# Teaching Versus Active Learning: A Computational Analysis of Conditions that Affect Learning

**Scott Cheng-Hsin Yang (scott.cheng.hsin.yang@gmail.com) & Patrick Shafto (patrick.shafto@gmail.com)**
Department of Mathematics & Computer Science, Rutgers University–Newark

**Abstract**

Researchers have debated whether instructional-based teaching or exploration-based active learning is better for decades with unsatisfying results. A main obstacle is the difficulty in precisely controlling and characterizing the pedagogical methods used and the learning conditions in empirical studies. To address this, we leveraged existing computational models of teaching and active learning to formalize the methods and the learning process. We compared the two pedagogical methods in a concept-learning framework and investigated their effectiveness under various scenarios. Our results show that when the learner and teacher are conceptually aligned, teaching is at least as effective as, and often much more effective than active learning, but when the alignment is broken, active learning can yield moderate improvement over teaching. We conclude by discussing our results' implications for the debate and the prospects of bringing computational models to bear on complex real-world problems that are resistant to simple experimental investigation.

**Keywords:** pedagogical methods; direct instruction; self exploration; Bayesian teaching; active learning

For centuries, the predominant pedagogical method has been instructional-based teaching such as lecturing. Over the past several decades, constructivism has received increased attention, and educational practitioners have been incorporating constructivist pedagogical methods that emphasize active learning, exploration, and discovery by the learners themselves (Bruner, 1961; Vygotsky, 1978). These methods are typically described as being opposite: the instruction-based method is teacher-centred and passive, while the exploration-based method is learner-centred and active.

Researchers have taken up the debate between instruction-based teaching and active learning, with unsatisfying results. Early cognitive science extolled the virtues of active learning (Bruner, 1961). However, more recently researchers have alternately found evidence for both teaching (Mayer, 2004) and active learning (Sweller, 1988; Gureckis & Markant, 2012). Researchers have also found instances of equivalence (Klahr & Nigam, 2004) and nuanced interplay between teaching and active learning when in sequence (Bonawitz et al., 2011; DeCaro & Rittle-Johnson, 2012). This has led researchers to propose new constructs, such as Guided Play (Weisberg, Hirsh-Pasek, & Golinkoff, 2013), which have moved forward the debate. However, when and why teaching or active learning may yield better outcomes remain largely unresolved.

One of the greatest barriers to resolving this debate is the difficulty in fully characterizing the pedagogical methods used and precisely controlling the conditions under which they are used (Prince, 2004). We argue that by abstracting away from the particulars of specific learning material (e.g., biology or mathematics) and idealizing the learning process (to be rational and normative), we may use computational models to formalize the methods and conditions and thus clarify when and why teaching may outperform active learning, and vice versa. Computational models of both teaching and active learning exist in the cognitive science and machine learning literature (e.g., Shafto & Goodman, 2008; MacKay, 1992). These models have been shown to describe human behavior well in a variety of simple learning settings and higher-level perceptions (Shafto, Goodman, & Griffiths, 2014; Castro et al., 2009; Gureckis & Markant, 2009; Yang, Lengyel, & Wolpert, 2016). However, the two types of models have never been compared in the same framework.

We provide the first computational comparison between teaching and active learning. Following previous computational work, we use a concept learning task to assess the effectiveness of the two pedagogical methods. To approximate some of the complexity of real-world learning problems, we design the concept space to be hierarchical and introduce *partial ambiguity* between concepts by introducing overlap between the concepts. As an illustration, the categories "bird" and "sea animals" are partially ambiguous due to the existence of examples that are in both categories. Such ambiguity makes certain examples uninformative for grasping distinctions between the concepts, thus limiting the performance of active learning. In these cases the guidance of a knowledgeable and helpful teacher may be of great import, by avoiding such ambiguous examples. Using varying degrees of ambiguity, we will investigate under what scenario is the effectiveness of teaching more pronounced.

Given perfect knowledge and rational inference, we know that teaching is at least as effective as active learning. This assumes that the teacher is knowledgeable and helpful and that the learner and teacher use the same inference scheme. There is little reason to believe that these assumptions are met in everyday educational settings (Chi, Siler, & Jeong, 2004). The effect of an unhelpful teacher is easy to imagine (random guidance), but the effects of a teacher with imperfect knowledge of a learner or a teacher operating with incorrect beliefs about the world is less apparent. To investigate this, we introduce a conceptually misaligned teaching model in which we vary the types and degrees of misalignment between the teacher's and learner's concept spaces. Using this, we explore scenarios under which exploration outperforms teaching.

## Framework

The most common and simplest concept learning tasks use concept spaces with a single-layer of Boolean concepts where the features can be discrete or continuous (e.g., Shafto & Goodman, 2008; Castro et al., 2009; Gureckis & Markant, 2009). While this is a sensible framework for certain questions, to accentuate differences between the models and ensure that we can model the kinds of challenging concept learning problems that appear in more realistic scenarios, we adopt a more complex concept space with two-layered Boolean concepts, where the features of the higher-level concepts are themselves concepts. In particular, a two-layered concept space allows us to introduce partial ambiguity between the concepts that only teaching, but not exploration, can resolve (explained later in more details with an example). Similarly, this allows us more interesting variations in the ways that teachers may be incorrect about learner's beliefs, or the true state of the world. Below we begin the description of the framework by describing the concept space and setting notations before presenting a detailed example, and the simulations results.

## Concept space

A concept space contains two concepts. A concept contains at least 1 and up to 6 distinct patterns. There are in total 6 types of patterns (Fig. 1 left), and they are all the Boolean concepts that have 4 features with balanced binary labels, that is, two features labeled 0 and two features labeled 1. Figure 1 provides two example concept spaces. Formally, we denote the concept space by $H$, the concepts by $h = \{h_1, h_2\}$, the $j^{\text{th}}$ pattern in a concept by $f_j$, the features by $x$, and the feature's binary label by $y = \{0, 1\}$.

The prior probability on the concept space is hierarchically uniform. This means that $P(h_1) = P(h_2) = \frac{1}{2}$ and that $P(f_j|h_k) = \frac{1}{N_k}$ for all $j$, where $N_k$ is the number of patterns in $h_k$. We say that the concepts are *ambiguous* when two concepts have shared patterns, and the degree of ambiguity, $a$, is defined by the number of shared patterns (see Fig. 1 for example). Recall that ambiguous concepts allow a stronger distinction to be made between teaching and active learning.

For the discussion of concept misalignment—where the teacher may be incorrect about the learner's beliefs or the true state of the world—we denote the learner's concept space by $H_L$, the teacher's concept space by $H_T$, and the true world space by $H_W$. We quantify the degree of misalignment, $m$, between two concept spaces by the the minimum number of pattern "moves" within a concept space to make the two concept spaces equivalent. The move operation includes moving a pattern between two concepts, removing a pattern, or adding a pattern. Misalignment with $H_W$ is sometimes referred to as *misconception*. Later we will investigate how the effectiveness of teaching degrades when the learner has misconceptions ($H_W = H_T \neq H_L$) and when the teacher has misconceptions ($H_W = H_L \neq H_T$).

## An example trial

First, a concept in $H_W$ is chosen as the correct answer; then, a pattern within that concept is chosen as the underlying pattern. On the first trial, a pedagogical method of choice (optimal exploration or teaching) computes scores (according to Eq. 3 or Eq. 6, respectively) for all four potential features.

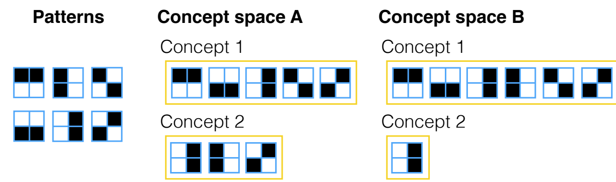


Figure 1: The six types of patterns, denoted $f$, are used to construct concepts. The positions—top left (TL), top right (TR), bottom left (BL), and bottom right (BR)—of the small squares with blue outline represent the four features. The colors—white or black—represents the binary feature labels which correspond to individual observations. For Concept space A, the prior for each pattern in Concepts 1 and 2 are $P(f|h_1)P(h_1) = \frac{1}{5} \times \frac{1}{2}$ and $P(f|h_2)P(h_2) = \frac{1}{3} \times \frac{1}{2}$, respectively. The degree of ambiguity, $a$, is 2 because $f_3$ and $f_5$ in $h_1$ are also in $h_2$. For Concept space B, $P(f|h_1)P(h_1) = \frac{1}{6} \times \frac{1}{2}$, $P(f|h_2)P(h_2) = 1 \times \frac{1}{2}$, and $a = 1$. The degree of misalignment between Concept spaces A and B is 2: one has to make two "moves" in Concept space A (delete the last pattern in Concept 2 and move the second pattern from Concept 2 to Concept 1) in order to make it equivalent to Concept space B.

The learner queries the world the feature with the highest score, and the world labels the query according to the predetermined underlying pattern. With this observation, the chosen pedagogical method computes the scores for the remaining three features. Then, the learner queries; the world labels; and the process repeats until every feature is queried. Before the first query and after each query, the learner's posterior belief (via Eq. 2) about the correct concept is recorded.

We now give a concrete example that compares optimal exploration with optimal teaching. We first name the features as Top Left (TL), Top Right (TR), Bottom Left (BL), and Bottom Right (BR; see Fig. 1 caption). Figure 2A shows the concept space under consideration, the target concept, and the target underlying pattern. In this case, the teacher's concept space, the learner's concept space, and the world are all aligned ($H_W = H_T = H_L$). Figure 2B shows the scores that exploration and teaching assign to each query. The reasoning behind the scoring is based on predicted outcomes. For example, if TL is queried and labeled black, then one can rule out all patterns in $h_2$ and be certain that $h_1$ is the answer. This is good. But if TR is queried and labeled black, one can only rule out $f_2$ in $h_1$ and $f_1$ in $h_2$, leaving the two concepts equally likely, which is not good. Following this reasoning, the active learner or teacher considers all the possible outcomes (if TL is white...; if TB is black...; if TB is white...; and so on) and chooses the one that best resolves the answer. In this case, before observing anything, both optimal exploration and optimal teaching scores TL or BR the highest. In this trial the learner chooses TL and observes white.

Given this data, $\mathcal{D} = \{x_1 = TL, y_1 = 1\}$, the learner rules out $f_1$ and $f_2$ in $h_1$; thus, at this point, the learner believes that $P(h_1|\mathcal{D}) = \frac{1}{4}$ and that $P(h_2|\mathcal{D}) = \frac{3}{4}$. Optimal exploration chooses a query that reduces the expected uncertainty about arriving at an answer. Intuitively, uncertainty is highest when $h_1$ and $h_2$ are equally likely and lowest when one is definitely correct. Following the above reasoning, an ac-

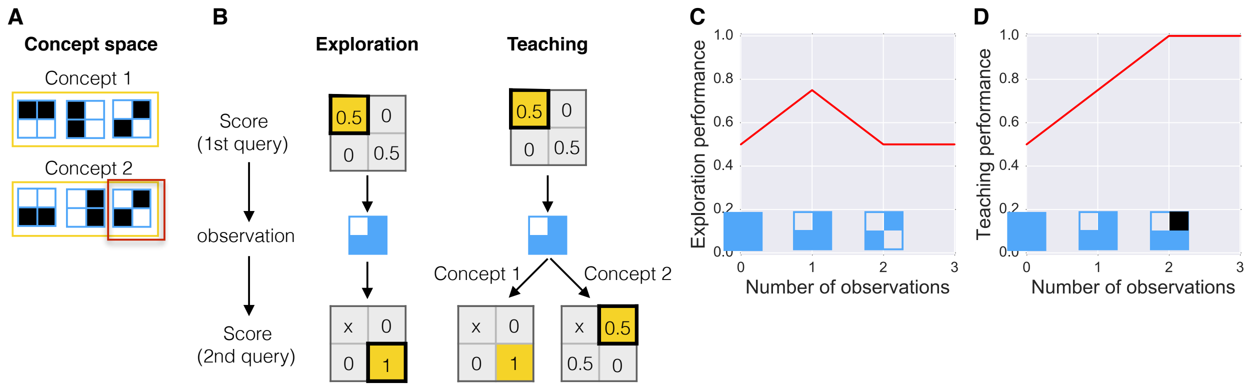


Figure 2: An example trial. **A.** The concept space under consideration. The predetermined underlying pattern is $f_3$ in $h_2$ (red box), which happen to be equivalent to $f_3$ in $h_1$. **B.** Query scorings for optimal exploration according to Eq. 3 (left) and for optimal teaching according to Eqs. 6 (right). The "x" indicates that an observation has been made on that feature, so that feature is excluded from the set of potential queries. The chosen query in each step is highlighted yellow and outlined with a thick border. **C-D.** Performances of optimal exploration and teaching following the observation sequences given in **B.** Blue squares indicate that the feature value has not yet been observed.

tive learner computes the expected uncertainty in each predicted case. After summing over all the expected uncertainty weighted by the chance those case would occur, the optimal explorer can then assign a score to each query to indicate how much certainty was gained (or uncertainty is reduced). Here, the sum shows that BR leads to the highest information gain, mainly because it helps reach certainty with a 50% chance.

Optimal teaching chooses a query to maximize the learner's inference for a desired concept. As such, the query depends on the answers that the teacher has in mind. The reasoning behind optimal teaching is again based on predicted cases and goes as follows: If the real answer is $h_1$, BR will be white, and TR and BL will be black. Hence, when BR is revealed, the learner will infer that both concepts are equally likely, and when TR or BL is revealed, the learner will consider $h_2$ to be more likely. Thus, to help the learner infer the hypothetical answer of $h_1$, the teacher will guide the learner to query BR, even though it is, in some sense, ambiguous. Following the same kind of reasoning, the teacher concludes that, for $h_2$, TR or BL is the better choice. Now, because the learner knows that the teacher is helpful, the learner can actually infer the answer with certainty just by the teacher's guidance because the guidance is answer-dependent.

This line of reasoning shows that optimal teaching can be better than optimal exploration for two reasons. First, the teacher helps *reduce irrelevant search* by tailoring guidance based on the answer. This is consistent with theories that support instructional-based teaching (Kirschner, Sweller, & Clark, 2006). Second, because the guidance depends on teacher's knowledge of the answer, the learner can leverage *pedagogical reasoning*—the fact that the teacher is knowledgeable and helpful—to make stronger inferences.

Figure 2C-D show the performance with optimal exploration and teaching, respectively. The performance is defined as the learner's posterior belief about the target concept after observing some data (Equation 2 is used for exploration, and Eq. 4 is used for teaching). The performance with optimal exploration eventually reaches chance level because the underlying pattern, $f_3$ in $h_2$, is ambiguous in this case. The performance with teaching reaches 1 even for the ambiguous pattern because in optimal teaching the learner can use pedagogical reasoning to break this ambiguity. This type of performance difference via pedagogical reasoning is not possible with single-layered concept space.

## Inference

The learner's inference follows Bayes' rule. Given some data $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$, the learner's joint posterior is

$$P(h,f|\mathcal{D}) = \frac{P(\mathcal{D}|h,f)\,P(h,f)}{\sum_{k,j} P(\mathcal{D}|h_k,f_j)\,P(h_k,f_j)} = \frac{1}{Z}P(\mathcal{D}|f)\,P(f|h)\,P(h)$$
$$= \frac{1}{Z}\prod_i P(y_i|x_i,f)\,P(f|h)\,P(h)\,. \quad (1)$$

In our framework, labelling is deterministic, so the likelihood $P(y_i|x_i, f_j)$ is either 0 or 1. The normalizing constant, $Z$, can be computed exactly by enumeration in our simple setting.

The joint posterior of Eq. 1 can be used to obtain the concept posterior,

$$P(h|\mathcal{D}) = \sum_j P(h, f_j|\mathcal{D})\,, \quad (2)$$

by marginalizing out $f$. It can also be used to obtain the pattern posterior, $P(f|\mathcal{D}) = \sum_k P(h_k, f|\mathcal{D})$, by marginalizing out $h$. This is used for computing the predictive distribution.

## Optimal exploration

We model optimal exploration following a Bayesian active learning model that chooses query $x$ to myopically maximize the expected information gain (MacKay, 1992). The probability of choosing an $x$ is

$$P(x|\mathcal{D}) = \lim_{\alpha\to\infty} \frac{1}{Z}\left[\langle \mathrm{H}[h|\mathcal{D}] - \mathrm{H}[h|\mathcal{D},x,y]\rangle_{P(y|x,\mathcal{D})}\right]^{\alpha}. \quad (3)$$

Here, $Z = \sum_{x'}\left[\langle \mathrm{H}[h|\mathcal{D}] - \mathrm{H}[h|\mathcal{D},x',y]\rangle_{P(y|x',\mathcal{D})}\right]^{\alpha}$ is the normalizer to produce a probability distribution over $x$, and

$H[h|\cdot] = -\sum_h P(h|\cdot) \log P(h|\cdot)$ is the Shannon entropy. Because $H[h|\cdot]$ is a measure of the uncertainty of the posterior, the difference in entropy before and after receiving a new observation pair $\{x, y\}$ in Eq. 3 quantifies the expected reduction in $h$ uncertainty, which is information gain. The expectation operator, $\langle \cdots \rangle_{P(y|x,\mathcal{D})}$, indicates that the learner does not know exactly whether the label for $x^*$ will be 0 or 1 but maintains a predictive distribution. The predictive distribution is given by $P(y|x,\mathcal{D}) = \sum_j P(y|x, f_j) P(f_j|\mathcal{D})$. The limit $\alpha \to \infty$ assigns probability uniformly over the $x$'s that produce the highest argument value. It returns the set of values that have the highest probability when there is more than one; it is equivalent to arg max when there is a single highest argument value.

## Optimal teaching

We define optimal teaching to satisfy three assumptions (Shafto & Goodman, 2008; Shafto et al., 2014). First, the teacher knows the correct answer. Second, $H_W = H_T = H_L$, and the teacher and learner use exactly the same inference scheme. Third, the teacher and the learner are cooperative. From the learner's perspective, this means that the learner reasons about how the teacher, knowing the answer, chooses the most helpful guidance. From the teacher's perspective, this means that the teacher provides guidance while being aware of the learner's inference.

We begin the formulation with the learner's inference. Given the guidance, $x$, from the teacher and the corresponding new observation, $y$, the learner's inference follows

$$P_L(h, f|x, y, \mathcal{D}) = \frac{P(y|x, f)\, P_T(x|h, \mathcal{D})\, P_L(f, h|\mathcal{D})}{\sum_{j,k} P(y|x, f_j)\, P_T(x|h_k, \mathcal{D})\, P_L(f_j, h_k|\mathcal{D})}. \tag{4}$$

Note that the teacher's guidance carries information about the answer via the likelihood, $P_T(x|h,\mathcal{D})$. The cooperative inference mentioned above can be modeled by combining Eq. 4 with

$$P_L(h|x,\mathcal{D}) = \frac{1}{Z} \left\langle \sum_j P_L(h, f_j|x, y, \mathcal{D}) \right\rangle_{P_L(y|x,h,\mathcal{D})} \tag{5}$$

$$P_T(x|h,\mathcal{D}) = \lim_{\alpha\to\infty} \frac{[P_L(h|x,\mathcal{D})\, P(x)]^\alpha}{\sum_{x'} [P_L(h|x',\mathcal{D})\, P(x')]^\alpha} \tag{6}$$

where $P_L(y|x,h,\mathcal{D}) = \sum_j P(y|x, f_j) P_L(f_j, h|\mathcal{D})$, $Z$ is a normalizer, and $P(x)$ is a uniform distribution over $x$. This system of equations (4-6) is first iterated until convergence, then an $x$ is sampled from Eq. 6 conditioned on the teacher's knowledge of the true concept. A sensible initial condition is a uniform $P_T(x|h,\mathcal{D})$ in Eq. 4. Compared to Eq. 3, the extra $h$ in the expectation operator, $\langle \cdots \rangle_{P_L(y|x,h,\mathcal{D})}$, shows that teacher and learner reasons about one $h$ at a time. The subscript $L$ in $P_L(\cdot)$ emphasizes that the concept-pattern joint prior and posterior are based on the learner's reasoning; the subscript $T$ in $P_T(x|h,\mathcal{D})$ emphasizes that the $x$ is based on the teacher's reasoning; and the unsubscripted $P(y|x, f)$ indicates that the label likelihood is provided by the true world. [1]

## Conceptually misaligned teaching

Optimal teaching is always at least as good as optimal exploration because the teacher's guidance offers extra information about the correct answer. But what happens when the assumptions of optimal teaching are violated? We consider two types of violation that breaks the second assumption of the learner and teacher sharing the same inference by introducing misconception in the learner (type 1: $H_W = H_T \neq H_L$) and misconception in the teacher (type 2: $H_W = H_L \neq H_T$). Note that the first and third assumptions of optimal teaching are still kept. The first assumption poses that regardless of whether the teacher has misconception, the teacher knows the correct concept label. The third assumptions poses that the teacher and learner still reason cooperatively. This leads us to introduce conceptually misaligned teaching, where the two agents, the learner and teacher, reason about each other while wrongly assuming the other agent's concept space. Computationally, the teacher provides $x$ by going through Eqs. 4-6 with $H_T$, thinking that the learner also operates with $H_T$. Having received $x$, the learner also goes through Eqs. 4-6, while assuming that the teacher also has $H_L$ in mind. The first type of violation (misconception in learner) is a common issue in education (Chi et al., 2004), and the second type (misconception in teacher but not in learner) is a natural counterpart simulation to do.

# Simulations: systematic comparison

In the Framework section, we gave a detailed example of how exploration compares to teaching given a particular underlying pattern and concept space. In this section, we compare the performance of the two pedagogical methods in a systematic manner over different classes of concept spaces. This will lead us to address more broadly the condition under which teaching is better than active learning and vice versa. To this end, we consider three scenarios.

For all three scenarios, a concept space with degree of ambiguity, $a$, contains $6 + a$ patterns. All 6 patterns in Fig. 1 are used at least once, but no pattern is used more than twice. Figure 1 shows two example concept spaces that satisfy the above criteria. The first scenario assumes that $H_W = H_L = H_T$ and entertains concept spaces of varying degree of ambiguity from $a = 0$ to 5. For a given $a$, the simulation includes all combinations of assignments of $6 + a$ patterns to two concepts, with isomorphic concept spaces counted only once. The second scenario assumes $H_W = H_T \neq H_L$ and a fixed $a = 1$. We consider all pairings of $H_T$ and $H_L$ with $a = 1$ up to concept-space pair isomorphism, and label each pair with their degree of misalignment, $m$, which can vary from 0 to 6. The third scenario assumes $H_W = H_L \neq H_T$ and $a = 1$. The

[1] The formulation here appears different from (Shafto et al., 2014) because the setup is different in two ways. First, the query and the label are kept distinct to match the setup of exploration. This gives rise the two likelihoods in Eq. 4, one for $y$ and one for $x$. It also gives rise to the expectation in Eq. 5 with respective to the predictive distribution over $y$. Second, in this setup, the teacher knows the concept from which the pattern is drawn from but not the pattern itself; thus, there is the marginalization over patterns in Eq. 5.

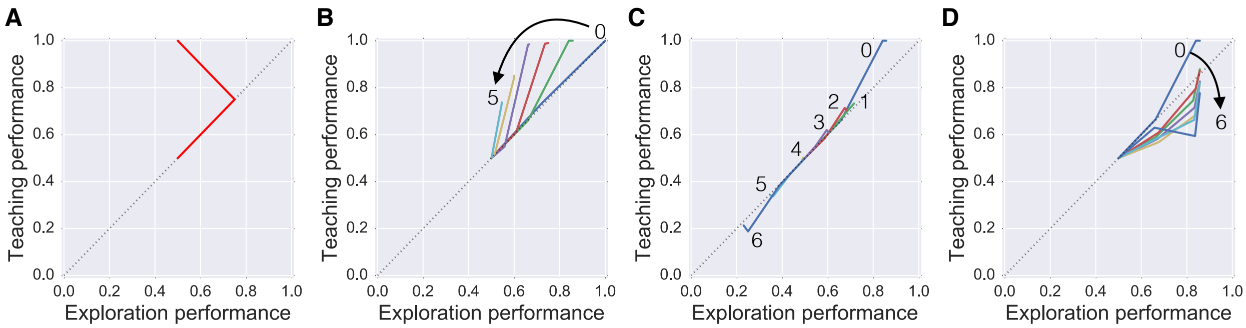


Figure 3: **A.** Double-performance plot for the trial in Fig. 2. **B.** Double-performance plot averaged over concept spaces with differing degree of ambiguity. The numbers indicate the $a$ of the curve, ranging from 0 to 5 in increments of 1 from the rightmost curve to the leftmost curve. **C.** Double-performance plot for a learner with differing degree of misconception, $m$, as described in the Concept Space section under Framework. The numbers indicate the $m$ of the curves. All concept spaces have $a = 1$, so the curve with $m = 0$ matches the curve with $a = 1$ in **A**. **D.** Double-performance plot for a teacher with differing degree of misconception. The numbers again indicate $m$ and increases from the top to bottom in order of the curves' end points.

pairing and labeling are done as in the second scenario.

To visualize the relative performances of exploration and teaching, we plot the two pedagogical methods against each other (Fig. 3A). On this double-performance plot, curves above the diagonal indicates that teaching is better than optimal exploration, and curves below indicates that teaching is worse. To reveal higher-order trends, we average single-trial performances firstly over the patterns in individual concept space, weighted by the patterns' prior probabilities, and secondly over concept spaces that have the same label; that is, the same $a$ or $m$, with each concept space contributing equal weight. Figure 3B, C, and D show the relative performances under the first, second, and third scenario, respectively.

Figure 3B shows that optimal teaching is better than optimal exploration when the concepts are partially ambiguous but no better or worse when the concept spaces are fully unambiguous. When the concept space is fully ambiguous, performances remain at chance level for both teaching and exploration. The advantage of of teaching in absolute performance, as judged by the end points on the double-performance plot, is most pronounced with a 30% improvement over exploration at a medium degree of ambiguity of $a = 3$. Overall, teaching is better than exploration under partial ambiguity because of the reduction in irrelevant search and pedagogical reasoning described in the example task.

Figure 3C shows that on average, teaching and exploration perform similarly when the learner's concept space is wrong. On the one hand, this shows that teaching is robust against learner's misconceptions (in terms of hurting learning) even when the misconception is strong; on the other hand, it shows that the benefits of teaching (in terms of boosting learning) diminishes in the face of a little misconception. The trial-by-trial performances reveal extreme cases when teaching is much worse than exploration (0% vs. $> 80\%$). This happens when the learner's concept space looks like Concept space B in Fig. 1, where $h_1$ contains all the patterns and $h_2$ contains only one pattern which is also in $h_1$. This suggests that the learner's strong prior bias for an ambiguous pattern ($f_3$ from $h_1$ or $f_1$ from $h_2$) being associated with a particular hypothesis (i.e., with $h_2$) can distort the effect of pedagogical reasoning when there is conceptual misalignment.

Figure 3D shows that teaching with the wrong concept space leads to evidently poorer performance (roughly 10-20% worse) during learning, but only somewhat poorer performance (up to about 10% worse) at the end of learning (after observing all feature values). The exceptions are when $0 < m < 3$; then teaching ends up very slightly ($< 5\%$) better. Thus, while the advantage of teaching quickly diminishes with misalignment, the final performance of teaching is rather robust. Interestingly, the first and third scenarios combine to show that exploration before teaching is better than the other way around, as the advantage of teaching comes in after the first observation (Fig. 3C), but its disadvantage due to potential misalignment comes in at the first query (Fig. 3E). This is consistent with previous findings on the interplay between exploration and teaching in learning mathematical concepts (Schwartz & Martin, 2004; DeCaro & Rittle-Johnson, 2012).

In summary, our analysis shows that if one knows little about the structures of $H_W$, $H_T$, and $H_L$ and their alignments, teaching is the preferred pedagogical method because it can potentially be much better and will unlikely be much worse than exploration. If one knows that there is moderate amount of misalignment, exploration is the preferred method. If one knows the detailed structures of $H_W$, $H_T$, and $H_L$, the alignments among them, and a particular target concept, detailed analysis should be done to choose the better method.

Lastly, it is worth considering our results in the context of popular explanations in favor of active learning. These reasons, including attentional control, enhanced memory, stress relief, and etc. (Springer, Stanne, & Donovan, 1999; Prince, 2004; Markant, Ruggeri, Gureckis, & Xu, 2016), all suggest that a learner who explores actively maintains an $H_L$ more consistent with $H_W$ than a leaner who receives passive teaching guidance. A model that includes pedagogical-method-dependent effect on the concept space is an important direction for future work.

## Discussion

Researchers have debated whether teaching or active learning is better for decades without reaching a consensus. A main barrier is the difficulty in precisely controlling the pedagogical method and learning conditions for meaningful comparison of results that support generalization. We argue that by formalizing the learning conditions and the pedagogical methods, we may clarify when and why which pedagogical method is more effective. We adopt existing models for optimal exploration and optimal teaching and introduce a model for conceptually misaligned teaching. Our computational analysis showed that optimal teaching is much better than exploration when the concepts in play are partially ambiguous, but this effect diminishes very quickly with conceptual misalignment between the learner and teacher.

We expect this trend in our results to generalize to larger concept spaces with richer structures. Optimal teaching should be increasingly more effective than optimal exploration because a larger space allows for greater reduction in irrelevant search and a richer structure allows for finer pedagogical reasoning. However, when there is conceptual misalignment, we expect this advantage of teaching to diminish quickly because the ways of misinterpreting guidance and observation also increases with the complexity of the concept space. The exact scaling between the rate of diminishing benefit and the size and complexity of concept space and is an interesting question for future work.

We have focused on the pedagogical method that best leads the learner to a specific concept. This approach is common to concept learning experiments. However, in many real-world scenarios, we may also be interested in generalization: what does performance on one task predict about future performance on related tasks. To capture this, we would need to consider concept spaces with much richer structure that would support incremental building of compositional concepts and/or transfer learning. Although beyond the scope of the current research, even defining what such conceptual structures should look like in order to capture some of the richness of real-world concept learning problems is another interesting question for future work.

Debates about the relative efficacy of different pedagogical methods have plagued the literature. Because of the complexity of concepts and the variability in the application of the pedagogical methods themselves, empirical tests have been largely inconclusive. Our approach has been to abstract away from some of the details and ask the question in an idealized setting: under what circumstances would we expect teaching or active learning to perform better. Our results yielded the surprising conclusion that, even when the assumptions of teaching are not perfectly met, it is quite robust. While there is more work to be done to capture the richness of psychological theories of active learning, our approach provides a way forward where empirical research has not been as successful as initially hoped. Considerable work remains, but systematic computational analysis of theories themselves provides a potentially promising complement to more traditional empirical methods for uncovering more optimal methods of delivering educational content.

## Acknowledgments

This research was supported in part by NSF CAREER award DRL-1149116 to P.S.

## References

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

Bruner, J. S. (1961). The act of discovery. *Harvard educational review*.

Castro, R. M., Kalish, C., Nowak, R., Qian, R., Rogers, T., & Zhu, X. (2009). Human active learning. In *Advances in neural information processing systems* (pp. 241–248).

Chi, M. T., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and instruction*, *22*(3), 363–387.

DeCaro, M. S., & Rittle-Johnson, B. (2012). Exploring mathematics problems prepares children to learn from instruction. *Journal of experimental child psychology*, *113*(4), 552–568.

Gureckis, T. M., & Markant, D. (2009). Active learning strategies in a spatial concept learning game. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 3145–3150).

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning a cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

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*(2), 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*, *15*(10), 661–667.

MacKay, D. J. (1992). Information-based objective functions for active data selection. *Neural computation*, *4*(4), 590–604.

Markant, D. B., Ruggeri, A., Gureckis, T. M., & Xu, F. (2016). Enhanced memory as a common effect of active learning. *Mind, Brain, and Education*, *10*(3), 142–152.

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

Prince, M. (2004). Does active learning work? a review of the research. *Journal of engineering education*, *93*(3), 223–231.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129–184.

Shafto, P., & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 1632–1637).

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71*, 55–89.

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of educational research*, *69*(1), 21–51.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257–285.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the development of children*, *23*(3), 34–41.

Weisberg, D. S., Hirsh-Pasek, K., & Golinkoff, R. M. (2013). Guided play: Where curricular goals meet a playful pedagogy. *Mind, Brain, and Education*, *7*(2), 104–112.

Yang, S. C.-H., Lengyel, M., & Wolpert, D. M. (2016). Active sensing in the categorization of visual patterns. *Elife*, *5*, e12215.