Leveraging Thinking to Facilitate Causal Learning from Intervention

Yuan Meng
yuan_meng@berkeley.edu
Department of Psychology
University of California, Berkeley

Fei Xu
fei_xu@berkeley.edu
Department of Psychology
University of California, Berkeley

Abstract

Intervention selection is at once crucial in causal learning and challenging for causal learners. While the optimal strategy is maximizing the expected information gain (EIG), both children and adults often combine it with suboptimal ones such as the positive test strategy (PTS). In the current study, we sought to facilitate causal learning from intervention by asking 5- to 7-year-olds to explain why they chose a certain intervention to identify the true structure of a three-node causal system that might work in one of two ways. Our findings suggest that while engaging in self-explaining did not help children select more informative interventions, asking them to think about their intervention choices (explaining or reporting) might help them better utilize interventional data to infer causal structures.
Keywords: causal learning; intervention; explanation; learning by thinking

Once upon a time in China, two men were accused of a murder yet no evidence could be found. The judge gave each of them a "magical" straw that was said to grow longer in the hands of the guilty. As the story goes, the man showing up with a *shorter* straw next day was put in jail. As you might have guessed, straws don't grow; the real magic is that the judge chose the most informative intervention centuries before informative theory came into being. He could not foresee which man would cut his straw in fear but whoever did it must be the murderer. This strategy allowed him to maximally reduce his uncertainty averaged across potential outcomes, or in other words, maximize his *expected information gain* (EIG)

EIG is widely regarded as a normative model for inquiry selection (Coenen, Nelson, & Gureckis, 2018; Nelson, 2005) but it only partially captures people's actual interventions. Both adults (Bramley, Lagnado, & Speekenbrink, 2014) and children (McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016) outperform models that intervene randomly but fall short of pure EIG maximization. On the computational level, a possible explanation is that adults (Coenen, Rehder, & Gureckis, 2015) and children (Meng, Bramley, & Xu, 2018) combine EIG maximization with a suboptimal strategy akin to the *positive test strategy* (PTS) in the rule learning literature (Klayman & Ha, 1989; Wason, 1960). In causal learning, Coenen et al. (2015) defined PTS as a tendency to generate the most expected effects under your current causal hypothesis. A minimal example of PTS is intervening on X when you try to discriminate between your hypothesis, $X \rightarrow Y \rightarrow Z$, and an alternative one, $Y \leftarrow X \rightarrow Z$. If the outcome (e.g., only X and Y are activated) happens to *falsify* your hypothesis, you get to rule it out; otherwise, both could still be true so you remain uncertain. By contrast, a high-EIG intervention (Y) reduces your hypothesis space (in this case, to 1) regardless of the outcome (all variables are activated or only Y and Z).

To ensure successful causal learning from intervention, learners should use an optimal strategy (e.g., EIG) to select interventions and make accurate inferences from interventional data. In the current study, we sought to facilitate both the intervention selection and the belief updating processes by prompting learners to *explain* why they choose a certain intervention to learn about an unknown causal system.

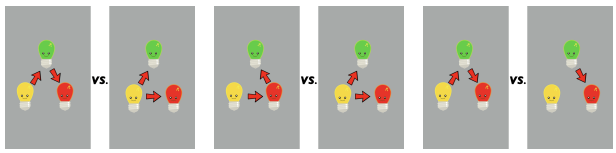
Explanation and intervention

Explaining requires no extra data or instructions; yet, it has profound downstream consequences for learning and inference in various domains (see Fonseca & Chi, 2011; Lombrozo, 2016, for reviews). Typically, learners achieve better learning outcomes simply by engaging in explanation (e.g., how a system works, why an effect occurred, etc.) even without feedback or generating accurate explanations (e.g., Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, De Leeuw, Chiu, & LaVancher, 1994; Walker, Lombrozo, Legare, & Gopnik, 2014; Walker, Lombrozo, Williams, Rafferty, & Gopnik, 2017). Many theories are proposed to explain why explaining facilitates learning, such as that it helps learners fill gaps in their knowledge, repair erroneous mental models, recruit criteria for "good" explanations (simplicity, breadth, or other "explanatory virtues") to constrain reasoning, etc..

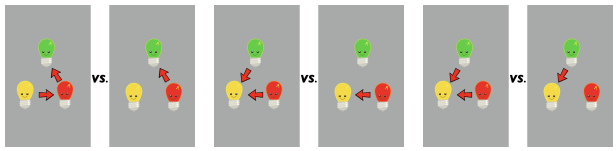
How do you explain an intervention? From an EIG perspective, to explain intervention selection, you must consider belief updating (Coenen & Gureckis, 2015): You choose an intervention because on average, it reduces the most uncertainty. Engaging in explanation may benefit both processes.

Explaining may facilitate intervention selection by promoting *comparison* and *abstraction*. A recent study (Edwards, Williams, Gentner, & Lombrozo, 2019) suggested that asking learners to explain exemplars' category membership (e.g., "Why is this robot a Glorp/Drent?") increased their comparison within and between categories. Should explainers compare more across different interventions and the outcomes of each intervention, they might be in a better place to select high-EIG interventions. Moreover, effective learners may realize that on an abstract level, informative interventions are ones that yield distinct outcomes under different hypotheses. Walker and Lombrozo (2017) found that explaining the outcome of a story (e.g., why a character is sad) helped children extract its underlying moral and go beyond the specifics. Should explaining help causal learners achieve such abstraction, it could largely reduce the cost of intervention selection.

Explaining may also facilitate belief updating by encouraging learners to apply their prior knowledge when interpreting interventional data (Williams & Lombrozo, 2013).



(a) Puzzle 1 (b) Puzzle 2 (c) Puzzle 3



(d) Puzzle 4 (e) Puzzle 5 (f) Puzzle 6

Figure 1: Light bulb puzzles used in the experiment.

Current study

In the current study, we investigated whether self-explaining could facilitate causal learning from intervention. We chose to test 5- to 7-year-olds because previous studies (McCormack et al., 2016; Meng et al., 2018) suggested that they were not yet able to reliably select informative interventions, leaving substantial room for improvement. This also allows us to compare our results directly to that in Meng et al. (2018).

Overview of experiments Our causal learning task was adapted from Meng et al. (2018). Children were tested on six unknown causal systems consisted of three light bulbs, some of which could turn on others if activated. Each system could work in one of two ways and children were allowed to turn on one light bulb to identify its correct structure. All causal connections were deterministic with no background noise.

In the first experiment (Experiment 1A), children were asked to explain their intervention choice ("Why did you turn on that light bulb?") after carrying it out. However, since those children observed the outcome before explaining, their explanation might be a *post hoc* justification of their choice ("Because it helped me solve the puzzle.") rather than the actual reason. To address this concern, we conducted a second experiment (Experiment 1B) where children pointed to the intervention they wanted to perform and were asked to explain their choice ("Why do you want to turn on that light bulb?") before carrying it out. In the respective control conditions, children were asked to *report* which intervention they carried out (Experiment 1A) or planned to choose (Experiment 1B).

Modeling intervention strategies To compare intervention strategies across conditions, we took a hierarchical Bayesian approach used by Coenen et al. (2015) and Meng et al. (2018). We compared models of three single strategies (EIG, PTS, and random selection) and a linear combination of EIG and PTS. Below is an overview of the four models.

Learners all begin with a set of causal hypotheses, each of which can be represented as a directed acyclic graph $g \in G$ (G is the space of possible graphs), or a *causal Bayesian network* (Pearl, 2000). In each graph, causal variables are presented as nodes and causal relationships as edges.

1. Expected information gain (EIG)

   The information gain (IG) after intervening on the node $n \in N$ is the difference between the initial entropy, $H(G)$, and the entropy conditioned on the outcome o, $H(G|n,o)$:

   $$IG(n,o) = H(G) - H(G|n,o). \tag{1}$$

   Since o is unknown, the expected information gain (EIG) over all possible outcomes O is used to estimate IG:

   $$EIG(n) = H(G) - \sum_{o \in O} P(o|n) H(G|n,o). \tag{2}$$

   Applying Shannon's entropy equation, we have

   $$H(G) = -\sum_{g \in G} P(g) log_2 P(g), \tag{3}$$

   and

   $$H(G|n,o) = -\sum_{g \in G} P(g|n,o) log_2 P(g|n,o). \tag{4}$$

   The prior probability $P(g)$ of each graph g is assumed to be equal and the posterior probability $P(g|n,o)$ is given by Bayes' rule, $\frac{P(o|g,n)P(g)}{\sum P(o|g,n)P(g)}$. $P(o|g,n)$ is the likelihood of an outcome o given a hypothesis g and an intervention n.

2. Positive test strategy (PTS)

   PTS manifests as the tendency to intervene the node $n \in N$ with the most of direct or indirect descendant links (normalize by the total number of links in each graph $g \in G$):

   $$PTS(n) = \max_g [\frac{DescendantLinks_{n,g}}{TotalLinks_g}]. \tag{5}$$

3. Random selection

   Random selection is equivalent to indiscriminately assigning the same value (e.g., 1) to all possible interventions.

4. Linear combination of EIG and PTS

   Rather than sticking to one strategy, learners may use multiple strategies such as EIG and PTS to select interventions. The value of each possible intervention is a linear combination of its EIG and PTS values (the weight of EIG is θ).

Under one strategy or another, each possible intervention is assigned a value $V(n)$. An ideal learner should always select the intervention with the highest value but due to noise τ in the decision process, an actual learner often does so probabilistically. According to the *softmax choice rule* (Luce, 1959), the probability that an intervention gets chosen, $P(n)$, is a function of its value $V(n)$ and the learner's decision noise τ:

$$P(n) = \frac{\exp(V(n)/\tau)}{\sum_{n \in N} \exp(V(n)/\tau)}. \tag{6}$$

When τ is 0, the learner selects interventions with the highest values; when τ approaches $+\infty$, they select randomly.

Experiments

Participants

Seventy-four 6- to 7-year-olds participated in Experiment 1A, 37 of whom were assigned to the Explanation condition (M = 85 months, range = 74–101 months, SD = 9 months) and 37 to the Report condition (M = 84 months, range = 64–96 months, SD = 8 months). Another forty-three 5- to 7-year-olds participated in Experiment 1B, 22 of whom were assigned to the Explanation condition (M = 77 months, range = 62–90 months, SD = 8 months) and 21 to the Report condition (M = 75 months, range = 50–101 months, SD = 14 months).

Equipment

Three light bulbs (yellow, green, and red) were presented on a laptop screen and controlled by three buttons of corresponding colors located on a response board. During practice, red arrows indicated the causal relationships among the light bulbs. During the test, the arrows were hidden but two possible structures were shown on two cards placed side by side.

Procedure

Both experiments included a familiarization phase, a practice phase, and a test phase. During familiarization, children were taught to use buttons on a response board to control light bulbs of corresponding colors on the computer. During practice, they saw four basic types of structures: Common Cause (*Yellow* ← *Green* → *Red*), Common Effect (*Yellow* → *Red* ← *Green*), Causal Chain (*Green* → *Red* → *Yellow*), and One Link (*Yellow* → *Red*). In Experiment 1A, the presentation order was randomized. For each structure, children decided when to turn on which light bulb and were asked to describe the outcome of each action. In Experiment 1B, each structure was one change apart from the previous one. The simplest structure, One Link, was presented first, which was followed by Causal Chain, Common Cause, and Common Effect. For each structure, children turned on the light bulbs in a designated order (*Yellow–Red–Green* in the first two trials and *Green–Red–Yellow* in the last two) and were asked to predict and then describe each action's outcome.

On each of the six test trials, children were shown two ways in which the three light bulbs might work and were told that they could only turn on one light bulb to find out the true structure. In Experiment 1A, children were asked to explain ("Why did you turn on that light bulb?") or report ("Which light bulb did you turn on?") the intervention that they had just carried out. In Experiment 1B, children were asked to first point to the light bulb they planned to turn on, then explain ("Why do you want to turn on that light bulb?") or report ("Which light bulb do you want to turn on?") their choice, and finally perform the intervention[1]. At the end of both experiments, children were asked to put a smiley face sticker on the correct causal structure. In order to avoid potential discouragement that we observed during piloting, feedback was only provided after the entire experiment.

[1] In the rare event that children's actual intervention differed from what they planned, we used the former for all our analyses.

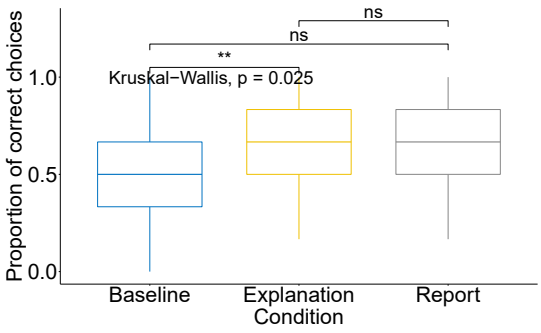


Figure 2: The proportion of causal structures that children correctly identified in each condition.

Results

Our initial analysis revealed no differences between the results of Experiments 1A and 2B, so data from these two experiments were pooled together in all subsequent analyses. To test whether explaining and reporting one's intervention choices could both influence causal learning, we used children in Meng et al. (2018) as our baseline. Apart from the additional explanation/report prompts, our procedure, stimuli, and population were identical to those in the previous study.

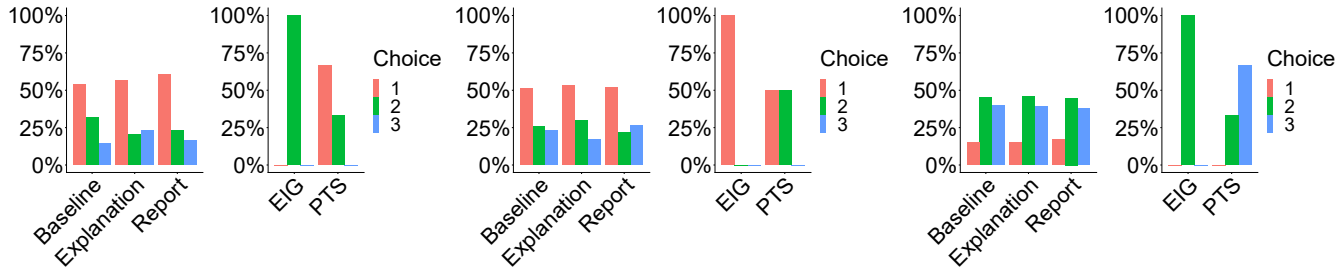
Inference accuracy To begin, we first looked at whether children were able to identify the correct causal structures in the end. As shown in Figure 2, those in the Baseline condition chose the correct structures 54% (SD = 22%) of the time, which was not distinguishable from chance (50%), $t(38)$ = 1.02, p = .31, Cohen's d = .16. However, children performed above chance in both the Explanation (M = 67%, SD = 25%) and the Report (M = 61%, SD = 23%) conditions, $t(58)$ = 5.13, $p < .001$, Cohen's d = .67 and $t(57)$ = 3.77, $p < .001$, Cohen's d = .50, respectively. The only significant difference between conditions was that explainers were more accurate than the baseline, $t(88.53)$ = 2.73, p = .007, Cohen's d = .55.

Intervention choices Before fitting models of intervention strategies, we examined children's interventions choices to see if they were random or biased towards EIG or PTS.

We compared the mean EIG and the mean PTS value of children's chosen interventions against the respective chance levels (.33 for EIG[2] and .55 for PTS[3]) of the two metrics. In the Baseline condition, only the mean PTS value (M = .74, SD = .22) was above chance, $t(38)$ = 5.37, $p < .001$, Cohen's d = .86, but not the mean EIG value (M = .39, SD = .28), $t(38)$ = 1.23, p = .23, Cohen's d = .20. Similarly in the Report condition, the mean PTS value (M = .74, SD = .20) was above

[2] Among all three possible interventions in each puzzle, only one was informative, i.e., having an EIG value of 1.

[3] This was the average PTS value across all interventions.



(a) Causal Chain vs. Common Cause (b) Causal Chain vs. One Link (c) Common Effect vs. One Link

Figure 3: Figures on the left show the proportion of children intervening on each node (n_1, n_2, and n_3) in each type of puzzles: (a) Causal Chain vs. Common Cause, (b) Causal Chain vs. One Link, and (c) Common Effect vs. One Link. Figures on the right show the probability of children intervening on each node in each type of puzzles predicted by EIG and PTS.

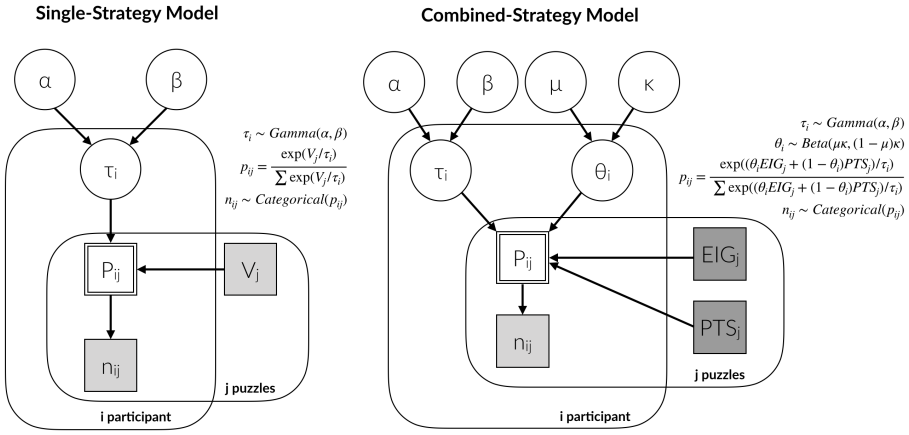


Figure 4: Hierarchical Bayesian models of single (left) and combined (right) strategies. In each puzzle j, each participant i chooses one node n_{ij} to intervene on. V_j, EIG_j, and PTS_j store the values of three possible interventions in each puzzle. p_{ij} stores probabilities of each participant choosing each intervention in each puzzle. τ_i and θ_i capture each participant's decision noise and weight of EIG. α and β are population-level hyper-parameters that generate τ_i; μ and κ generate θ_i.

chance, $t(57)$ =7.17, $p < .001$, Cohen's d = .94, but not the mean EIG value (M = .39, SD = .25), $t(57)$ = 1.63, p = .11, Cohen's d = .21. In the Explanation condition, however, both the mean EIG (M = .44, SD = .32) and the mean PTS (M = .75, SD = .18) value were above chance , $t(58)$ = 2.52, p = .014, Cohen's d = .33 and $t(58)$ = 8.4, $p < .001$, Cohen's d = 1.09, respectively. Neither the mean EIG or the mean PTS value differed significantly across conditions.

We also compared the proportion of children intervening on each node in each puzzle against what EIG and PT would predict. Since the mapping between node positions and light bulb colors is arbitrary, we re-coded Puzzles 1 and 2 as $n_1 \rightarrow n_2 \rightarrow n_3$ (Chain) vs. $n_2 \leftarrow n_1 \rightarrow n_3$ (Common Cause), Puzzles 3 and 4 as $n_1 \rightarrow n_2 \rightarrow n_3$ (Chain) vs. $n_2 \rightarrow n_3$ (One Link), and Puzzles 5 and 6 as $n_2 \rightarrow n_1 \leftarrow n_3$ (Common Effect) vs. $n_3 \rightarrow n_1$ (One Link). As Figure 3 shows, children deviated the most from EIG predictions in "Chain vs. Common Cause". In the other two types of puzzles, children's choices were split between EIG and PTS predictions. A small but non-negligible proportion of interventions were on nodes whose EIG and PTS values were both 0, suggesting that children occasionally chose interventions randomly.

Intervention strategies We used two hierarchical Bayesian models to capture children's intervention strategies (Figure 4). The single-strategy model draws from a single source to evaluate interventions—be it EIG, PTS, or always "1" in the case of random selection. The combined-strategy model assigns a weighted mean of EIG and PTS (the weight of EIG is θ) to each intervention. In both models, each child's decision noise τ_i is sampled from a population-level gamma distribution with two hyper-parameters α (shape) and β (rate). In the combined-strategy model, each child's weight of EIG θ_i is sampled from a population-level beta distribution with two hyper-parameters μ (mean) and κ (standard deviation. Uninformative priors are chosen for all hyper-parameters: α = .001, β = .001, $\mu \sim$ Beta(.5, .5), $\kappa \sim$ Gamma(.001, .001). The probability of selecting a given intervention is a function of its value $V(n)$ as well as the child's decision noise τ. Actual interventions are sampled from a categorical distribution of these probabilities. Parameter values were estimated using

Table 1: The deviance information criteria (DIC) of each model and the weight of EIG θ in three conditions.

	Baseline		Explanation		Report	
Model	DIC	θ	DIC	θ	DIC	θ
Random	514.15	–	777.82	–	764.64	–
EIG	481.98	–	727.00	–	769.45	–
PTS	469.62	–	**706.87**	–	**706.93**	–
EIG + PTS	**454.95**	.24	634.43	.31	746.25	.19

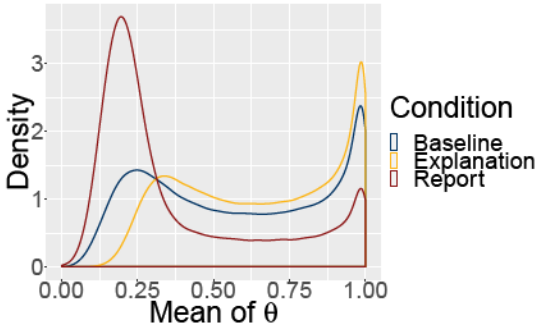


Figure 5: The distributions of the group-level hyper-parameter μ (the mean of θ) under the three conditions.

Markov chain Monte Carlo (MCMC) samples generated by the JAGS program[4] (Plummer, 2003). The deviance information criterion (DIC, Spiegelhalter et al., 2002) was used for model comparison. Models that fit data better (smaller posterior mean of the deviance $\bar{D}$) or are simpler (smaller effective number of parameters p_D) have lower DIC ($= \bar{D} + p_D$). As a common practice, a difference over 10 is substantial.

As shown in Table 1, the combined-strategy model (EIG + PTS) best captured children's intervention strategy in both the Baseline and the Explanation conditions. However, the PTS-only model turned out to be the best fit in the Report condition. Children in all three conditions relied more on PTS than EIG, with the mean weight of EIG being .24, .31, and .19, respectively. Figure 5 illustrates the distributions of μ—the population-level hyper-parameter that captures the mean of θ—in all three conditions. To see whether μ differed across conditions, we sampled 10,000 estimates of μ in each condition. For each contrast between conditions (Explanation vs. Report, Explanation vs. Baseline, Report vs. Baseline), we paired the estimates randomly and calculated the differences. Since the 95% Highest Density Interval (HDI) of all three difference distributions contained 0, we couldn't claim with confidence that μ differed across three conditions.

[4]In keeping with Meng et al. (2018), we ran MCMC for 100,000 iterations, discarding the first 1,000 samples and drawing one sample every 10 iterations. To ensure that samples were from a stationary distribution, we repeated this process 30 times with different initial parameter values and results from each sequence of samples (or *chain*) successfully converged since Gelman and Rubin's diagnostic $\hat{R}$ (Gelman & Rubin, 1992) of all parameters was smaller than 1.05.

Intervention and inference Lastly, we looked at whether children's intervention choices and strategies predicted if they could accurately identify the true causal structures.

First, for each puzzle, we performed a logistic regression using the EIG value (0 or 1) of children's chosen intervention to predict whether they identified the correct structure later. In the Baseline condition, EIG values did not predict inference accuracy in any puzzles. However, in the Explanation condition, high-EIG interventions strongly predicted successes at identifying the correct structures in all six puzzles. In the Report condition, EIG values predicted inference accuracy in four of the six puzzles (except Puzzles 2 and 6).

We examined the correlation between the weight of EIG θ and children's average accuracy across all puzzles. θ and average accuracy were uncorrelated in the Baseline condition, $F(1,37) = 1.14$, $p = .29$, $\bar{R}^2 = .0038$, but positively correlated in the Explanation and the Report conditions, $F(1,57) = 30.73$, $p < .001$, $\bar{R}^2 = .34$ and $F(1,56) = 25.27$, $p < .001$, $\bar{R}^2 = .30$, respectively. Correlations in the Explanation and the Report conditions were both stronger than that in the Baseline condition, $z = 2.31$, $p = .02$ and $z = 2.08$, $p = .04$, respectively.

Discussion

In the current study, we investigated whether asking children to explain their intervention choices facilitated causal learning from intervention. Specifically, we looked at 1) whether explainers were better able to select informative interventions and 2) make accurate inferences based on interventional data.

Our first hypothesis was not supported by the results. Neither children's weight of EIG θ nor the group-level hyper-parameter μ that captures the mean of θ differed significantly across the Baseline, the Explanation, and the Report conditions; this suggests that children used similar strategies to select interventions across three conditions. Curiously, asking children to report their intervention choices might have slightly "backfired": While a linear combination of EIG and PTS best captured children's intervention strategy in the Baseline and the Explanation conditions, the PTS-only model turned out to best characterize the strategy used in the Report condition. Moreover, unlike in the other two conditions, the distribution of μ was right skewed in the Report condition, indicating heavier reliance on PTS. However, since differences in μ were not statistically significant, further investigation is needed to examine whether this finding was due to random noise or potential drawbacks of the report prompts.

Compared to the chance performance in the Baseline condition, children in both the Explanation and the Report conditions were more accurate at identifying the correct causal structures after performing interventions. Since children in the latter two conditions did not choose more informative interventions, a possible explanation is that when prompted to think about their intervention choices, children were better able to utilize interventional data that were already available. This explanation was supported by our findings: In the Explanation and the Report conditions, children's intervention choices (EIG value = 0 or 1) and interventions strategies

(weight of EIG θ) predicted their inference accuracy, which was not the case in the Baseline condition.

General Discussion

In the current study, we looked at whether asking children to think about their intervention choices might facilitate their causal learning from intervention. In Experiments 1A and 1B, 117 5- to 7-year-olds solved six puzzles where they performed one intervention to identify the true structure of three light bulbs that might be connected in one of two ways. Those in the Explanation condition were asked to explain why they chose certain interventions whereas those in the Report condition were simply asked to report their choices. Meng et al.'s (2018) previous study served as our Baseline condition where children solved the same puzzles without being prompted. Using hierarchical Bayesian models developed by Coenen et al. (2015), we captured children's intervention strategy mainly in terms of how much they relied on the normative strategy, which is maximizing the expected information gain (EIG) of their chosen interventions, and the suboptimal positive test strategy (PTS). Children in all conditions relied more on PTS than EIG; there was no difference across conditions. However, compared to those in the Baseline condition who performed at chance, children in both the Explanation and the Report conditions were more accurate at identifying the correct structures after interventions. Crucially, children's intervention choices and strategies only predicted their accuracy at inferring the true causal structures in the Explanation and the Report conditions but not in the Baseline condition.

Taken together, our findings suggest that while engaging in self-explaining did not help children select more informative interventions, asking them to *think* about their intervention choices (explaining or reporting) might help them better utilize interventional data that were already generated.

Revisiting the self-explaining effect

The major motivation behind this study was the plethora of self-explaining effects in education (Fonseca & Chi, 2011) and cognitive development (Lombrozo, 2016). Given what we found, two questions stood out: Why did self-explaining have no effect on intervention selection? Why was the improvement on causal inferences not unique in explainers?

Further investigation is needed to provide precise answers. Here we offer some speculations. Explaining an intervention is not an easy feat: Not only do you need to contrast the value of your intervention with that of other interventions, but more fundamentally, you need to contrast your strategy of evaluating interventions with other strategies. The cognitive process of generating a good explanation may be too challenging for 5- to 7-year-olds given their limited working memory capacity, knowledge about causal systems and experiments, and metacognitive skills (Horne, Muradoglu, & Cimpian, 2019). A recent study (Ruggeri, Xu, & Lombrozo, in press) suggested that the quality of explanations might matter after all. In their study, 4- to 7-year-olds were asked to explain phenomena in a domain before playing Twenty Questions in that domain; the accuracy of explanations was correlated with the efficiency of question-asking. Since reasonable explanations may be more difficult to generate in our study than in past studies (Walker et al., 2014, 2017), it might limit the benefit children can reap from self-explaining. Regarding the second question, it might be that when asked to reflect on (i.e., explaining or reporting) their intervention choices, children became aware that their interventions played an important role for solving puzzles later and therefore paid closer attention to the intervention outcomes when making causal inferences.

Future directions

Given the importance of intervention selection in causal learning, we seek to explore more effective scaffolding methods in the future. To begin, we can provide feedback after each intervention. A recent study (Liquin & Lombrozo, 2017) found that explaining had greater effects when evidence contradicted what learners' beliefs. Another way to strengthen the scaffolding may be asking children to explain why *each* possible intervention may or may not be useful, rather than just their chosen intervention. Since belief updating is inherently linked to intervention selection (Coenen & Gureckis, 2015), we may help children choose more informative interventions by correcting errors in their belief updating process.

Conclusion

Rather than passively absorbing correlations and crunching numbers, active learners generate explanations and design interventions to learn about causality. Our study is among the first to bridge "thinking" and "doing" in causal learning. While self-explaining did not show benefits of improving children's intervention strategy, prompting children to think about their intervention choices in some way (explaining or reporting) may help them better utilize interventional data generated by themselves to infer unknown causal structures.

References

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2014). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182.

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477.

Coenen, A., & Gureckis, T. M. (2015). Are biases when making causal interventions related to biases in belief updating? In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 411–416). Austin, TX: Cognitive Science Society.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2018). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, 1–41.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Edwards, B. J., Williams, J. J., Gentner, D., & Lombrozo, T. (2019). Explanation recruits comparison in a category-learning task. *Cognition*, *185*, 21–38.

Fonseca, B. A., & Chi, M. T. (2011). Instruction based on self-explanation. *Handbook of research on learning and instruction*, 296–321.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Horne, Z., Muradoglu, M., & Cimpian, A. (2019). Explanation as a cognitive process. *Trends in Cognitive Sciences*, *3*(23), 187–199.

Klayman, J., & Ha, Y.-W. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 596–604.

Liquin, E. G., & Lombrozo, T. (2017). Explain, explore, exploit: Effects of explanation on information search. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2598–2603). Austin, TX: Cognitive Science Society.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748–759.

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

McCormack, T., Bramley, N. R., Frosch, C., Patrick, F., & Lagnado, D. A. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, *141*, 1–22.

Meng, Y., Bramley, N. R., & Xu, F. (2018). Children's causal interventions combine discrimination and confirmation. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 762–767). Austin, TX: Cognitive Science Society.

Nelson, J. D. (2005). Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979–999.

Pearl, J. (2000). *Causality*. New York: Oxford University Press.

Plummer, M. (2003). A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing.*

Ruggeri, A., Xu, F., & Lombrozo, T. (in press). Effects of explanation on children's question asking. *Cognition*.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Walker, C. M., & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition*, *167*, 266–281.

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*(2), 343–357.

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child Development*, *88*(1), 229–246.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology*, *66*(1), 55–84.