new set of nine potential widgets, in the same format as the previous stimuli, ranging in length in uniform increments from 1 mm to 28 mm, in random order.

8.2. Results & discussion

To analyze people's choices of examples, we will consider the distribution of which examples were chosen, as well as the triads that people chose. To analyze the triads, we collapsed the 27 examples into three bins based on their distance from the mean. The smallest six examples were considered small (S). The largest six examples were considered large (L). The remaining fifteen examples were considered medium (M). Fig. 7 shows the model predictions and results for the triplets of examples. Strong sampling (Fig. 7a) predicts that examples should be drawn from the true distribution in proportion to their probability—the three examples should tend to be medium-sized. In contrast, pedagogical sampling (Fig. 7b) predicts that examples should be chosen to emphasize the mean and extent of the concept, resulting in one small, one medium, and one large. Nearly half of participants chose triplets composed of one small, one medium, and one large (Fig. 7c). To quantify the differences between the models, we computed the probability of the set of 28 triads under each model and compared the model fits via a likelihood ratio test. The pedagogical model provided a significantly better fit to the data than strong sampling, $\chi^2(1) = 69.04,\ p < .001$.

Fig. 8 shows the model predictions and experiment results for the learning task. To test the prediction that generalization should be broader under the random sampling cover story, we computed a variance score for each participant by converting their responses into a probability distribution by normalizing their ratings. These variances were then submitted to a 2×2 ANOVA with sampling (random or teaching) and cover story (random or teaching) as variables. As predicted, there was a main effect of cover story, $F(1,80) = 10.70,\ MSE = .196,\ p < .005$. There was also a main effect of sampling,

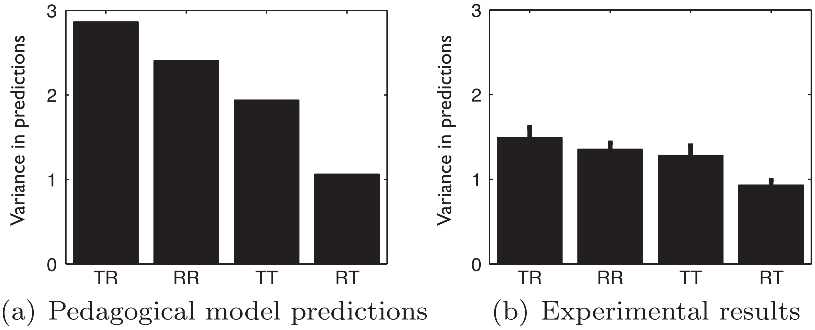


Fig. 9. The variances of predictions by the pedagogical model and people for Experiment 2. The vertical axis represents (a) predicted variance and (b) observed variance in predictions (see Fig. 8). The errorbars represent one standard error of the estimates. The abbreviations on the horizontal axes refer to the *Teacher/Random* (TR), *Random/Random* (RR), *Teacher/Random* (TR), and *Random/Teacher* (RT) conditions, where the labels indicate how the examples were sampled, and what the learner was told about how the data were sampled. The model predicts that the variances of the conditions should be ordered such that $TR > RR > TT > RT$, consistent with the experimental results. Note that if people did not modify their inferences based on the sampling process, we would expect that $(TR = TT) > (RR = RT)$. These results suggest that people's inferences are sensitive to the sampling process.

$F(1,80) = 6.41$, $MSE = .196$, $p < .05$, and no interaction, $F(1,80) = 1.23$, $MSE = .196$, $p = .27$ (see Fig. 9 for plots showing the average variances for each condition).

Together, the results of teaching and learning task provide further support for the predictions of the pedagogical model and the claim that people's inferences differ in pedagogical and non-pedagogical settings. The results of the teaching task showed that when teaching prototype concepts, people do not choose examples randomly, but instead purposefully select examples that indicate the mean and extent of the true distribution, consistent with the predictions of the pedagogical model. The results of the learning task showed that people use knowledge of how data are sampled to guide their inferences. In conditions where examples were ostensibly chosen by a teacher, people's inferences were less broad than in cases where the same examples were described as randomly sampled, consistent with the predictions of the pedagogical model.

9. Causally-structured concepts

In this final section, we investigate the implications of pedagogical reasoning for inferences about causally-structured concepts. Though formal models of causal knowledge are a relatively recent development (Pearl, 2000; Spirtes, Glymour, & Schienes, 1993), they have taken on special prominence in the concept learning literature (Gopnik et al., 2004; Griffiths & Tenenbaum, 2005; Waldmann, Holyoak, & Fratianne, 1995). A variety of researchers have used Bayes nets to capture the causal knowledge that supports reasoning (e.g. Rehder, 2003; Rehder & Hastie, 2001; Shafto, Kemp, Baraff, Coley, & Tenenbaum, 2008).

Bayes nets specify how features of a concept are related and provide a generative model for the values, v, of features, f. The standard relations, which we focus on in this experiment, are noisy-or relations. Noisy-or causal relationships specify that causes are probabilistically sufficient for bringing about their effects. These models are typically associated with two parameters. The first is a background rate, which indicates how likely any individual feature will take the value $v = on$ due to reasons that are not accounted for by the causal model. The second is a transmission rate, which indicates the probability with which, when the node is on, it will cause its children (any node(s) at the end of an arrow) to also turn on.

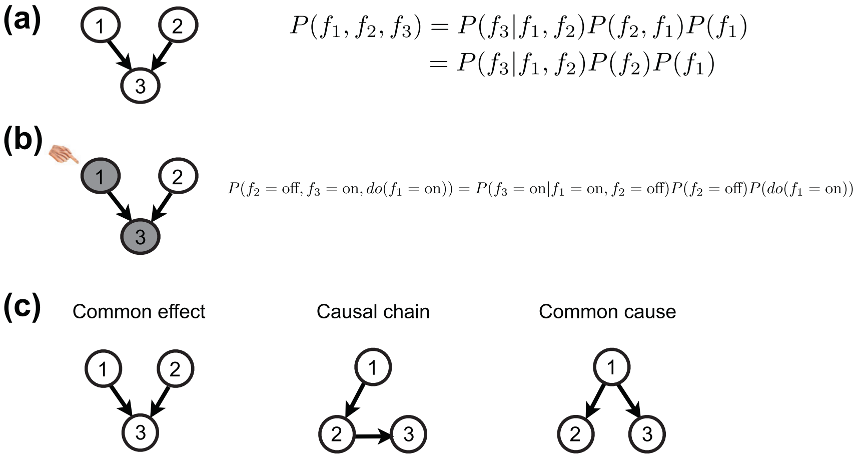
For example, consider the case shown in Fig. 10a. This structure, known as a common effect, specifies how features of the concept are related: features f_1 and f_2 have a common effect, feature f_3. Under noisy-or causal relations, this means that either f_1 or f_2 can individually cause f_3. Causal structures specify dependence and independence relationships among features. In the figure, the fact that there is no arrow connecting features f_1 and f_2 indicates that they are a priori independent, whereas features f_1 and f_3 are dependent, specifying the belief that f_1 can affect f_3 but not f_2. In contrast with prototype concepts, which typically assume that features are independent or simply correlated, causal structures allow construction of hypotheses that capture directional causal relationships. For example, flapping of wings and flying are not merely correlated features of birds, flapping of wings can cause flying.

The equations in Fig. 10a demonstrate the implications of the causal structure for reasoning. The expansion of the $P(f_1, f_2, f_3)$ follows by the chain rule for probabilities. Because features f_1 and f_2 are independent, this equation simplifies. Feature f_2 does not depend on feature f_1, and therefore we can ignore feature f_1 when assessing the prior probability of feature f_2.

In addition to expressing directional probabilistic relationships, Bayes nets can be extended to support reasoning about intervention, resulting in causal Bayes nets. The $do(\cdot)$ operator expresses an intervention from outside the causal network. This has the effect of setting the intervened variable to a particular value, and breaking incoming causal links to that node. Fig. 10b shows one possible intervention and the likely results. The equation shows how the probability of this state would be evaluated, given the common effect hypothesis. The critical difference between evaluating the scenario in Fig. 10a and 10b is in understanding how interventions are generated, for example assessing the probability $P(do(f_1 = on))$.

Research on causal learning and reasoning has typically focused on learning causal relationships among sets of three variables, as these represent the basic forms that are combined to create larger causal networks. Fig. 10c shows the three basic cases, called common effect, causal chain, and common cause. Each has a different characteristic structure and the implications of interventions on variables differ for each case. For instance, in the case of the common effect structure, no single intervention is likely to turn on all of the variables, while in the causal chain and common cause structure, interventions on f_1 will tend to turn all features on.

In the standard formulation of causal learning, learners observe interventions i and the resulting values of the features v. Together these constitute the data, and the generative model gives $P(v, i|h)$. The goal of learning is to infer the latent causal structure. That is, given the interventions and the observed results, infer the causal structure that was most likely to have generated these data. In these approaches, interventions are assumed to be independent and chosen at random, as in weak or strong sampling, and the resulting values for the features are generated according to the causal model, h.



$$P(f_1, f_2, f_3) = P(f_3|f_1, f_2)P(f_2, f_1)P(f_1)$$
$$= P(f_3|f_1, f_2)P(f_2)P(f_1)$$

$$P(f_2 = \mathrm{off}, f_3 = \mathrm{on}, do(f_1 = \mathrm{on})) = P(f_3 = \mathrm{on}|f_1 = \mathrm{on}, f_2 = \mathrm{off})P(f_2 = \mathrm{off})P(do(f_1 = \mathrm{on}))$$

Fig. 10. Figure showing a hypothesis about the causal structure of a concept. (a) A common effect hypothesis where feature f_1 or feature f_2 can cause feature f_3, with the probability of observing data given the hypothesis. (b) The probability of observing data given the hypothesis and an intervention. (c) Three basic causal structure templates.

If a teacher was allowed two interventions, which ones should they choose? The pedagogical model can be used to generate predictions about which sets of interventions teachers will choose and the inferences that learners will make based on pedagogical interventions (see Fig. 11 for predictions and Appendix for full details about model implementation). For common effect networks, the answer should be fairly clear—given two interventions, the teacher should intervene on feature f_1 showing the learner that feature f_3 will likely turn on but feature f_2 will not, and then intervene on feature f_2 showing that feature f_3 will likely turn on but feature f_1 will not. Similarly, for the causal chain, the model predicts that the teacher will choose features 1 and 2. For the common cause, the situation is somewhat more interesting, as there is not a pair of interventions that would fully disambiguate the possibilities. The model predicts two possible strategies. The most likely is that the teacher will intervene on feature f_1 twice. Note that this is ambiguous: the data are consistent with both a common cause and two causal chains. The second most likely is choosing feature f_1, and then feature f_2 or feature f_3. These are also both ambiguous between the common cause and causal chain cases.

Unlike previous cases, causal inferences from a teacher's interventions are generally strong whether the learner assumes the teacher is helpful or not. Consider the case of the common effect structure. Given the teacher's interventions on features f_1 and f_2 and the likely data that would be observed, the inference to a common effect structure is relatively straightforward. However, knowledge of the teacher's intent is important in two kinds of situations: when evidence is ambiguous and when surprising events occur. These cases are shown in Fig. 11b and 11c. The ambiguous cases revolve around the common cause structure, where the data are consistent with both the common cause and causal chain. If the teacher intended to teach the causal chain, then they would have chosen an intervention that turned on all of the nodes, then one that turned on only the bottom two. Similarly,

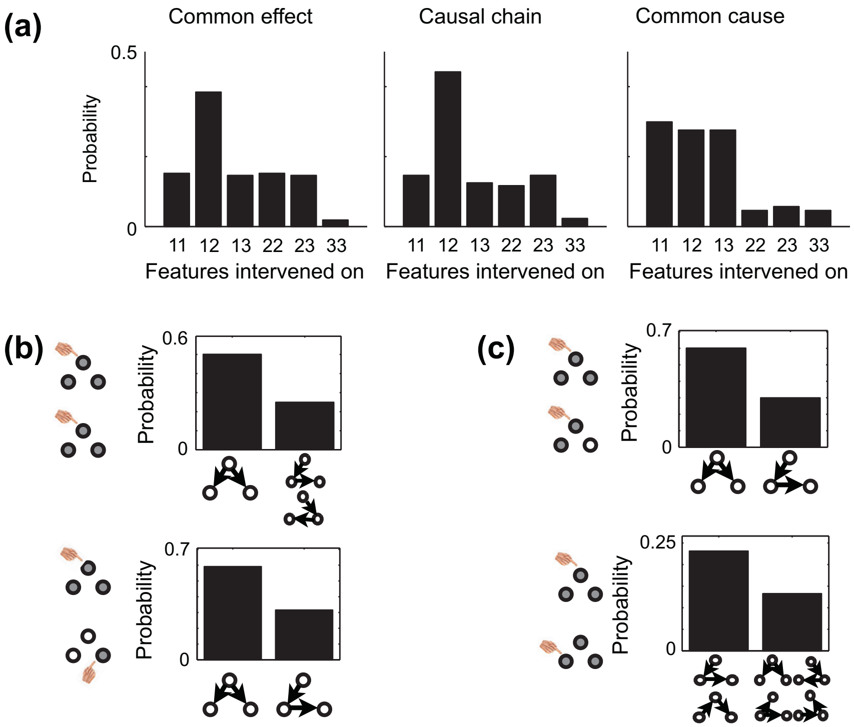


Fig. 11. Predictions of the pedagogical model for Experiment 3: (a) the *Teaching* condition and (b & c) the *Pedagogical Learning* condition. Numerical labels in (a) correspond to the features shown in Fig. 10(b) and (c). The predicted probability of possible causal structures for *Pedagogical Learning* from (b) two kinds of ambiguous data and (c) two kinds of surprising data. For the cases in (b) & (c), random sampling predicts no differences.

for the surprising events, the teacher's intention can be used to infer the correct structure. Fig. 11c (top) shows a case where the same intervention brings about two different results that are equally consistent with a common cause and chain allowing one surprise event. However, the choice of the same intervention twice allows the learner to infer that the intended structure must have been common cause. The complementary case is shown in Fig. 11c (bottom) where the choice of two different interventions can be used to make the opposite inference.

10. Experiment 3: Teaching and learning causally-structured concepts

To test the model predictions, we contrasted pedagogical causal reasoning with non-pedagogical causal reasoning in three conditions as in Experiment 1: *Teaching-Pedagogical Learning*, *Pedagogical learning*, and *Non-Pedagogical Learning.*

10.1. Method

10.1.1. Participants

Eighty-six University of Louisville undergraduates participated in exchange for course credit. Participants were randomly assigned to one of two conditions: *Teaching-Pedagogical Learning* ($n = 30$), *Pedagogical Learning* ($n = 30$), or *Non-Pedagogical Learning* ($n = 26$).

10.1.2. Design

The experiment consisted of two parts. In all conditions, participants saw different causal structures depicted on a computer and were allowed to familiarize themselves with them by intervening on variable and observing the effects of their intervention. They then participated in the first part, either a teaching task (in the teaching condition) or an exploration task (in the other conditions), followed by the second part, a learning task.

10.1.3. Procedure

Participants were seated at a Apple Mac Pro desktop computer for the familiarization task and told that they were going to learn about causal relationships between sets of three variables. Participants saw three variables with novel names on the screen (e.g. "ziffing") arranged in a triangle shape. Arrows between variables depicted causal relationships between variables. Experimenters explained that by turning one variable on, participants could observe which other variables would likely turn on. The experimenter also explained that causal relationships were probabilistic—sometimes causes did not turn on effects and sometimes variables turned on for no reason. Participants were encouraged to try setting different variables and to intervene multiple different times. In the *Teaching-Pedagogical Learning* condition, when people were comfortable with the causal relationships, participants were then asked to choose two different interventions to show a learner (who could not see the arrows) how the variables were related. Participants did not observe the results of their teaching interventions. Each individual participated in six trials of the teaching task in random order, with two trials for each of three causal structures (common cause, common effect, chain). For each structure, the causal relationships between the three variables was randomly permuted to control for preferences among different orientations of the structures. In the *Pedagogical Learning* and *Non-Pedagogical Learning* conditions, participants were asked to choose two interventions to try to discover how three variables were related. They also saw a total of six trials.

When they completed the teaching task, participants moved to the learning task. The learning task was conducted on paper. Each participant was given a booklet with ten scenarios. Each scenario included two pictures. Each picture indicated an intervention, and for each intervention, which other variables were on or off (the intervened upon variable was always on). Scenarios were presented in four pseudorandom orders. In the *Pedagogical Learning* conditions, participants were told that the interventions were chosen by a teacher, with the intention of helping them learn. In the *Non-Pedagogical Learning* condition, participants were told that the interventions were chosen by someone who was trying to figure out the causal relationships. Participants were reminded that because events were probabilistic, sometimes surprising things happen. For each scenario, after seeing the data, participants indicated which of 12 possible causal structures were most likely on a scale of $0 - 100$, with 0 indicating definitely incorrect and 100 indicating definitely correct. When participants completed the task, they were debriefed and thanked.

10.2. Results & discussion

In the teaching task, the key question is which pairs of interventions people choose. The pedagogical model predicts that certain pairs of interventions are better than others for each of the three causal structures (see Fig. 12). For the common effect condition, people's choices were highly non-random, $\chi^2(5) = 115.08,\ p < 0.001$. The pedagogical model predicts that intervening on each of the two causes is the best solution. People's choices were strongly correlated with the predictions of the pedagogical model, $r = 0.96$. For the causal chain condition, people's interventions were again highly non-random, $\chi^2(5) = 124.32,\ p < 0.001$, and were strongly correlated with the predictions of the pedagogical model $r = 0.94$. Similarly, in the common cause condition, the interventions that people chose were highly non-random, $\chi^2(5) = 74.38,\ p < 0.001$, and strongly correlated with the pedagogical model, $r = 0.97$. These data show that people's interventions were highly non-random and that the pedagogical model accurately predicts which pairs people choose in order to teach a learner about the causal structure.

For the learning task, people provided numerical ratings on a $0 - 100$ scale. To facilitate comparison with the model predictions, each participant's responses for each question were converted into probabilities by normalizing the ratings for each question, so that the ratings for the 12 causal structures summed to one.

Do learners make use of pedagogical situations to make stronger inferences? To test whether learners appreciate the implications of pedagogical situations, we turn to the cases where the models make different predictions: the two ambiguous data questions, and the two surprising event questions. In all cases, the model based on random selection is ambivalent about the best causal structure given the data. In contrast, the pedagogical model predicts that learners' inferences should reflect an understanding of the intent of the teacher, allowing learners to make confident inferences about which of the causal structures the teacher meant to teach.

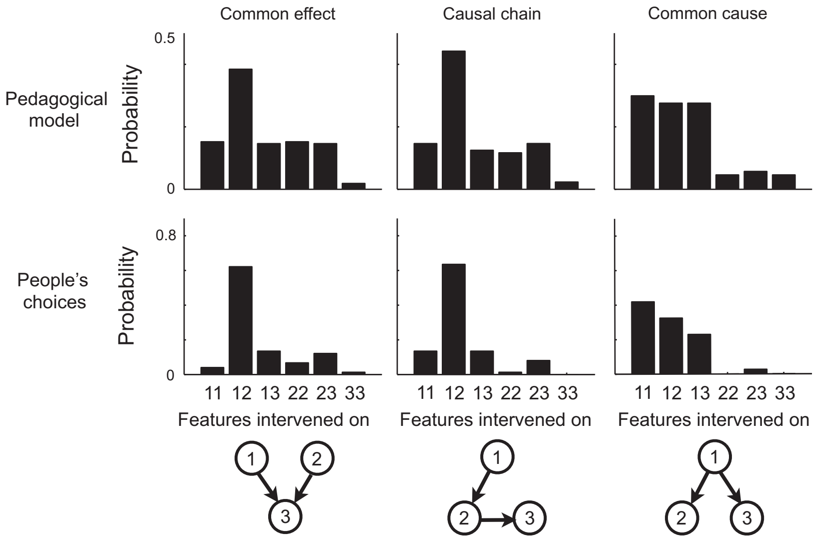


Fig. 12. Results for the teaching task from Experiment 3. Bar charts showing the pairs of interventions chosen by participants in the teaching task. Below the horizontal axis are the causal structures with the nodes labeled for reference. On the horizontal axis are possible pairs and on the vertical axis is the proportion (as a probability) of choices.

Fig. 13 shows the data from the two ambiguous cases. For the *Teaching-Pedagogical Learning* condition, there were significant differences, $t(88) = 2.22,\ p < 0.05$ and $t(58) = 4.21,\ p < 0.001$. For the *Pedagogical Learning* condition, there were significant differences, $t(88) = 5.11,\ p < 0.001$ and $t(58) = 2.98,\ p < 0.005$. For the first case, the *Non-Pedagogical Learning* condition showed no significant difference, $t(76) = 0.73,\ p = 0.47$. For the second case, there was a significant difference, $t(50) = 2.24,\ p < 0.05$. To test whether the differences in the *Pedagogical Learning* conditions were larger than that in the *Non-Pedagogical Learning* condition, we ran two 2 × 2 ANOVA with condition as a between-subjects variable and question as a within-subjects variable. If people in the *Pedagogical Learning* conditions showed a larger preference for the common cause structure, then we would expect a significant interaction between question and condition. We found that, indeed, people in the *Teaching-Pedagogical Learning* condition showed a larger preference for the common cause explanation, $F(1,54) = 9.33,\ p < 0.01$, as did people in the *Pedagogical Learning* condition, $F(1,54) = 3.88,\ p = 0.054$.

Fig. 14 shows the model predictions and experimental results for learning from surprising events. In the *Teaching-Pedagogical Learning* condition, we found the differences predicted by the pedagogical model, $t(58) = 2.19,\ p < 0.05$ and $t(178) = 2.01,\ p < 0.05$. In the *Pedagogical Learning* condition, we found no differences, $t(58) = 0.40,\ p = 0.69$ and $t(178) = -0.63,\ p = 0.53$. In the *Non-Pedagogical Learning* condition, there were no significant differences, $t(50) = 0.15,\ p = 0.88$ and $t(154) = -0.51,\ p = 0.61$. As in Experiment 1, engaging in teaching appears to have affected pedagogical learning.

Together these results suggest that people understand the implications of pedagogical situations for teaching and learning causal structures and the pedagogical model captures the logic underlying their inferences. When asked to choose interventions for the purpose of teaching a learner, people choose the pairs of interventions that are predicted to be most helpful by the model. In learning situations, people reason differently when examples are provided by a teacher. People capitalize on their understanding of the teacher's intention to help to disambiguate potentially uncertain situations. For surprising events, having previously engaged in teaching facilitates pedagogical learning. In

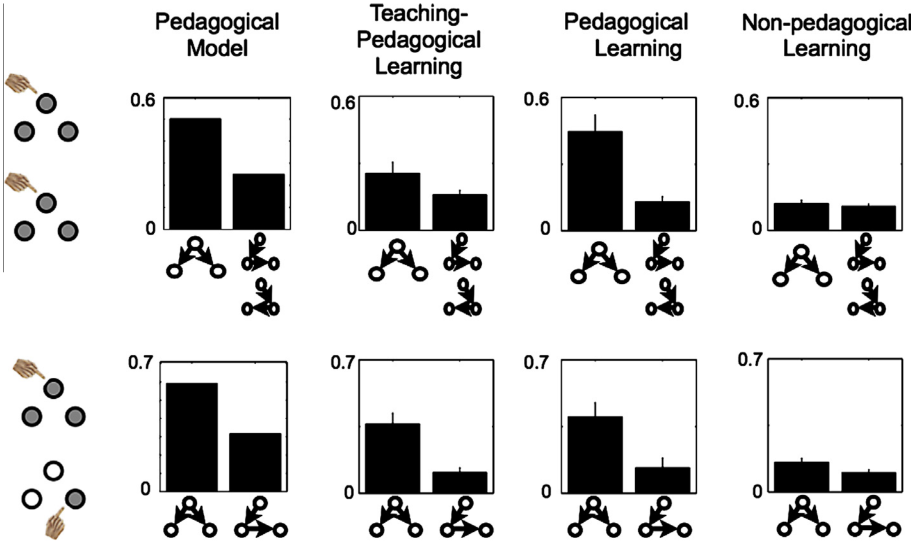


Fig. 13. Model predictions and human data from the *Teaching-Pedagogical Learning*, the *Pedagogical Learning*, and *Non-Pedagogical Learning* conditions for the two ambiguous scenarios in Experiment 3. The top row shows the case where the teacher chooses the same node twice and all nodes turn on (as indicated in the picture to the left of the graphs). These data are ambiguous between the common cause and the causal chain structures shown on the horizontal axis. The bottom row shows the case where the teacher chooses one node resulting in all nodes turning on and a second node and no other nodes turn on. These data are also ambiguous between the common cause and the chain structures.

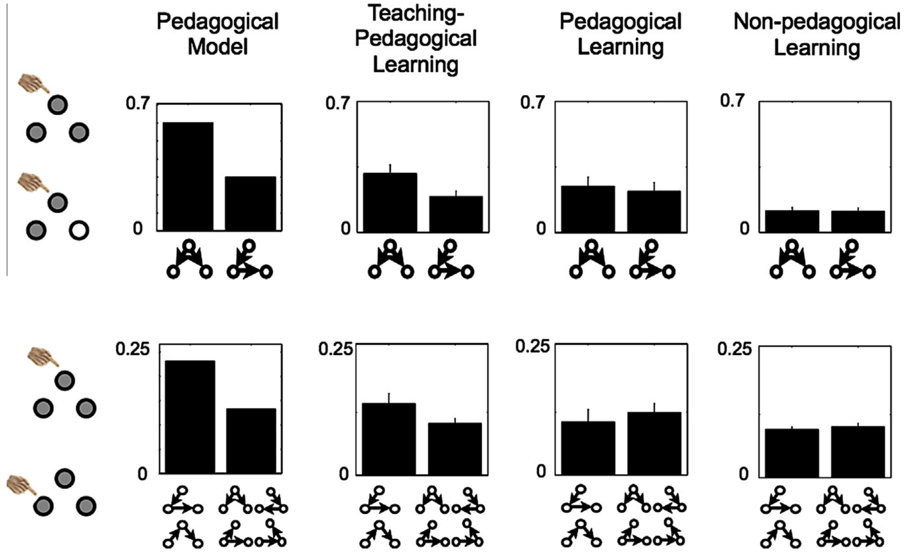


Fig. 14. Model predictions and human data from the *Teaching-Pedagogical Learning*, *Pedagogical Learning*, and *Non-Pedagogical Learning* conditions for the two surprising data scenarios in Experiment 3. The top row shows the case where the teacher chooses the same node twice and all nodes turn on once, but not the second time. These data are ambiguous between the common cause and the causal chain structures shown on the horizontal axis—each requires the failure of one causal link. The bottom row shows the case where the teacher chooses one node resulting in all nodes turning on and a second node and all nodes turn on again. These data are also ambiguous between the two the chains and the structures shown on the horizontal axis.

contrast, when told that the same data were generated at random, people's inferences reflect the ambiguity of the situation.

11. General discussion

We have presented a computational model of pedagogical reasoning, addressing which examples teachers should choose to teach concepts and what inferences learners should make based on this purposefully sampled data. The model predicts the examples that people will choose to teach different hypotheses, and that people will draw systematically stronger and qualitatively different inferences in pedagogical learning situations. We presented three experiments testing the predictions of this model using the teaching games method. Each experiment investigated a different kind of concept: rule-based, prototype, and causally-structured. Together the results of these experiments showed that when teaching, the examples that people choose are well-predicted by the model. Similarly, in pedagogical situations, people's inferences are predicted by the pedagogical model, while in non-pedagogical situations, people's inferences are instead consistent with predictions based on random sampling. Taken together, the results suggest that people differentiate pedagogical and non-pedagogical situations and the pedagogical model accurately captures inferences in pedagogical situations. In the following, we connect these results to previous findings, then consider broader implications.

11.1. Connections to previous results

Our research builds on considerable research investigating learning concepts from examples, as well as research investigating how explicit teaching might affect such learning. Specifically, we consider relationships to previous research on learning, including research investigating learning from positive examples and previous models of concept learning. We also consider connections to previous research on teaching by example, pragmatics, and order effects in learning. We conclude this section by discussing connections to previous research on pedagogical reasoning.

11.1.1. Learning from positive examples

Our findings build on and extend results suggesting that word learning reflects an assumption that sampling is constrained by knowledge of intentionality (Xu & Tenenbaum, 2007a, 2007b). In these experiments, participants' generalizations of the learned labels differed when the examples were chosen by a teacher from when the examples chosen by an agent who was ignorant of the true meaning. Consistent with these results, we found that people systematically differentiate between helpfully sampled data and data sampled by an agent ignorant of the true concept across three qualitatively different learning problems. We also extend their results by showing that, when asked to choose examples for teaching, people choose data that are helpful to the learner. This is the key difference between strong sampling and pedagogical sampling: the assumption that examples are sampled with the purpose of helping the learner.

Differences between pedagogical sampling and strong sampling are somewhat less stark for learning. Our findings in the learning condition of Experiment 1 are broadly consistent with the qualitative predictions of strong sampling, when strong sampling applies. For the case of two positive examples, strong sampling also predicts that the inferred concept should be relatively constrained, though not as strongly as the pedagogical model predicts.

Experiments 2 and 3 provides cases where the predictions of pedagogical and strong sampling diverge for the learning task. For Experiment 2, examples in both the random and the teaching condition are drawn from the true concept, and therefore strong sampling does not make differential predictions. Thus, the finding that people generalize less broadly in the teaching condition cannot be explained by strong sampling. Similarly, for Experiment 3 strong sampling predicts that interventions should be chosen at random from the possible interventions with effects and therefore the intervention itself provides no information about the underlying structure to the learner. But in pedagogical causal learning, our evidence suggests that the teacher's intent to help plays the critical role in disambiguating ambiguous and surprising data, as predicted by the pedagogical model.

The pedagogical model provides more accurate description of both teaching and learning by formalizing how teachers choose data to help learners. It is important to reinforce that the pedagogical model is a generalization of random sampling and can therefore capture the same phenomena by allowing that the teacher not choose data helpfully (setting $\alpha = 0$). An interesting extension within this framework is learning whether an individual is helpful by inferring a value of α based on the data they provide.

11.1.2. Previous results on teaching by examples

Though very few studies have investigated teaching by examples in well-controlled settings, two notable examples exist, which demonstrate adult teaching of linearly separable concepts (Avrahami et al., 1997) and children teaching in a property induction task (Rhodes, Gelman, & Brickman, 2010). We turn our attention to their results, and the relationship between the predictions of the pedagogical model and their results.

Avrahami et al. (1997) investigated adults' teaching of linearly separable concepts. In these experiments, people were presented with stimuli that varied on two dimensions, size of the semicircle and the angle of a radial line in the semicircle (varying from acute to obtuse). They were taught one of two kinds of categories: categories for which the boundary was defined in terms of a single dimension (e.g. small semicircles) or categories for which the boundary was defined in terms of both dimensions (e.g. small semicircles with an acute line). Notably, although these categories are rule-based, discriminating nearby exemplars is difficult. In the teaching task, people were shown a subset of the possible examples and asked to choose examples to teach someone else the concept. Their results show that people's choices were systematic (see Fig. 15). People show a general preference for positive examples and tend toward examples that are away from the boundary.

To test whether our model predicted a similar pattern of results, we implemented a version in which the learner inferred the boundary in the presence of perceptual noise (using a logistic function, see Reed, 1972). The results (see Fig. 15) show that there is a good qualitative fit between the

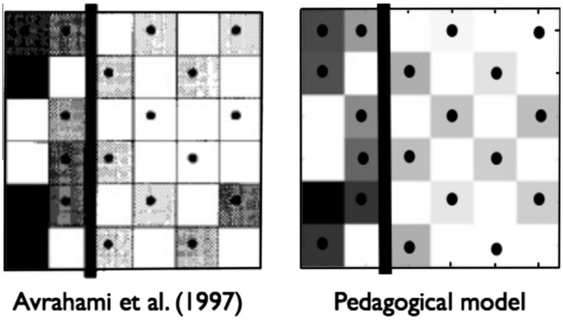


Fig. 15. Comparison of sets of three examples chosen in Avrahami et al. (1997) and the predictions of the pedagogical model. In the experiment, participants were teaching linearly separable concepts using examples. The grid represents the possible examples on each dimension; the black line indicates the boundary. Dots indicate examples that were available to use for teaching. This task is notable for using rule-based concepts with perceptually confusable stimuli. Dark patches indicate examples that were more likely to be chosen. The results show that people have a broad preference for positive examples, and for examples that are away from the boundary. In the pedagogical model, we formalize learning a boundary between perceptually confusable (noisy) stimuli as learning a logistic function. The logistic function captures the fact that because the stimuli are noisy, the examples that are close to the boundary risk being confused with the other category. Though there are quantitative differences, the pedagogical model captures the broad preference for positive examples, and for examples that are away from the boundary.

predictions of the pedagogical model and the results observed by Avrahami et al. (1997), with both preferring examples that are far from the boundary and therefore less confusable. These results provide further support for the pedagogical model as a model of teaching by example.

Rhodes et al. (2010) investigated six-year-old children's teaching in a property induction task. In their teaching task, children were told about a novel property (e.g. has a four-chambered heart) that was true of a subset of animals (e.g. dogs). The children were then asked to choose a set of three examples by which to teach the concept, and were given the choice of three examples of a subset of the concept (e.g. dalmations) or diverse set of examples (e.g. a golden retriever, a dalmation, and a collie). As in the rectangle game, where the preferred positive examples are in opposite corners of the concept, the pedagogical model predicts that a diverse set of examples is preferred for teaching the concept. Rhodes et al. (2010) find that children were significantly more likely to choose the diverse set of examples, suggesting that children understand the implications of different sets of examples for teaching concepts.

11.1.3. Teaching versus communicating: connections to pragmatics

Considerable research in language has focused on how choice of utterances are related to inferences of listeners (e.g. Brennan & Clark, 1996; Fussell & Krauss, 1989; Krauss, 1987), how people negotiate common ground (e.g. Garrod & Anderson, 1987; Pickering & Garrod, 2004; Wilkes-Gibbs & Clark, 1992) and common reference (e.g. Brennan & Clark, 1996; Fusaroli et al., 2012; Voiklis & Corter, 2012), and formalizing pragmatics via signaling games (Benz, Jager, & van Rooij, 2005; Frank & Goodman, 2012). Among the most closely related of these works is a recent paper by Voiklis and Corter (2012), which investigated whether negotiating reference affected category learning. Participants engaged in an paired interaction in which one described an object and the other predicted what category the object was in based on the description. Participants engaged in multiple trials, alternating roles with each trial. The results showed that participants in this dialogue condition learned the categories better than participants in a monologue condition (in which they only engaged in self-talk).

There are many similarities between our work and this literature, most notably the focus on how people come to arrive at common beliefs. There are also a variety of ways in which our focus, and teaching more generally, differs from communication. In our experiments, we have focused on cases in which the goal is to select examples a priori. Thus, we have not focused on iterative process related to negotiating how to refer. This reflects a difference between teaching and communication, namely that the object of teaching is some objective truth, while the object of communication can be a great variety of things including to simply come to agreement (whether true or not). Consistent with this, we have focused on natural signs—examples that have meaning—while the communication literature most often focuses on arbitrary signs—such as words—for which meaning must be learned or negotiated. Similarly, our model also differs from those proposed in the language literature on signaling games. The focus of that work is on the communication of specifics while in ours the focus is on teaching of generalities (Benz et al., 2005; Frank & Goodman, 2012). These are examples of how teaching, and our experimental approach, differ from communication more generally; however, a more precise formalization of how these two context are similar and different is an important direction for future work.

11.1.4. Selecting examples to facilitate learning

Many researchers have noted that certain examples or orderings of examples could facilitate or impede learning (Elio & Anderson, 1981, 1984; Goldstone & Sakamoto, 2003; Goldstone & Son, 2005; Mathy & Feldman, 2009; Medin & Bettger, 1994; Stewart, Brown, & Chater, 2002). Aside from the fact that data in our experiments were presented together, rather than ordered, our approach differs from this research in two main respects. First, the pedagogical model highlights the effects of pedagogical context in interpreting examples; we have shown that the very same data lead learners to different inferences in pedagogical versus non-pedagogical situations. Second, our approach to facilitating learning is derived from an analysis of the implications of social inferences—inferences about others' knowledge and intent—for learners and their learning.

Similarly, heated debates continue in the education literature about the relative merits of direct instruction (i.e. teaching) versus discovery learning. In this literature, direct instruction is commonly perceived as facilitating learning by providing 'good' data (Dean & Kuhn, 2006; Kirschner, Sweller, & Clark, 2006; Klahr & Nigam, 2004; Mayer, 2004; Rittle-Johnson, 2006). In related work, we have shown that this common perception overlooks an important consequence of teaching that is predicted by the pedagogical model—teaching not only shows what is true, but also provides information about what is not true (Bonawitz et al., 2011). In this work, we show that preschool-aged children understand the implications of teaching situations; they engage in less exploratory play after demonstrations by a knowledgeable teacher, but not after observing the same data presented by a not knowledgeable teacher. This provides another demonstration that learning depends on inferences about the demonstrator's knowledge and intent, and the consequences of these social inferences are predicted by our pedagogical model.

11.1.5. Two aspects of pedagogical reasoning

Establishing a pedagogical context can have two different kinds of consequences for the assumptions that a learner might make. First, it can affect the learner's expectations about the way in which the data provided by the teacher have been sampled, with learners expecting that teachers will provide informative data. Second, it can affect what kinds of concepts the learner might expect a teacher to convey. Our model of pedagogical reasoning has focused on the first of these two aspects of pedagogical reasoning, examining how people sample data in order to teach concepts. However, the second aspect of pedagogical reasoning could be equally important in determining the conclusions that learners reach, and in establishing a role for pedagogy in supporting cumulative cultural evolution.

Csibra and Gergeley (Csibra & Gergely, 2009, 2006; Topal et al., 2008) have focused on this second aspect of pedagogical reasoning. Their proposal is that when children are engaged in a pedagogical context through ostensive cueing, an assumption of semantic generalizability of information is engaged (Csibra & Gergely, 2009). For example, Topal et al. (2008) explain the A-not-B task results by arguing that the child assumes that the A bin is 'for toys'—that is, that the pedagogical situation was intended to teach information about the kind of thing that the A bin is, as opposed to simply indicating that the toy is in the A bin. Under this proposal, pedagogy provides a mechanism for communicating information about kinds and their properties, dealing with the challenge of communicating such generalizations.

Our sampling-based approach and Csibra and Gergely's work represent different, complementary approaches to the same problem. From our perspective, their proposal that pedagogical situations engage an assumption of semantic generalizability is a kind of context-sensitive prior (Shafto, Kemp, Baraff, Coley, & Tenenbaum, 2005, 2008) that applies in pedagogical settings. In contrast, the sampling assumptions we have emphasized influence the likelihood assumed by the learner, since they determine the relationship between hypotheses and data. However, these two aspects of pedagogy can be brought together quite naturally within the more general Bayesian framework that we have used to develop our account. Bayes' rule provides a way to combine this prior expectation of generalizability with the assumption that the teacher will choose data to help the learner infer the intended concept. Investigations of how to implement a prior on semantic generalizability and the implications of combining an assumption of generalizability with the assumption of pedagogical sampling are interesting directions for future research.

11.2. Implications

We consider three implications of our results. First, we consider the implications of the order effects observed in Experiments 1 and 3. Second, we discuss the implications of these results for debates on learning more generally, which have mainly focused on the representation of concepts. Finally, we discuss implications for cultural evolution and conclude with a brief summary.

11.2.1. Effects of teaching first on later learning

One interesting aspect of the current results is the finding that participants who played the role of teacher first made inferences that more closely approximated the model predictions later. In Experiment 1, participants in the *Pedagogical Learning* condition showed two of the three predicted effects, while participants in the *Teaching-Pedagogical Learning* conditions showed all three of the predicted effects. In Experiment 3, participants in the *Pedagogical Learning* condition showed two of the four predicted effects, while participants in the *Teaching-Pedagogical Learning* condition showed all four of the predicted effects. Importantly, in both experiments, control conditions showed none of the predicted effects. Together, these results suggest that engaging in teaching first facilitated pedagogical learning.

One possible explanation for this result is that the underlying cognitive mechanisms provide solutions that approximate the predicted responses. The model suggests recursive reasoning where the learner must consider what conclusions the teacher would use for possible concepts (and so on). If we imagine this recursive reasoning as a process-level account, it becomes clear that it is extremely demanding and the resource demands grow with the recursion depth (cf. Colman, 2003; Hedden & Zhang, 2002). One way to deal with these demands is to use approximate or pre-computed values instead of recursing. In particular, if a participant first has experience in the role of teacher, she may store some representation of what examples are likely to be used to convey different concepts, and then when put in the role of learner she may use this "cached" information rather than computing the full recursion on the fly. If this is right, we expect participants to be more optimal to the extent that they have previous experience teaching in a given domain. An important direction for future research is considering possible algorithmic-level approximations that adapt based on experience and whether they explain these effects.

11.2.2. Concept learning as an interaction between concept representation and sampling

Previous approaches to understanding concept learning have focused on how representational commitments provide biases for learning. Connectionists have focused on how the conjunction of simple network structures and connection weight update rules can allow effective extraction of information from observed data (Rogers & McClelland, 2004). Simplicity-based approaches have explored how a generic approach based on compressing redundant information out of data sets can learn predictive relationships (Pothos & Chater, 2002). Similarly, the theory-based Bayesian approach has explored how learning structured generative models of observed data can allow meaningful learning from relatively minimal data (Tenenbaum et al., 2006). The overarching commonality across these approaches is the search for generic learning biases that allow a learner to generate predictions about new situations based on relatively limited data.

A contribution of our work is in highlighting the influence of how data are sampled on learning (see also Tenenbaum, 1999; Xu & Tenenbaum, 2007a, 2007b). One of the strengths of the Bayesian approach is that it has, in part, motivated these questions and provides a natural framework in which to integrate information about how data are sampled with our prior beliefs to produce new beliefs about the world. It is not obvious to us what the most natural way of integrating knowledge about sampling in either connectionist or simplicity-based approaches to learning.

The interaction between sampling and knowledge representation turns out to be non-trivial, a point that may not be obvious when one focuses on randomly (weak) sampled data. Consider, for example, the problem of learning in the rectangle game. Assuming randomly sampled data, it will take a very long time on average to infer the correct answer. Indeed, to arrive at the correct conclusion with complete certainty requires six particular examples—four negative examples that mark the outside boundaries of the concept, and two positive examples to mark the interior boundaries—and it could take quite a long time to get those examples by random chance. In contrast, for a pedagogical learner receiving helpfully sampled data, a strong inference can be made on the basis of only two examples. Similarly, in both prototype learning and causal learning, learning can be expedited by understanding which data a teacher is likely to sample. In each of these domains, qualitatively different patterns of evidence are expected based on which hypothesis is being taught and the relationship among different competing hypotheses.

Teaching is not the only kind of situation in which an understanding of how data are sampled may play a role in learning. Many researchers have noted that people can learn by observing other's actions (Baldwin, 1993; Gergely, Bekkering, & Kiraly, 2002; Lyons, Young, & Keil, 2007; Meltzoff & Moore, 1977; Shafto, Goodman, & Frank, 2012). Recent research has formalized how inferences about the goal of another agent can facilitate causal learning (Goodman, Baker, & Tenenbaum, 2009). Learning from goal-directed actions differs from learning from pedagogically sampled data in the strength of the inferences that they afford (Shafto et al., 2012). Consider a causal learning scenario, as in Experiment 3. If the agent has the goal to turn on one of the variables, then the most direct course of action is to turn it on. Similarly, if the agent's goal is to turn on some pair with a single action, then they should intervene on the cause, rather than the effect. These situations provide information about the relationship between interventions and effects; however, they do not provide the kind of disambiguating information that is predicted by the pedagogical model and observed in our experiments. Pedagogical situations afford stronger inferences because the teacher's goal is to teach the hypothesis to the learner, and this leads the teacher to sample data that disambiguate different possible hypotheses.

11.2.3. Implications for cultural evolution

Csibra (2007) has argued that a uniquely human ability to understand pedagogical situations allows us to accumulate knowledge over generations, an ability that separates us from even our closest evolutionary relatives. He argues that not only do people naturally teach, thus providing better data to learn from, but learners naturally understand the implications of pedagogical situations for learning. We have presented a computational framework that formalizes which examples people should choose and the inferences that learners should infer in pedagogical contexts. The model formalizes the claim that simply receiving good data does not capture all that pedagogical situations have to offer—when learners are aware that they are in pedagogical situations, they can make stronger inferences than if they are unaware that the teacher is being helpful. Each of our experiments provides evidence that learners in pedagogical situations use the knowledge that teachers are being helpful to guide their inferences, and those inferences are stronger than those that are warranted based on the data alone.

It is an open and important question whether this kind of pedagogical reasoning supports the accumulation of knowledge through generations. Recent research has formalized the cultural evolution within a Bayesian framework (Griffiths & Kalish, 2007) and developed experimental methods for testing the model predictions. Their results confirm that in standard learning settings, where data are randomly sampled, knowledge is not accumulated over generations. These methods of studying cultural evolution, together with our model and experimental methods, provide formal computational and experimental methods for testing whether pedagogical situations allow for the accumulation of knowledge over generations.

11.3. Conclusions

Understanding how explicit teaching affects learning is a critical issue in the study of human learning. We have presented a computational model of pedagogical teaching and learning, and evidence from three experiments suggesting that the model predicts people's behavior in teaching and learning situations. We have focused on cases of simple concept learning, cases that are much simpler than those that may be encountered in educational contexts. A critical issue for future work will be generalizing these results to more ecologically valid domains and where the problem of learning includes the possibility of encountering novel concepts. These results provide a first step toward understanding how pedagogical situations affect learning, and present a new framework within which we may explore implications for education and cultural evolution.

Appendix A

To formalize the model, we must specify four basic elements that capture the experimental tasks. First, we must specify the hypothesis space, $\mathcal{H}$, the set of possible hypotheses that one may have about the true concept. Second, we must specify the prior probabilities of each hypothesis, $P(h)$; throughout, we have designed situations in which all hypotheses are equally plausible to focus on the effects of sampling. Third, we must specify the set of possible data, D; the examples the teacher may choose. Fourth, we must specify the likelihoods, $P(d|h)$, the probability of randomly sampling each possible datum under each possible hypothesis, assuming random sampling.

Given these elements, the model can be used to generate predictions about the behavior of teachers and learners. In the following, we first work through an illustrative example in detail, specifying the four elements and working through how the model generates predictions. We then provide modeling details for each of the experiments. Because the experiments include many hypotheses and many possible data, we cannot work through these in the same detail as the example; instead, we provide the elements necessary to implement the models in each case and describe how the predictions in the figures were derived.

A.1. Computational modeling: An illustrative example

To illustrate how the pedagogical model derives predictions, we will describe in detail a simpler case of the rectangle game, a line game. In the line game, concepts correspond to line segments on an interval. Fig. 16 shows a simple version of this problem, where the hypothesis space consists of 6 possible hypotheses (see Fig. 16a). We assume a uniform prior over the hypotheses, i.e. $P(h) = \frac{1}{6}$. Assuming data, D, are labeled and chosen in pairs, there are 12 possible examples (see Fig. 16b). The likelihood, $P(d|h)$, is based on sampling examples uniformly at random from among those consistent with a given hypothesis (see Fig. 16c).

To see how the model generates predictions, we begin be considering what would happen if the data were sampled randomly, $P(d|h)$ (Fig. 17a). For rule-based concepts, random sampling implies choosing among the examples that are consistent with the hypothesis with uniform probability. Consider h_3, there are three data that are consistent with $h_3 : d_1,\ d_5$, and d_9. Thus, the probability of choosing each is $\frac{1}{3} \approx .33$. Indeed, for all of the hypotheses, there are exact three data consistent, though which data are consistent varies by hypothesis.

The problem from the perspective of the teacher is to choose the two best examples with which to teach the target concept (here, h_3). Eq. (2) suggests that the teacher should choose the examples that tend to maximize the learner's belief in the correct hypothesis. Eq. (2) tell us how to identify these examples, as well. For example, to compute the probability of choosing d_1, we must first find the probability of the learner inferring the correct hypothesis, given this example, $P(h|d)$. Assuming that hypotheses are equally likely a priori, $P(d_1|h_3) \times P(h_3) = \frac{1}{3} \times \frac{1}{6} \approx .33 \times .17 \approx .056$. The denominator depends on other hypotheses that are consistent with the data. The data d_1 are only consistent with two hypotheses, the intended hypothesis h_3, and h_2. Because the prior probability of all hypotheses are the same and the data are all selected at random from three possibilities, the $P(d_1|h_2) \times P(h_2) = P(d_1|h_3) \times P(h_3)$. Putting the pieces together,

$$P(h_3|d_1) = \frac{P(d_1|h_3) \times P(h_3)}{P(d_1|h_3) \times P(h_3) + P(d_1|h_2) \times P(h_2)} \approx \frac{.056}{.056 + .056} = .5.$$

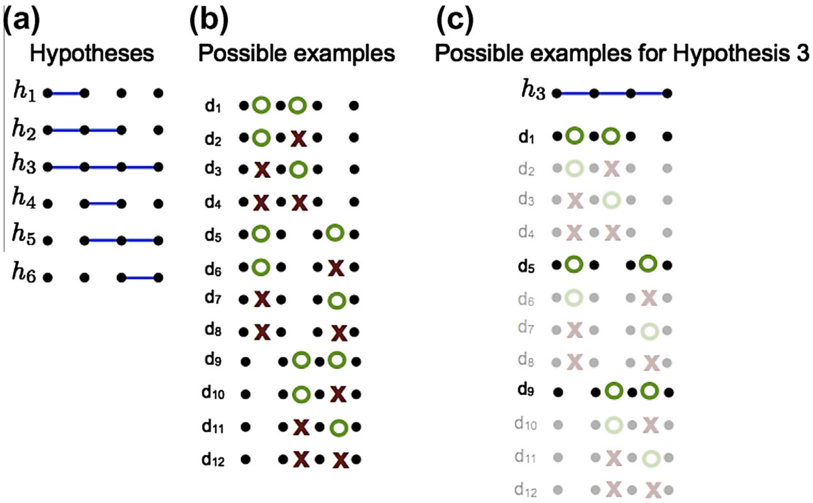
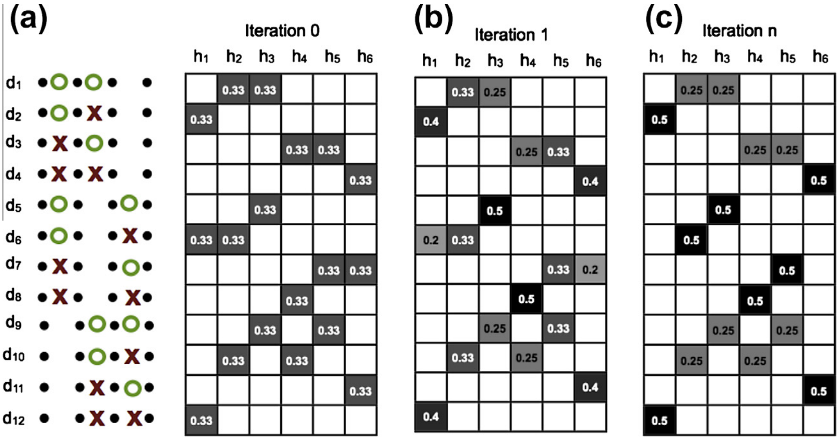


Fig. 16. Illustration of the line game. (a) The possible hypotheses are indicated by blue lines, where areas without the line are outside the concept. (b) The possible data are combinations of two labeled examples, with green circles indicating areas inside the hypothesis and red Xs indicating areas outside the hypothesis. (c) The possible data for a specific hypothesis, h_3. For this hypothesis, only some data are possible: d_1, d_5, and d_9. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



Iteration 0

	h1	h2	h3	h4	h5	h6
d1		0.33	0.33			
d2	0.33					
d3				0.33	0.33	
d4						0.33
d5			0.33			
d6	0.33	0.33				
d7					0.33	0.33
d8				0.33		
d9			0.33		0.33	
d10		0.33		0.33		
d11						0.33
d12	0.33					

Iteration 1

	h1	h2	h3	h4	h5	h6
d1		0.33	0.25			
d2	0.4					
d3				0.25	0.33	
d4						0.4
d5			0.5			
d6	0.2	0.33				
d7					0.33	0.2
d8				0.5		
d9			0.25		0.33	
d10		0.33		0.25		
d11						0.4
d12	0.4					

Iteration n

	h1	h2	h3	h4	h5	h6
d1		0.25	0.25			
d2	0.5					
d3				0.25	0.25	
d4						0.5
d5			0.5			
d6		0.5				
d7					0.5	
d8				0.5		
d9			0.25		0.25	
d10		0.25		0.25		
d11						0.5
d12	0.5					

Fig. 17. Predictions of the pedagogical model for a simple case of rule-based concepts. The matrices represent the probability of choosing the data given the hypothesis, and therefore each column sums to one. Zeros are omitted for clarity. (a) The probability of sampling data given hypotheses assuming random sampling. (b) The probability of sampling data given hypotheses after one step of the pedagogical model. (c) The optimal probability of sampling data for each hypothesis according to the pedagogical model.

Note that different data lead to different conclusions. For instance, a learner who observed d_5 would infer h_3 with probability 1, because there are no other hypotheses that are consistent with d_5.

Now consider which of these data the teacher should choose, $P_{teacher}(d|h)$. Eq. (2) suggests that teachers should choose among the data that are consistent with the hypothesis, favoring those that are more likely to lead the learner to the correct inference. Here, a teacher would choose among d_1, d_5, and d_9. The probability of a correct inference given d_1 and d_9 is .5, while the probability of a correct inference given d_5 is 1. Following Eq. (2),

$$P_{teacher}(d_5|h_3) = \frac{P_{learner}(h_3|d_5)}{P_{learner}(h_3|d_1) + P_{learner}(h_3|d_5) + P_{learner}(h_3|d_9)} = \frac{1}{.5+1+.5} = .5$$

(see Fig. 17b).

If both the teacher and learner are privy to this reasoning strategy, then one iteration is not enough. If the learner knows that the teacher will be helpful, then the teacher would want to reason about what the learner will do, given this information. Thus, the optimal choice of examples is based on the stable solution to the system of equations; essentially this amounts to running the computations described in the previous paragraphs over and over until the answer is (nearly) stable (see Fig. 17c).

This sampling strategy outlines a generative process for choosing data that optimizes the learner's chances to invert the process to guess the correct hypothesis. These results are for a teacher that chooses probabilistically, i.e. for whom $\alpha = 1$, and who assumes that examples. Note that pedagogical sampling predicts two basic effects that can be seen in Fig. 17 and Experiment 1: positive pairs should mark the boundaries (see h_3) and negative examples should be near the boundaries (see h_1).

A.2. Computational details: Experiment 1, Rule-based concepts

As mentioned above, there are four basic pieces required to formalize the model: hypothesis space, priors, possible data, and likelihoods. For the rectangle game, we approximate the full reasoning problem by discretely approximating the game board with a 6 × 6 grid. The hypothesis space is then every rectangle from 2 × 2, through 2 × 5 up to 5 × 5, a total of 196 rectangles. We assume a uniform priors over the possible hypotheses, i.e. $P(h) = \frac{1}{196}$ for all rectangles. We consider pairs of labeled examples. The set of possible single data points include each of the 36 locations on the 6 × 6 grid. This gives $36 \times 35 = 1260$ pairs and $2^2 \times 1260 = 5040$ possible labeled pairs. The likelihoods are based on random sampling from among the labeled pairs that are consistent with a given hypothesis. As in the example above, each hypothesis is consistent with the same number of labeled pairs.

Although there are considerably more hypotheses and data, the model predictions are qualitatively similar to those in the example above. Specifically, focusing on positive examples, because examples in opposite corners of the rectangle rule out the largest number of alternative hypotheses, these are the examples that the model predicts teachers will choose. Similarly, turning to negative examples, because examples that are just outside the rectangle eliminate the most possible alternatives, these are the examples that the model predicts teachers will choose. Because learners reason about teachers' behavior, the model predicts that learners should infer rectangles in which positive examples are in the corners and negative examples are on the boundaries. To generate the predictions for the teaching task, we averaged the predictions based on the four approximately centered 3 × 3 rectangles. Because learners reason about the teachers choices, the model predicts the same qualitative effects for learning.

A.3. Computational details: Experiment 2, Prototype concepts

As for the other experiments, to model Experiment 2, we must specify the possible hypotheses, the priors, the possible data, and the likelihoods. Recall that in the teaching task, participants chose three examples from among 27 line segments whose lengths were sampled from a normal distribution. In the learning task, participants were shown three examples and they rated a series of 9 examples varying in length in uniform increments, on whether they were from the same category. To capture this in the model, we start with a collection of 9 uniformly-spaced examples. The set of possible data includes all possible triplets of 9 (where order does not matter), 165 examples. We derive the possible hypotheses from the possible examples. The possible means were set to correspond to all possible means of three stimuli. The standard deviations were set to correspond to the one-half of the standard deviations of all possible sets of three examples. (Standard deviations of 0 were set to .01.) Thus, there were a total of 165 mean and standard deviation pairs. The prior probabilities, $P(h)$, are assumed to be uniform, $\frac{1}{165}$. The likelihood is a discretized Gaussian, where the probabilities are assigned for each of the 9 segments based on the particular mean and standard deviation and renormalized to ensure that they sum to 1. Probabilities for triplets are computed by assuming examples are selected independently.

To generate predictions for the teaching task, the examples were binned with the smallest 2 segments called small, the largest 2 segments called large, and the remaining 5 called medium. The probabilities for a given triplet (e.g. SSS) by summing the probabilities of all examples that fell in that category. We generated predictions based on examples generated from the true distribution by either strong sampling (random condition) or pedagogical sampling (teaching condition). Participants' choices were contrasted with the predictions of the two models in Fig. 7.

The learning task included manipulations of the examples provided (*Teacher* or *Random*), as well as the learners' beliefs about the examples (*Teacher* or *Random*). Thus, to model the learning task, we must model the choice of examples and the learner's beliefs about the choice of examples. We then must generate the learner's ratings, given the data they observe and the inferred probability of hypotheses. To model choice of data and inferences, posterior beliefs were computed based on the learner's assumption that the data were strong sampled (random cover story) or pedagogically sampled (teaching cover story) by marginalizing over the data that were chosen according to the true sampling process. Predictions for the 9 data points were then generated using the posterior predictive distribution—by computing the probability of randomly choosing that data point from each hypothesis, weighted by the probability of that hypothesis given the observed data. To reflect uncertainty in people's perception of the line lengths, predictions were generated based on uncertain estimates of the true length. Uncertainty was implemented as a normal distribution centered on the example with a standard deviation 1.5 times the difference between examples. The perceptual noise does not affect the qualitative ordering of the model predictions, but does increase the variance of predictions across all four conditions. For Fig. 8, the probabilities were then scaled such that the largest value was one (by dividing all values by the maximum). The variances in Fig. 9 were also derived from the posterior predictive distribution.

A.4. Computational details: Experiment 3, Causally-structured concepts

For the causal experiment, the hypothesis space includes all possible acyclic graphs over three variables, giving 12 hypotheses (3 common cause, 3 common effect, and 6 chains). The prior probabilities, $P(h)$, are uniform, $\frac{1}{12}$. The set of possible data include all possible pairs of interventions on three variables (including duplicates), $3 \times 3 = 9$. The likelihood is assumed to be uniform over possible individual interventions, $\frac{1}{3}$, and pairs are sampled independently, $\frac{1}{3} \times \frac{1}{3}$. Because we are considering causal relationships, we must additionally specify how variables affect each other. Based on the training, the relationships were noisy-or and the strength of the causal relationships were set to .9 and the base rate was set to .05.

Predictions for Fig. 12 were generated by running the model and plotting the likelihoods. Predictions for Fig. 13 and 14 were generated by running the model and plotting the posterior probabilities of the depicted hypotheses. Where there is more than one hypothesis represented by a single bar, the posterior probabilities were averaged.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409–429.

Ashby, F. G., & Alphonso-Reese, L. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39*, 216–233.

Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., & Lerner, S. (1997). Teaching by examples: Implications for the process of category acquisition. *Quarterly Journal of Experimental Psychology, 50A*, 586–606.

Baldwin, D. (1993). Early referential understanding: Infants' ability to recognize acts for what they are. *Developmental Psychology, 29*, 832–843.

Benz, A., Jager, G., & van Rooij, R. (2005). *Game theory and pragmatics*. Hampshire, UK: Palgrave Macmillan.

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

Brennan, S., & Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1482–1493.

Bruner, J. R., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Chase, C., Chin, D. B., Oppezzo, M., & Schwartz, D. L. (2009). Teachable agents and the protege effect. *Journal of Science Education and Technology, 18*, 334–352.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science, 3*, 57–65.

Colman, A. M. (2003). Depth of strategic reasoning in games. *Trends in Cognitive Sciences, 7*, 2–4.

Csibra, G. (2007). Teachers in the wild. *Trends in Cognitive Sciences, 11*, 95–96.

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and cognitive development. Attention and performance XXI* (pp. 249–274). Oxford: Oxford University Press.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 14*, 148–153.

Dean, D., & Kuhn, D. (2006). Direct instruction versus discovery: The long view. *Science Education, 91*, 384–397.

Elio, R., & Anderson, J. R. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 7*, 397–417.

Elio, R., & Anderson, J. R. (1984). The effects of information order and learning mode on schema abstraction. *Memory and Cognition, 12*, 20–30.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*, 998.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 234–257.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., et al (2012). Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological Science, 23*, 931–939.

Fussell, S., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology, 25*, 203–219.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition, 27*, 181–218.

Gergely, G., Bekkering, H., & Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature, 415*, 755.

Goldstone, R. L., & Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology, 46*, 414–466.

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the Learning Sciences, 14*, 69–110.

Goodman, N. D., Baker, C. L., & Tenenbaum, J. B. (2009). Cause and intent: Social reasoning in causal learning. In *Proceedings of the 31st annual meeting of the cognitive science society*.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*, 108–154.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 1–31.

Gosselin, F., & Schyns, P. G. (2001). Why do we slip to the basic-level? Computational constraints and their implementation. *Psychological Review, 108*, 735–758.

Griffiths, T. L., & Kalish, M. L. (2007). A Bayesian view of language evolution by iterated learning. *Cognitive Science, 31*, 441–480.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric bayesian density estimation. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for rational models of cognition*. Oxford: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 354–384.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*.

Gureckis, T. M., & Goldstone, R. L. (2006). Thinking in groups. *Pragmatics and Cognition, 14*, 293–311.

Hedden, T., & Zhang, J. (2002). What do you think I think you think? Strategic reasoning in matrix games. *Cognition, 85*, 1–36.

Hsu, A., & Griffiths, T. L. (2009). Differential use of implicit negative evidence in generative and discriminative language learning. In *Advances in neural information processing systems 22*.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge University Press.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*, 75–86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 7*(4), 661–667.

Krauss, R. M. (1987). The role of the listener: Addressee influences on message formulation. *Journal of Language and Social Psychology, 6*, 81–97.

Luce, R. D. (1959). *Individual choice behavior*. New York: John Wiley.

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences, 104*, 19751–19756.

Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General, 127*, 331–354.

Marr, D. (1982). *Vision*. New York: W.H. Freeman.

Mathy, F., & Feldman, J. (2009). A rule-based presentation order facilitates category learning. *Psychonomic Bulletin and Review, 16*, 1050–1057.

Mayer, R. E. (2004). Should there be a three-strikes rule against discovery learning. *American Psychologist, 59*(1), 14–19.

Medin, D. L., & Bettger, J. G. (1994). Presentation order and recognition of categorically related examples. *Psychonomic Bulletin and Review, 1*, 250–254.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 100*, 254–278.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198*, 75–78.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Nosofsky, R. M. (1991). Relation between the rational model and the context model of categorization. *Psychological Science, 2*, 416–421.

Nosofsky, R. M., Gluck, M., Palemeri, T. J., & McKinley, S. C. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition, 18*, 211–233.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Piaget, J. (1955). *The construction of reality in the child*. New York: Basic Books.

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*, 1–57.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 241–248.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26*, 303–343.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382–407.

Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science, 27*, 709–748.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 130*, 323–360.

Rhodes, M., Gelman, S. A., & Brickman, D. (2010). Children's attention to sample composition in learning, teaching, and discovery. *Developmental Science, 13*(3), 421–429.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-exploration and direct instruction. *Child Development, 77*, 1–15.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition. Bradford books*. Cambridge: MIT Press.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117*, 1144–1167.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science, 7*, 341–351.

Shafto, P., Kemp, C., Baraff, E., Coley, J. D., & Tenenbaum, J. B. (2005). Context-sensitive reasoning. In *Proceedings of the 27th annual conference of the cognitive science society*.

Shafto, P., Kemp, C., Baraff, E., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition, 108*, 175–192.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, 75.

Spirtes, P., Glymour, C., & Schienes, R. (1993). *Causation, prediction, and search*. New York: Sperner-Verlag.

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 3–11.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. Kearns, S. A. Soller, T. K. Leen, & K. R. Müller (Eds.), *Advances in neural processing systems 11* (pp. 59–65). MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001a). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*, 629–641.

Tenenbaum, J. B., & Griffiths, T. L. (2001b). The rational basis of representativeness. In *Proceedings of the 23nd annual conference of the cognitive science society* (pp. 1036–1041). Hillsdale, NJ: Erlbaum.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*, 309–318.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Topal, J., Gergely, G., Miklosi, A., Erdohegyi, A., & Csibra, G. (2008). Infants' perseverative search errors are induced by pragmatic misinterpretation. *Science, 321*, 1831–1834.

Voiklis, J., & Corter, J. (2012). Conventional wisdom: Negotiating conventions of reference enhances category discovery. *Cognitive Science, 36*, 607–634.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the aquisition of category structure. *Journal of Experimental Psychology: General, 124*, 181–206.

Wilkes-Gibbs, D., & Clark, H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language, 31*, 183–194.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, 97–104.

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

Xu, F., & Tenenbaum, J. B. (2007b). Word learning as Bayesian inference. *Psychological Review, 114*, 245–272.