When teaching breaks down: Teachers rationally select what information to share, but misrepresent learners' hypothesis spaces

Rosie Aboody[1] (rosie.aboody@yale.edu), Joey Velez-Ginorio[2] (joeyvelez@knights.ucf.edu), Laurie R. Santos[1] (laurie.santos@yale.edu), & Julian Jara-Ettinger[1,3] (julian.jara-ettinger@yale.edu)

[1] Department of Psychology, Yale University. New Haven, CT 06520 USA.
[2] Department of Electrical Engineering, UCF. Orlando, FL 32816 USA.
[3] Department of Computer Science, Yale University. New Haven, CT 06520 USA.

Abstract

Although we possess intuitions about pedagogy from early in life, adults commonly fail to teach effectively in real-world situations. Why might adults struggle in more complex teaching tasks? Here we develop a simple teaching task where adults fail to teach naïve learners, despite reporting high confidence that they taught effectively. Using a formal model of a rational teacher, we analyze the sources of our adult teachers' failures. Our model-based analyses reveal that teachers successfully provided high-quality examples, but failed to address hypotheses that naïve learners find plausible. We validate these results in a second experiment, where we find that constraining learners' hypothesis space increases their performance in the task. Our findings help bridge the gap between children's teaching proficiency in constrained tasks, and adults' teaching failures in more naturalistic tasks.

Keywords: social cognition; theory of mind; pedagogy; computational modeling.

Introduction

Our propensity to share what we know with others allows our species to compile extensive bodies of knowledge over time, ameliorating the need for each generation to acquire information firsthand. But despite the ubiquity of our species' pedagogical interactions, the act of sharing what we know is far from straightforward. Explaining too much is tedious and inefficient; explaining too little is ineffective. To teach well, we must decide what and how much to share. And to avoid providing the wrong amount of information, we must take into account what learners already know. But in choosing what to share, we face an epistemic problem: we cannot see other people's knowledge. Thus, to teach well, we must infer what others know, and rely on these inferences to decide what and how much information to share.

Impressively, children seem to solve this challenging problem from early on. Young children can reason about others' beliefs even when those beliefs are incorrect (Wellman, Cross & Watson, 2001), and they are sensitive to features of teaching that make for a successful pedagogical interaction (i.e., whether teachers provide data that matches their learners' needs; Gweon & Asaba, 2017; Gweon, Shafto & Schulz, 2014). Toddlers will point to share information with ignorant (but not knowledgeable) adults (Liszkowski, Carpenter & Tomasello, 2008), and by the late preschool years, children can identify and address learners' errors during a simple teaching task (Ronfard & Corriveau, 2016; Strauss, Ziv & Stein, 2002). Children also gauge the utility of different pieces of information in the context of learners' goals and costs, and choose to share the utility-maximizing information (Bridgers, Jara-Ettinger, & Gweon, 2016).

Puzzlingly, adults do not always teach as proficiently as one might expect, given children's early pedagogical successes. Most of us can probably recall failing to get a point across despite our best intentions, and many studies have documented that adults often fail to teach effectively in unconstrained interactions (Chi, Siler, Jeong, Yamauchi & Hausmann 2001; Chi, Siler & Jeong, 2004; see VanLehn, 2011 for review).

If children already possess intuitions about successful teaching during the first few years of life, why does adult teaching so often break down in naturalistic tasks? Tasks where adults fail are often more complex than those where children and adults succeed. As such, there are many more pieces of information a teacher could choose to provide. Perhaps when teachers are confronted with so many options, they struggle to decide what information to share. Under this account, teachers may fail to teach effectively because they cannot distinguish helpful from unhelpful information. Alternatively, teachers may struggle to represent the specific hypotheses their learners consider plausible, especially in complex domains. Under this account, teachers succeed in simple tasks because they can represent learners' hypothesis spaces easily, but fail in more complex tasks because they do not grasp the additional hypotheses learners consider.

Naturally, both types of difficulties likely affect performance in complex tasks. Unfortunately, identifying the role that these limitations play has historically been challenging. Naturalistic tasks that elicit adult teaching failures are too complex to analyze formally. For example, Chi et al. (2004) asked college students to tutor 8th graders about the human circulatory system. These tutoring sessions lasted between 1.5 – 2 hours, with dialogue and questions encouraged. Despite the ability to query their learners and correct misunderstandings, Chi et al. (2004) found that after the tutoring session concluded, tutors tended to overestimate how much their students knew. And over the course of the tutoring session, tutors sometimes failed to detect, diagnose, and correct misconceptions. Because these tutoring sessions

were so long and unconstrained, it is difficult to identify the sources of teachers' failures.

Conversely, tasks that are susceptible to analysis are generally simpler, and teachers tend to succeed. Therefore, they do not reveal the sources of teachers' difficulties. For example, Shafto, Goodman and Griffiths (2014) asked teachers to indicate the location and size of a rectangle by placing two markers on a screen. Learners would see only the markers, and would have to guess the location and size of the rectangle. Teachers in this task perform incredibly well, almost always placing markers in the two opposing corners of the rectangle (enabling learners to infer both relevant dimensions: size and location). In sum, tasks where teachers succeed tend to be simple and constrained. Tasks where teachers fail are more complex, and this complexity makes it difficult to identify the sources of teachers' failures.

In this paper, we bridge this divide by developing a teaching task complex enough to elicit teaching failures, but also simple enough to identify the causes of teachers' difficulties. Relying on a standard model of pedagogy that explains how teachers and learners communicate successfully, we show that participants' failures to teach cannot be explained by the quality of the examples they provide. Instead, our formal analyses suggest that teachers fail because learners' hypothesis spaces are substantially larger than teachers assume. We validate our model conclusions through a second experiment, where we find that constraining learners' hypothesis space increases their quantitative performance on their learning task. Altogether, our results are consistent with literature suggesting that pedagogy is impaired by a "curse of knowledge" (wherein participants' ability to reason about naïve minds is impaired by their own privileged knowledge; Camerer, Loewenstein & Weber, 1989; Hinds, 1999; Nickerson, 1999). Our findings begin to shed light on how this curse arises.

Computational framework

Our computational framework is based on previous research investigating how people share information (Shafto et al., 2014). Teachers can be formalized as generating data, given their knowledge of the correct hypothesis, and learners can be formalized as inferring the correct hypothesis, given the data that they receive. The process of teachers tailoring their data to learners, and learners reasoning about why the teacher provided that particular data, can be formalized through a pair of recursive equations:

$$p_{teacher}(D|H) \propto p_{learner}(H|D) \quad (1)$$

and

$$p_{learner}(H|D) \propto p_{teacher}(D|H), \quad (2)$$

where $\boldsymbol{p_{teacher}(D|H)}$ is the probability that the teacher will generate certain data, *D*, given the true hypothesis *H*, and $\boldsymbol{p_{learner}(H|D)}$ is the probability that the learner will infer the correct hypothesis given the data that they observe. These equations formalize the idea that rational teachers select data that will allow rational learners to infer the right hypothesis, and that rational learners infer this hypothesis by reasoning about why the teacher chose the data they did.

The learner's success in recovering the right hypothesis (Eq. 2) depends on two factors: the teacher's data, and the learner's hypothesis space. Here, our main interest is in using the model to evaluate teachers' data. Thus, $\boldsymbol{p_{teacher}(D|H)}$ is obtained from Study 1, and this data is evaluated using Equation 2. To gauge the quality of the data, we designed a set of hypothesis spaces that sequentially increase in size, by combining basic primitive hypotheses using two logical operators: *AND* (&), and *OR* (|). In our study, primitive hypotheses correspond to beliefs that a specific block must be on top of a machine for it to activate (see Study 1 methods for details). In the simplest hypothesis space, hypotheses consist of up to two primitive hypotheses combined by one logical operator (e.g. E; A&C; B|D).[1] These hypotheses correspond to simple beliefs participants might hold about the machine (e.g., that block E makes the machine go; that blocks A and C together are required to make the machine go; that either block B or block D is required to make the machine go).

This hypothesis space can be expanded by increasing the number of primitive hypotheses that can be combined (called the *ceiling*), and it can be modified depending on whether primitive hypotheses can only be combined by a single logical primitive (called *single-primitive* space; e.g. A&B&C; A|B|C), or by more than one primitive (called *dual-primitive* space; *e*.g., A&(B|C); (A&B)|C).

Study 1

Study 1 consisted of a teaching task and a learning task (run across participants). Participants in the teaching task learned how to activate a machine, and were then asked to generate examples that would show a naïve participant how the machine works. Participants in the learning task saw a set of these examples, and then were asked to infer how the machine works. Teachers' performance was assessed based on the proportion of naïve participants who uncovered how the machine worked. The sources of any failures were analyzed by feeding teacher-selected data to our model.

Methods

Participants 220 participants were recruited from Amazon's Mechanical Turk platform. The first 20 participants (mean age = 35.4; range = 21-70) were assigned to the teacher condition, and the last 200 participants (mean age = 34.4; range = 19-68) were assigned to the learner condition. Five additional participants were recruited but not included in the study because they failed an inclusion question (teacher n = 2; learner n = 2) or because they did not follow task instructions (n = 1).

Stimuli Stimuli consisted of images of a machine with a triangle on the front, and of 5 blocks. The color of the

[1] We do not consider the hypothesis space that consists only of hypotheses with no logical primitives (A, B, C, D, E), because it does not contain the true hypothesis (B&E) used in our task.

triangle signaled whether the machine was on or off (see Figure 1). When two particular blocks (B and E) were placed on top of the machine together, it activated (henceforth referred to as the "B&E" rule). The presence or absence of other blocks did not affect the outcome.

Procedure ***Teachers.*** Participants assigned to the teacher condition were told how the machine worked, and asked to generate between 3 and 20 unique examples that would teach a naïve learner the B&E rule. After generating their examples, teachers were asked to rate their confidence (on a Likert scale) that a naïve learner would learn the B&E rule from their examples. Critically, teachers were explicitly told that learners would know nothing about how the machine worked, but that their understanding of the machine would be tested after they saw teachers' examples.

Learners. Participants assigned to the learner condition were first familiarized with the machine, and taught how to distinguish between an inactive and an active machine. Unlike participants in the teacher condition, however, learners were *not* told how the machine works. Instead, learners were shown the examples that one of the teachers generated (with ten learners assigned to each of the twenty teachers) and they were asked to infer the underlying activation rule. Understanding was assessed in two ways: In the quantitative task, participants were shown every possible combination of block(s) on the machine, and indicated whether each combination would activate the machine or not (31 possible combinations). In the qualitative task, participants were asked to explain how the machine worked.

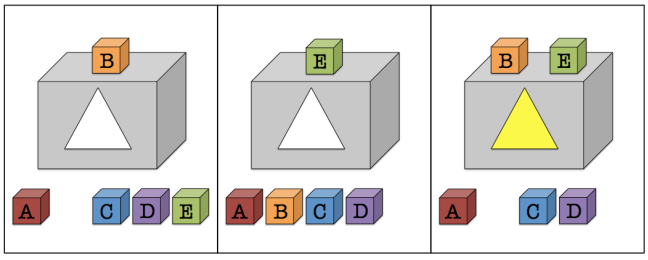

Figure 1: Stimuli used in the study. The far right panel shows the activation rule teachers had to communicate (blocks B and E together made the machine turn on).

Results

Although participants in the teacher condition only had to produce a minimum of three examples, they produced an average of 8.2 examples (range = 3-20; SD = 4.4), with only two participants providing the minimum three unique examples. On the post-task confidence rating, teachers were confident that a naïve participant would successfully learn from their examples, with a mean confidence rating of 6.05 on a 7-point scale (range = 5-7; SD = 0.76).

Despite teachers' confidence, not all participants in the learner condition succeeded in learning how the machine worked. Only 50% (n = 100) of participants performed at or near ceiling in the quantitative task, correctly identifying which block combinations activated the machine with one mistake or less (n = 88 performing perfectly). Nonetheless, the remaining 50% of participants showed evidence that they had partially learned how the machine worked. On average, these participants predicted whether the machine would be on or off correctly in 70.6% of trials, performing significantly above chance (n = 100; $t(99) = 17.58$; $p < .001$). Although it is possible that learner performance reflects only differences in motivation, a Monte-Carlo permutation test revealed this is not the case. Learners' performance is significantly predicted by the teacher they were assigned to learn from ($p = 0.03$, 10,000 samples).

To uncover the sources of learners' difficulties, we next examined their qualitative explanations. 52% (n = 104) of learners correctly described how to activate the machine (answers independently coded by the first and second authors; Cohen's $\kappa = 0.85$; $p < .001$). These learners were largely the ones who performed well on the quantitative task (86 performed perfectly, 6 made one error). Nine learners gave uninformative explanations (e.g., "When the machine is on, the triangle turns to the yellow color") and were excluded from analyses.

The remaining 43.5% of explanations most naturally fell into one of two categories: either referencing the *right kinds* of hypotheses, or the *wrong kinds* of hypotheses. Participants who provided the right kinds of explanations understood that the machine was activated by placing a certain combination of blocks on top, but didn't quite figure out which ones (n = 42). Participants who provided the wrong kinds of explanations did not identify the relevant features of the task, producing explanations that referenced incorrect activation mechanisms (for example, believing that the distance between the blocks determined whether it would be activated or not, or that the blocks needed to form an English word for the machine to go; n = 45).

Model-based Analysis

To formally evaluate whether learners erred because teachers failed to infer their complete hypothesis space, we next analyzed teachers' examples through our computational model. Given teachers' examples, our model computes what learners should conclude from these examples, as a function of the hypotheses they are considering. If learners struggled in our task because teachers provided bad or confusing data, then our model should be unable to infer how the machine works, even under constrained hypothesis spaces. However, if learners struggled in our task because teachers provided helpful data that simply did not target their beliefs, then our model should find this data to be sufficient to infer how the machine works under constrained hypothesis spaces.

We first evaluated the data using the simplest hypothesis space: a single-primitive hypothesis space with ceiling = 2 (see Computational Framework for explanation of the ceiling and single/dual primitive parameters). Hypotheses in this space could be single blocks (e.g., B), and two blocks, combined with an *AND* or an *OR* operator, (e.g., B&E, or B|E; 30 hypotheses total). The model inferred the correct rule for 75% of teachers (n = 15), placing over 95% of the

posterior probability mass on the correct hypothesis (these results are identical when the mass threshold is decreased to 50%). For the remaining 5 participants, the model continued to place the highest posterior probability on the correct hypothesis (B&E; on average 12%), but similar hypotheses were rated as equally plausible, preventing the right hypothesis from accruing a probability mass above 50%."

Next, the hypothesis space was sequentially increased to identify the first space where the model usually failed to infer the underlying activation rule. The model continued to succeed for the same 75% of teachers in all single-primitive hypothesis spaces, for all ceiling values. By contrast, in the dual-primitive hypothesis spaces, the model concluded that participants should fail to learn the activation rule with a ceiling as low as 2. This hypothesis space contained hypotheses built from up to two primitive units, and combined with both types of logical primitives (e.g., (B*&*(E|C)); 120 hypotheses total). For this hypothesis space, the model inferred the activation rule for 25% of teachers (n = 5). The examples from the remaining teachers did not sufficiently narrow down the posterior hypothesis space. While no other hypotheses were ever rated as more likely than the B&E rule, there continued to be too many hypotheses left that were consistent with teachers' data.

Our results demonstrate that teachers provided examples informative enough for a rational learner model to infer the correct activation rule under constrained hypothesis spaces. This model assumed that examples were chosen by a teacher attempting to maximize the learner's belief in the right hypothesis (Equation 2). Past work has established that, when learning from minimal data, this pedagogical assumption is critical (Shafto et al., 2014). To evaluate if this assumption was also critical in our analyses, we reanalyzed the data using an impaired model that did not treat the data pedagogically. In contrast to our main model, the impaired non-pedagogical model computed the posterior probability of the hypothesis space based on whether each hypothesis was consistent with the observed data (without any assumptions about how this data was selected). Intuitively, this corresponds to an assumption that the examples were randomly generated, rather than selected by a knowledgeable teacher.

At the 50% probability mass threshold, the impaired model produced identical results, but it was less successful at the 95% probability mass threshold. For all single-primitive hypothesis spaces, this model only placed over 95% of the posterior probability mass on the correct hypothesis for 65% of teachers (n = 13). And for the dual-primitive hypothesis space with a ceiling of 2, the impaired model only placed over 95% of the posterior probability mass on the correct hypothesis for 10% of teachers (n = 2). Additionally, the impaired model "learned" more slowly than the full pedagogical model. To succeed in placing over 95% of the posterior probability mass on the true hypothesis, this model needed to observe 1 more example than the pedagogical model on average.

In sum, given single-primitive hypothesis spaces, both models successfully learned the correct activation rule from the majority of teachers' examples (with the impaired model performing more weakly at the 95% threshold). And given a larger, more complex dual-primitive hypothesis space, both models failed (even at the lowest possible ceiling of 2). That is, under some circumstances, both models found most teachers' data to be sufficient. The fact that teachers' examples were informative in the single-primitive hypothesis spaces (even at a conservative 95% probability mass threshold) suggests that teachers did not generate poor examples. Instead, this suggests that participants may have failed to teach well because they did not consider the entire space of learner hypotheses.

Discussion

Although teachers were confident they provided good data, many learners struggled to infer the activation rule. While these failures could arise from a lack of teacher motivation, teachers generally provided more examples than they needed to. Furthermore, our model-based analyses suggest that teachers' data was informative, but only useful under narrow hypothesis spaces: because teachers failed to represent the breadth of learners' beliefs, they failed to teach effectively. If this is the case, then learners whose beliefs are constrained to match teachers' representations should be able to learn more effectively from this data.

Study 2

Learners were presented with the examples obtained from the teacher condition in Study 1. First, however, learners were shown the activation rules of similar machines, thus constraining their hypothesis space. By demonstrating that similar machines are always activated by one or more blocks, learners should now only consider hypotheses that consist of block combinations - independent of irrelevant features such as their spatial arrangement. Consequently, we predict that learner performance will improve, even given the same examples as learners in Study 1.

Methods

Participants 200 participants were recruited from Amazon Mechanical Turk (mean age = 34.4; range = 18 - 66).

Stimuli We created images of two additional light-up machines (of a different size and color), each with a corresponding set of blocks).

Procedure First, we constrained learners' beliefs by introducing them to two other light-up machines. One machine was introduced with three blocks (one of which made it go), and the other with six blocks (three of which were needed to make it go). After learning the activation rules of these two machines, participants were introduced to the target machine from Study 1. The experiment then proceeded identically to the learner condition in Study 1.

Results

Quantitatively, learners in Study 2 performed significantly better than learners in Study 1 (mean = 92%, SD = 16; $p < .001$, permutation test with 10,000 samples; Figure 2). Average learner scores (calculated by teacher) improved the most drastically for teachers whose learners had produced the most wrong kind explanations in Study 1: the difference between average learner performance in Study 1 and Study 2 is marginally predicted by the number of wrong kind explanations learners gave in Study 1 ($\beta = 0.7724$; $p = .064$).

As in Study 1, the first and second authors independently coded qualitative explanations (Cohen's $\kappa = 0.94$; $p < .001$). Four participants provided uninformative explanations (e.g., "The machine turns on by the yellow triangle") and were therefore excluded from qualitative analyses. As predicted, the number of correct qualitative explanations was significantly higher than in Study 1 (75% of participants giving the right activation rule; Fisher's exact test, $p < .001$). Furthermore, while participants produced only 3 fewer right kind explanations in contrast to Study 1 (n = 39), they produced 38 fewer wrong kind explanations (n = 7).

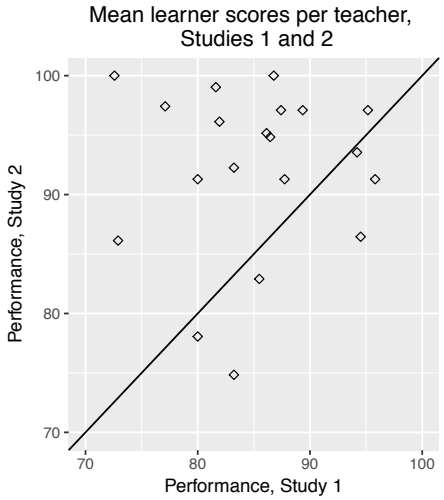

Figure 2: Learner performance across studies. Each point represents average learner performance in Study 1 (x-axis) and Study 2 (y-axis) for a given teacher's examples. Points above the diagonal indicate that learners with a constrained hypothesis space (Study 2) performed better than learners with an unconstrained hypothesis space (Study 1).

Discussion

Although learners in Study 2 saw the same examples as learners in Study 1, participants performed better on both the qualitative and quantitative measures. Moreover, performance improved the most for teachers whose examples had produced the most wrong kinds of explanations in Study 1. We investigated the mechanism behind this improvement, and found that learners in Study 2 provided significantly fewer explanations that focused on the wrong kinds of hypotheses. These results indicate that we succeeded in constraining learners' hypothesis spaces (because they were no longer considering the wrong kinds of hypotheses). These results also dovetail with those of our model-based analysis, providing strong evidence that teachers in Study 1 indeed chose helpful data, but misjudged learners' initial hypothesis spaces.

General Discussion & Conclusion

Despite the ubiquity of teaching interactions, we often struggle to share information effectively (e.g., Chi et al., 2004). Two types of teaching deficits could explain these suboptimal outcomes: perhaps teachers cannot decide what information will be helpful to share when they have many options. Or, perhaps teachers inaccurately represent the possibilities that learners are considering, and therefore provide good data that isn't matched to learners' beliefs.

To distinguish between these two accounts, participants in our first study were assigned to a teaching or learning task. Although teachers in this task were confident they had taught well, naïve participants struggled to learn from their data. A formal analysis indicated that teachers chose informative data, but misrepresented the breadth of learners' potential beliefs (and thus failed to address possibilities learners found likely). In Study 2, we validated these results with an additional behavioral experiment, finding that when learners' hypothesis spaces are constrained, all teachers' examples become quite effective. Although these participants saw the exact same data as learners in Study 1, they performed significantly better on both our qualitative and quantitative measures. These findings indicate that, broadly, teachers did not struggle to choose informative data. Rather, they failed to infer the breadth of hypotheses learners were considering, and thus did not produce enough data to provide evidence against many of these possibilities.

Although these findings are an exciting step towards better understanding pedagogy (and its boundaries), there are several limitations. First, in our formal analysis, our modeled 'learners' began by considering a particular hypothesis space, which became constrained as they observed teachers' examples. But intuitively, real learners probably don't start off with a large array of hypotheses they are considering. Rather, learners likely come up with a space of possibilities over time. Although this is a point where our model is likely incongruent with real learners' reasoning, our model is not intended to capture the algorithms people use to build hypothesis spaces. It instead provides a computational-level analysis (Marr, 1982) of how teacher-provided data constrains hypothesis spaces of different sizes. Future work will investigate how learners generate hypotheses as a function of the examples they see.

Second, the experiments reported in this paper were conducted on Amazon's Mechanical Turk platform. Although teaching failures are often found in lab-based samples (e.g. Chi et al., 2004), it is possible that participants online performed poorly because they were unmotivated, or due to task-based constraints on their ability to generate examples for learners. Overall, it appears unlikely that teachers were unmotivated: most teachers provided more examples than required (even those the model failed to learn

from). And even learners who struggled usually performed far above chance. However, it is possible that the online format prevented teachers from producing certain types of examples, and thus teaching to their fullest capacity. Future work will address both of these possibilities by replicating Study 1 with an in-lab sample of teachers (with a real machine and blocks that they can use for demonstrations).

Our findings also introduce some important future directions. For example, do teachers struggle to generate the types of hypotheses learners might be considering? Or do they generate an appropriate range of possibilities, but fail to evaluate which ones learners still find likely? Future work will address this question by providing teachers with different hypotheses a learner might consider on this task. Teachers will only have to *rate* the probability that a learner might consider each hypothesis. These judgments will be contrasted to those obtained from actual learners.

It is also unclear exactly why teachers might have an overly narrow representation of learner's hypothesis spaces. While our findings are certainly consistent with proposals that the curse of knowledge may affect our pedagogical abilities (Hinds, 1999; Camerer, 1992), there have been few formal investigations into how the curse of knowledge arises. How do agents with privileged information decide what the hypothesis space of a naïve agent looks like? One method might be to recall the hypotheses *they* themselves considered when they had been naïve. If this is the case, it would suggest that the more recently a participant has become knowledgeable, the better they will teach, because they will have better access to the range of hypotheses they were considering (e.g., Hinds, 1999). A different method might be to simply independently generate sets of plausible (but incorrect) alternatives to the truth. Future work should distinguish between these possibilities. By further clarifying the mechanism underlying teachers' difficulties, this work could also help us understand how to ameliorate them.

We began by noting an apparent inconsistency in the pedagogical literature: teachers appear to excel in constrained teaching tasks (e.g. Shafto et al., 2014), but fail in more naturalistic tasks (e.g. Chi et al., 2004). Across two studies, we find that when knowledgeable adults teach, they often fail to consider the breadth of hypotheses a naïve learner may be considering. Although these teachers provide excellent data, they fail to provide enough of it for some naïve participants to learn from. Our results unify prior findings, suggesting that teachers should succeed in tasks where the kinds of hypotheses learners can consider are relatively constrained – but fail in more naturalistic tasks, where learners are considering many possibilities. Because most real-world teaching occurs in naturalistic, unconstrained settings, these findings suggest that teachers would benefit from putting more effort into gauging learners' beliefs – or constraining them.

Acknowledgments

This work was supported by a Google Faculty Research Award to JJE.

References

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2016). Children consider others' expected costs and rewards when deciding what to teach. In *Proceedings of the 38th Cognitive Science Society meeting*.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*.

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*.

Chi, M. T., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction*.

Gweon, H., & Asaba, M. (2017). Order Matters: Children's Evaluation of Underinformative Teachers Depends on Context. *Child Development*.

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

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *JEP: Applied*.

Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. Cambridge, MA: MIT Press.

Nickerson, R.S. (1999) How we know – and sometimes misjudge – what others know: imputing one's own knowledge to others. *Psych. Bulletin.*

Ronfard, S., & Corriveau, K. H. (2016). Teaching and preschoolers' ability to infer knowledge from mistakes. *Journal of Experimental Child Psychology*.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*.

Strauss, S., Ziv, M., & Stein, A. (2002). Teaching as a natural cognition and its relations to preschoolers' developing theory of mind. *Cognitive Development*.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development*.