Explanation Supports Hypothesis Generation in Learning

Erik Brockbank (ebrockbank@ucsd.edu)
Department of Psychology, 9500 Gilman Dr.
La Jolla, CA 92093-109 USA

Caren Walker (carenwalker@ucsd.edu)
Department of Psychology, 9500 Gilman Dr.
La Jolla, CA 92093-109 USA

Abstract

A large body of research has shown that engaging in explanation improves learning across a range of tasks. The act of explaining has been proposed to draw attention and cognitive resources toward evidence that will support a good explanation—information that is broad, abstract, and consistent with prior knowledge—which in turn aids discovery and generalization. However, it remains unclear whether explanation acts on the learning process via improved hypothesis *generation*, increasing the probability that the correct hypothesis is considered in the first place, or hypothesis *evaluation*, the appraisal of the correct hypothesis in light of evidence. In the present study, we address this question by separating the hypothesis generation and evaluation processes in a novel category learning task and quantifying the effect of explanation on each process independently. We find that explanation supports the generation of broad and abstract hypotheses but has less effect on the evaluation of hypotheses.

Keywords: explanation; learning; inference; hypothesis generation; hypothesis evaluation

Introduction

Though every student knows the fear of being asked to explain their answer in front of the class, the benefits of explanation for learning have been shown across a broad range of domains. In particular, our unique abilities to reason abstractly and to infer causal structure in the world are heavily influenced by the process of seeking and generating explanations for our observations (Lombrozo, 2016). Children as young as three years of age are more likely to generalize on the basis of causal properties rather than salient perceptual features when prompted to explain (Walker, Lombrozo, Legare, & Gopnik, 2014; Legare & Lombrozo, 2014), five and six year-olds are better able to abstract the moral of a story when they explain key events (Walker & Lombrozo, 2017), and adolescents learning biology concepts constructed better mental models and showed improved abstraction when they self-explained (Chi, De Leeuw, Chiu, & LaVancher, 1994).

Explanation also appears to support learning among adults. Williams and Lombrozo (2013) showed participants a set of eight novel robots that each belonged to a particular category and varied along a number of perceptual dimensions (e.g., body shape, color). Participants who were asked to explain *why* the robots in an initial learning phase belonged to their respective categories were more likely to infer the correct categorization rule and apply it to a set of new robots than control participants (Williams & Lombrozo, 2013).

How might explanation benefit learning across such a broad range of domains? Prior research has proposed that learners who are prompted to explain will privilege hypotheses that support "good explanations", focusing on simplicity, breadth, and consistency with prior knowledge (e.g., Bonawitz & Lombrozo, 2012; Lombrozo, 2016; Walker, Bonawitz & Lombrozo 2017). In other words, explaining has been suggested to act like a bottleneck, constraining the set of hypotheses to favor those that exhibit *explanatory virtues*. This supports better causal inference and abstract reasoning in some domains, but can also make learners less attentive to conflicting evidence, biasing them too strongly towards abstract and generalizable hypotheses (Williams & Lombrozo, 2013; Williams, Lombrozo & Rehder, 2013; Walker, Bonawitz & Lombrozo, 2017). Broadly, this research suggests that a learner's goals play a selective role in the learning process by influencing which hypotheses they are most likely to privilege (Williams & Lombrozo, 2010).

If explanation supports learning by influencing the hypotheses that learners endorse, this raises a question: Does explanation modify the set of hypotheses that learners initially entertain, or does it change how learners appraise the hypotheses they are considering? In other words, does explanation facilitate reasoning and abstraction via hypothesis *generation* or hypothesis *evaluation*?

A growing body of research points to the role that environmental and contextual factors play in determining which hypotheses are generated during a particular task, including working memory capacity, perceived likelihood,

the number of alternatives available, and the design of the learning environment, among others (e.g., Dougherty & Hunter, 2003; Koehler, 1994; Walker, Rett, & Bonawitz, 2020). If the effectiveness of explanation in learning lies in directing attention and cognitive resources to hypotheses that are consistent with explanatory virtues, we might expect this to predominantly impact the process of hypothesis generation. However, there is no prior research indicating that explanation might not *also* affect hypothesis evaluation, by, for example, causing learners to over-weight the evidence they observe in favor of a particular type of hypothesis.

In the current study, we examine the role of explanation in hypothesis generation and evaluation by modifying existing methods that have been developed to pull these interrelated cognitive processes apart. Specifically, Bonawitz & Griffiths (2010) show that when participants are given a simple prime before performing a rule learning task, the prime impacts the proportion of participants who correctly infer the rule but does *not* impact how likely participants rate the correct rule to be. In this way, priming can be interpreted as constraining hypothesis generation, but not evaluation. Building on these results, we hypothesize a similar effect of explanation during learning.

To test this, we presented participants with a category learning task which requires them to generate and evaluate hypotheses about which kinds of fishing lures are most likely to catch fish. Participants in an explanation condition were prompted to explain the evidence they saw during the task, while control participants simply described it. All participants were then presented with a hypothesis generation task drawing on the "rule report" method used in prior research on explanation (Williams & Lombrozo, 2010), followed by a separate hypothesis evaluation task modeled after Bonawitz & Griffiths (2010). Across these separable hypothesis generation and evaluation prompts, we compare the performance of explanation with the control description condition to understand the role that explanation plays in both hypothesis generation and evaluation processes. By combining prior research on explanation with investigations of hypothesis generation and evaluation, we aim to provide a more precise description of the impact that explanation has on the learning process.

In line with existing theoretical proposals, we predict that explanation will support learners' generation of broad and abstract hypotheses in this task. It is conceivable that explanation might also influence how learners evaluate hypotheses, but our predictions were conservative with respect to this possibility. In order to investigate how explanation impacts hypothesis evaluation, participants evaluated both an abstract "target" rule which was consistent with 100% of the evidence they had seen, as well as a "distractor" rule, which was also consistent with all of the evidence observed but was considerably more complex. These rules appear in Table 1. If explanation influences

learners' evaluation of candidate hypotheses, we predict that explainers may be more likely to privilege the abstract target rule that better reflects explanatory virtues.

Participants

Participants were 86 undergraduate students from a large West Coast university who received course credit for their participation. Participants were randomly assigned to either *explain* or *describe* conditions.

Procedure

Participants completed the experiment in a web browser on laboratory computers[1]. All participants were given instructions indicating that they would see a number of different fishing lure combinations, and that their task was to determine which combinations were most likely to catch fish. The fishing lures used throughout the experiment were composed of two stacked shapes: one smaller shape on the bottom of the fishing lure and one larger shape on the top. Each of the top and bottom shapes belonged to one of six possible shapes, three of which were *rounded* (circle, oval, and teardrop) and three of which were *pointy* (triangle, diamond, and four-pointed star). Each shape in the fishing lure combination had one of four possible colors: red, blue, green or yellow. In addition, each shape either did or did not have a purple dot. The fishing lure combinations that caught fish were determined by the following rule: *lures with pointy shapes on the bottom catch fish*. This rule was chosen based on prior work suggesting that explainers are more likely to propose more abstract rules (i.e., *pointy*, rather than triangle, diamond, or star) (Williams & Lombrozo, 2010), and rules that are more consistent with prior mechanistic knowledge (i.e., pointed objects are used to catch fish) (Williams & Lombrozo, 2013). This was also confirmed in a pilot study of the present experiment.

The experiment itself was composed of a trial phase, a hypothesis generation phase, a hypothesis evaluation phase, and a memory check.

Trial Phase In the trial phase, participants observed eight fishing lure trials, each consisting of an evidence component, a description or explanation component, and a prediction component. These are illustrated in Figure 1.

In the evidence component of each trial, participants were shown a novel fishing lure combination and told whether or not this combination successfully caught a fish. In the subsequent explanation or description component, participants in the *explain* condition were prompted to provide a written explanation for the evidence they had just seen ("*Explain why your friend might have {not have} caught a {any} fish with this lure combination*"), while in the control *describe* condition, participants were simply asked to describe the evidence they had just seen ("*Describe this lure*
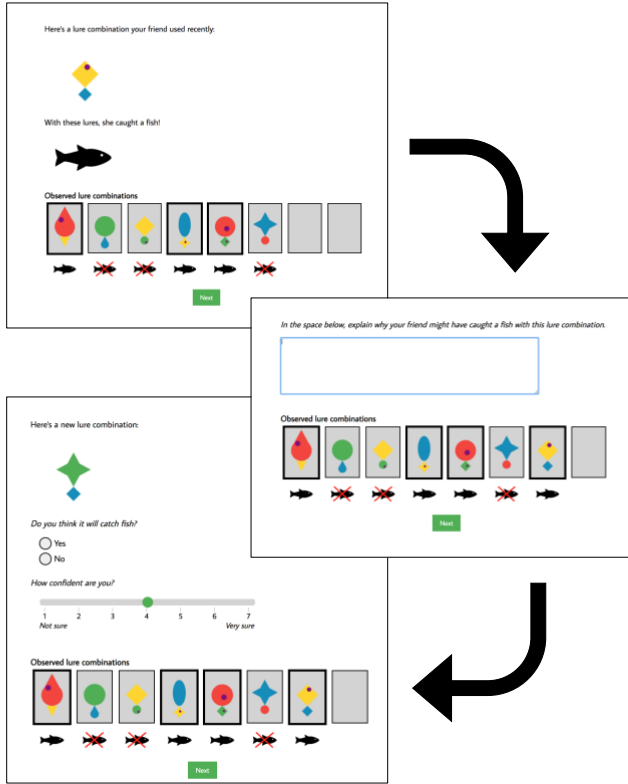
Figure 1: A sample trial seen during the trial phase of the experiment. Top, a sample *evidence* component in which participants see a lure combination that does or does not catch fish. Middle, a sample *response* component for participants in the *explain* condition. Control participants saw a similar prompt asking them to describe the lure combination. Bottom, a sample *prediction* component for a new lure similar to the previous evidence.

combination that your friend caught a fish {didn't catch a fish} with"). This was the only difference between conditions. In the prediction component of each trial, participants were shown a novel fishing lure combination which retained only one of the elements of the earlier lure combination they observed and asked to indicate whether they thought this new combination would catch fish or not. Participants were not given feedback about their predictions.

Accumulated evidence from previous trials remained visible at the bottom of the screen throughout all subsequent trials to help participants recall which fishing lure combinations did and did not catch fish. The evidence and prediction segments of the trials included four fishing lure combinations that did catch fish and four that did not. The fishing lures chosen and the order they appeared in were identical across participants. The decision to present fishing lures in the same order for all participants allowed for tight control over when participants saw each negative exemplar, and therefore when various hypotheses could be ruled out by the evidence. Any order effects resulting from the presentation of evidence would therefore impact participants in both conditions equally.
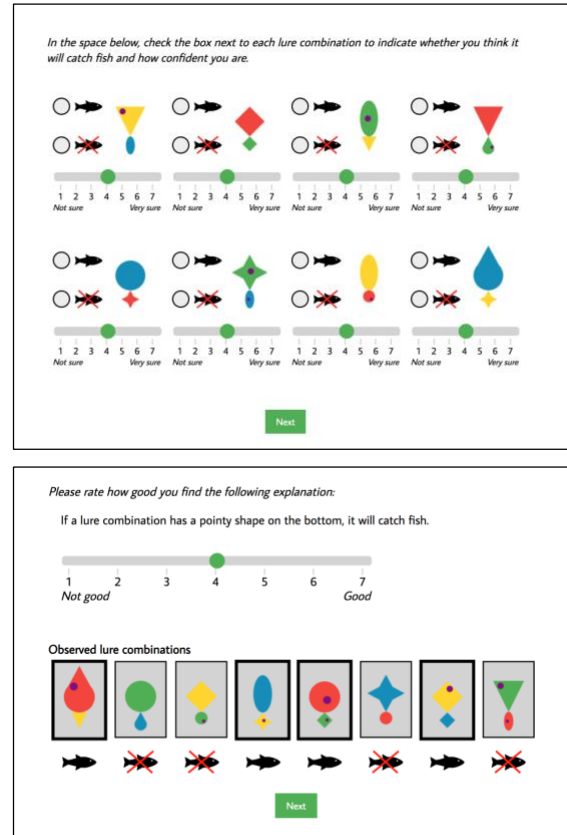

Figure 2: Top, the hypothesis generation classification task used to test whether participants had inferred the correct rule for categorizing fishing lures. At bottom, the hypothesis evaluation task for a sample rule.

Hypothesis Generation Phase After completing the eight trials, participants were tested on hypothesis generation. First, they were given a free response prompt to assess whether they had figured out the target rule: "*Describe the single best rule you used for deciding whether or not each lure combination will catch fish*". Next, they were shown a set of eight novel fishing lure combinations and asked to indicate whether each of these combinations would catch fish, along with a confidence rating from 1 to 7 (see Figure 2). During the hypothesis generation phase, the evidence from earlier trials was *not* available for reference; this ensured that the rules participants provided were generated during the trial phase, rather than by careful study during the generation phase.

Participants' free responses were coded as either correct or incorrect, depending on whether they were able to provide a rule which was consistent with 100% of the evidence and would allow them to successfully classify a novel fishing lure combination. By this criterion, participants who were explicit about the shapes that caught fish (noting the triangle, diamond, and star), but did not refer to them as "pointy," were still coded as correct. Responses that provided insufficient evidence that the learner generated the correct rule (e.g., *"I used the lure's shape"*) were coded as incorrect. A second

researcher who was blind to condition coded the responses for reliability. Agreement was 99%. Though it was possible to have come up with a rule—other than the target rule—which was consistent with all of the evidence (e.g., the distractor rule), no participant provided such a rule.

Hypothesis Evaluation Phase Next, participants were tested on hypothesis evaluation. Participants were shown a series of six possible rules representing candidate hypotheses about which types of fishing lure combinations catch fish. Of the six rules, the target rule and a distractor rule were both consistent with 100% of the evidence. Three additional miscellaneous rules were consistent with five or six pieces of evidence, and one additional rule suggested that it was randomly determined which fishing lure combinations caught fish. Table 1 lists each of the rules that participants evaluated. The target rule was chosen to highlight the sort of abstract reasoning recruited during explanation, while the distractor rule, while also consistent with all of the evidence, was significantly more complex. Participants were asked to rate the strength of each rule as an explanation of the evidence on a 1 to 7 scale (see Figure 2). During this phase, participants were again provided with a visual reminder of the outcome of each of the eight trials at the bottom of the screen. Consistent with prior work (Bonawitz & Griffiths, 2010), this was done to assess participants' appraisal of each hypothesis in light of the evidence.

Table 1: Rules used during evaluation.

Rule	Category	Evidence
If a lure combination has a red shape or a blue shape, it will catch fish	Misc.	62.5%
If a lure combination has a diamond, it will catch fish	Misc.	62.5%
If a lure combination has a pointy shape on the bottom, it will catch fish	Target	100%
There is no pattern to which lure combinations catch fish: the results are random, but there are approximately equal numbers that catch fish and don't	Random	NA
If a lure combination has a yellow shape or a diamond on the bottom, it will catch fish	Distractor	100%
If a lure combination has a purple dot on at least one of the lures, it will catch fish	Misc.	75%

Memory Check Finally, participants were given a memory probe in which they were shown a set of eight fishing lure combinations, including four novel fishing lure combinations and four that had previously appeared during the training phase. Participants were prompted to indicate whether they had seen each fishing lure combination at any point during the experiment.

Results

To understand the role of explanation on hypothesis generation and evaluation, we compare the *explain* and control conditions on the hypotheses they generate, their accuracy at classifying novel fishing lure combinations based on these hypotheses, and their subsequent evaluation of candidate hypotheses about which combinations catch fish. Our results provide evidence that explanation supports hypothesis generation while having a minimal effect on hypothesis evaluation.

Hypothesis Generation

We first examine the effect of explanation on hypothesis generation. In line with our hypothesis, a significantly greater proportion of participants in the *explain* condition provided the correct hypothesis in their free response: 51.2% in the *explain* condition versus 18.6% in the control ($\chi^2 (N = 86, 1) = 8.65, p = 0.003$). Consistent with their performance on the free response item, participants in the *explain* condition were also better able to classify novel fishing lure combinations in the judgment task, compared with controls (Mean accuracy: *explain*: 0.81; *control*: 0.65, $t(84) = -2.84, p = 0.006$). Figure 3 shows accuracy on the judgment task by condition. Finally, if we treat people who had 100% accuracy on the judgment task as most likely to have generated the correct rule (whether or not they were able to articulate it in their free response), this proportion is significantly greater in the *explain* condition (*explain*: 0.54; *control*: 0.30; $\chi^2 (N = 86, 1) = 3.87, p = 0.049$). These results are robust to lower cutoff scores
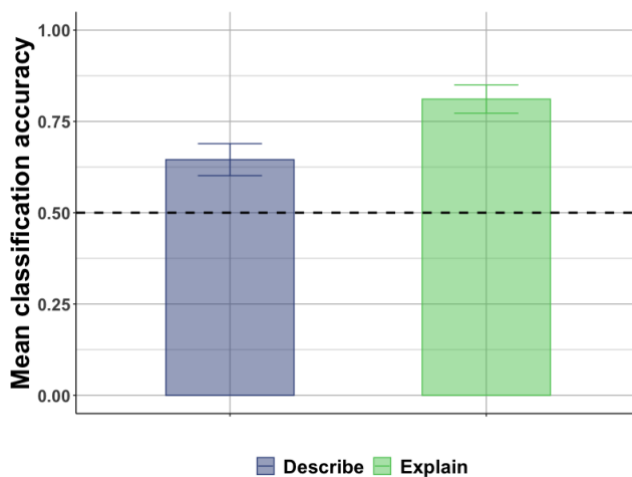

Figure 3: Accuracy on classification task by condition. Error bars indicate one SEM.

including 7/8 and 6/8 correct. In sum, both the free response and classification measures indicate that participants in the *explain* condition were more likely to produce and apply the target hypothesis, suggesting that explanation plays a role in hypothesis generation.

Hypothesis Evaluation

Figure 4 shows evaluation ratings for the target rule, the distractor rule, and the average ratings across all remaining rules. The difference between conditions in target rule rating is significant, suggesting that explanation may also play a role in hypothesis evaluation (Mean ratings: *explain*: 6.6; *control*: 6.09, t (84) = -2.04, p = 0.045). However, both conditions provide target rule ratings that are close to ceiling and evaluate the set of all other rules, including the distractor, rule to be far worse explanations (Wilcoxon Signed-Rank: *explain*: z = -5.58, $p < 0.001$; *control*: z = -5.41, $p < 0.001$).

To further investigate effects of explanation on hypothesis evaluation, we examine the evaluation ratings of the distractor rule across conditions and relative to the target rule. We might expect participants in the *explain* condition to find the distractor hypothesis less appealing than *controls* since this rule lacked the explanatory virtues of both simplicity and breadth, which they may be more attuned to. However, we find no significant difference between conditions in their average rating (Mean ratings: *explain*: 4.12; *control*: 4.74, t (84) = 1.53, p = 0.13). Notably, all participants *did* rate this rule significantly lower than the target rule, despite the fact that both rules were consistent with all the evidence observed (*explain*: t (84) = -7.54, $p < 0.001$; *control*: t (84) = -3.85, $p < 0.001$). This likely reflects a general prior preference for explanations which are not only consistent with the evidence, but also simple and easily generalizable (Williams, Lombrozo & Rehder, 2013). Further, in a comparison of mean ratings of the target and distractor rule for participants in each condition, the interaction between condition and rule type is significant, suggesting that the differential between target and distractor rules was different for participants in the *explain* condition than control participants (F(1, 82) = 5.6, p = 0.02). This suggests that there may be an effect of explanation in producing a greater disparity between target and distractor rules.

Finally, to ensure that these results do not depend on whether participants successfully generated the target hypothesis, we look at evaluation ratings of the target rule among participants who did *not* generate the target hypothesis. We find no difference in rule ratings across conditions (Mean ratings: *explain*: 6.24; *control*: 5.91, t (54) = -0.86, p = 0.39). More importantly, evaluation ratings for the target rule remain significantly higher than those for all other rules including the distractor (Wilcoxon Signed-Rank: *explain*: z = -3.83, $p < 0.001$; *control*: z = -4.75, $p < 0.001$). In line with the results for all participants, among participants who did not generate the target hypothesis, there is no significant difference between conditions in ratings of the distractor rule (Mean ratings: *explain*: 4.62; *control*: 5.17, t (54) = 1.14, p = 0.26). Finally, participants in both conditions
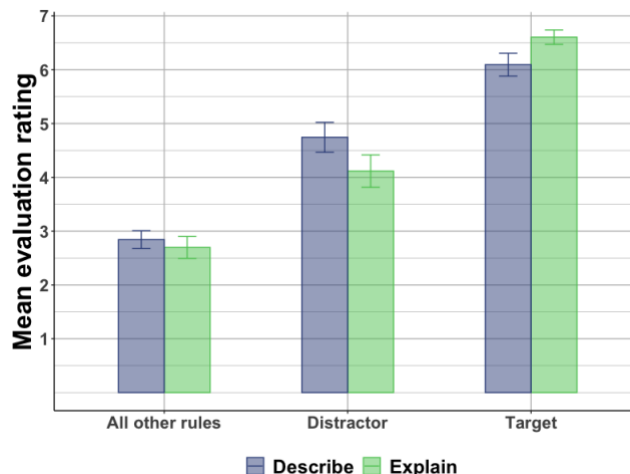

Figure 4: Evaluation ratings for target rule, distractor, and all other rules by condition. Error bars indicate one SEM.

rate the target rule significantly higher than distractor rule (*explain*: t (40) = -3.24, p = 0.002; *control*: t (68) = -2.02, p = 0.048). This effect appears to be larger for explain participants, as with the full participant set; however, in this case the interaction between rule type and condition is not significant (F(1, 52) = 2.04, p = 0.16). In sum, we find evidence that is broadly consistent with the hypothesis that participants across both conditions evaluate the target rule based on the available evidence, as well as its generality.

Memory

Results of the hypothesis generation and evaluation tasks strongly support the proposal that explanation facilitates generation of broad and abstract hypotheses. To assess whether the observed effects of explanation on hypothesis generation could be due to a general increase in attention or engagement, we looked for condition differences in memory for fishing lure combinations. Participants in both conditions performed above chance in their overall memory accuracy (*explain*: t (42) = 5.17, $p < 0.001$; *control*: t (42) = 3.67, $p < 0.001$), with no evidence of higher accuracy for participants in the *explain* condition (Mean accuracy: *explain*: 0.61; *control*: 0.59, t (84) = -0.61, p = 0.55). This suggests that the benefits of explanation on hypothesis generation cannot be reduced to a general increase in attention or depth of processing. Collapsing across conditions, participants who successfully generated the target hypothesis did not perform better on the memory task than those who did not (t (84) = 0.32, p = 0.75). This further suggests that effects of effort or processing cannot account for accuracy in hypothesis generation.

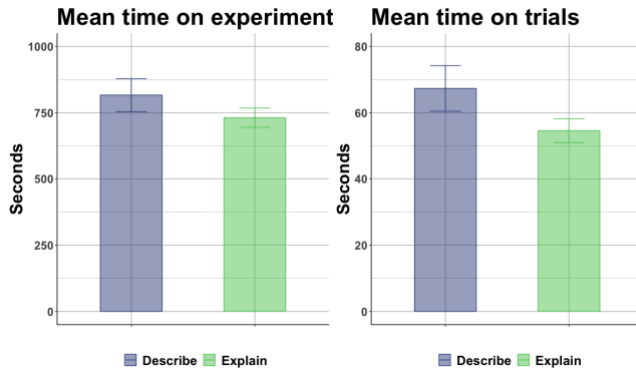Figure 5: Mean experiment completion time and mean trial completion time by condition. Error bars indicate one SEM.

Time on Task

To understand whether performance on the hypothesis generation tasks could simply be attributed to time spent on the task, we examined time spent on the experiment as a whole, as well as time spent on individual trials within the evidence phase. Figure 5 shows overall time on the experiment across conditions. Participants in the *explain* condition did not spend any longer on the task than control participants; in fact, though it may appear that control participants spent *longer*, this difference is not significant (Mean time on task: *explain*: 731s; *control*: 816s, $t(84) = 1.18$, $p = 0.24$).

Looking exclusively at time spent on the evidence trials themselves (i.e., the segment of the experiment which differentiates the two conditions), the data are similar to the overall experiment time: control participants appear to spend slightly longer describing the evidence than explainers spent explaining it, but this difference is not significant (Mean time on trials: *explain*: 54.6s; *control*: 67.3s, $t(67) = 1.63$, $p = 0.11$) (see right side of Figure 5)2. Further, participants across conditions who generated the target hypothesis did not spend longer on the task than those who failed to generate the target hypothesis, ($t(84) = 0.99$, $p = 0.33$), and did not spend longer on evidence trials, ($t(67) = 0.93$, $p = 0.35$). This further suggests that time spent on the task is not accounting for the accuracy in hypothesis generation at the outset.

Discussion

In this experiment, we investigated the role explanation plays in the process of hypothesis generation and evaluation. Using a category learning task with novel stimuli, we find that participants who are prompted to provide explanations for the evidence they observe are more likely to generate the (correct) abstract rule for category membership than control participants who were simply asked to describe the evidence.

This suggests that the benefits that explanation confers on learning arise initially by constraining the set of hypotheses generated by the learner. By comparison, the effect of explanation on hypothesis evaluation is less clear. Participants in both conditions rated the target rule significantly higher than all other rules, including the distractor rule which was consistent with 100% of the evidence. In other words, the pattern of results was roughly similar for participants in both *explain* and *control* conditions. However, we also find a significant difference between the two conditions in their rating of the target rule; since these ratings were near ceiling in both cases, this may obscure a larger difference between the two groups in how they evaluate hypotheses.

The results of the memory probe and time analysis allow us to rule out a variety of alternative explanations for the observed condition differences in hypothesis generation. First, it's possible that participants in the *explain* condition simply tried harder. Generating explanations for the evidence observed during the eight trials is undoubtedly more challenging than simply describing that same evidence. The increased effort and attention in this condition, rather than the act of generating explanations, could have plausibly accounted for these results. However, if this were the case, we would expect participants in the *explain* condition to have better memory for the fishing lure combinations they did or did not see at any point in the experiment. The results from the memory probe do not support this explanation.

Second, if participants in the *explain* condition had spent more time on the task, this longer exposure to the stimuli and additional time spent generating responses might have made them more likely to generate the target hypothesis. However, the time on task results also rule out this alternative account. In short, the lack of differences in the amount of time spent on the task or memory for task stimuli rule out alternative proposals that the process of explaining increases overall attention or engagement (e.g., Siegler, 2002).

Though the hypothesis generation results were consistent with our initial predictions, the modest impact of explanation on hypothesis evaluation deserves further attention. One possibility which might account for these results relates to the overlapping demands of generating explanations with hypothesis generation and evaluation. In particular, the prompts to explain and the hypothesis generation task may themselves involve some amount of tacit hypothesis evaluation. Further, the evaluation task may rely to some degree on explanatory reasoning, such that when participants were prompted to explicitly evaluate the provided rules, differences between the two conditions were more difficult to discern. A second possibility (noted above) is that participants' evaluations of the target rule were close to ceiling in both conditions. This was presumably due to the fact that participants were able to reference the evidence while evaluating hypotheses. Future work might address

---

2 Trial completion time was not recorded for the first 17 participants (20% of the total sample), balanced across conditions (8 control, 9 *explain*).

these concerns in several ways. First, the evaluation task could be modified to remove the evidence from view, increasing memory demands. Given that all participants performed similarly on the memory probe, any effects observed on this revised evaluation task should be due to condition differences. Second, the current study could be replicated without including the generation task, thus limiting the amount of evaluation that occurs prior to the explicit evaluation prompt. Finally, future work will consider the qualitative content of the explanations and descriptions provided during the evidence phase to shed light on the specific hypotheses considered during learning.

In sum, we find strong initial evidence that explanation primarily supports learning via improved hypothesis generation. This work opens the door for further research defining the precise cognitive mechanisms underlying the effects of explanation in learning and inference, as well as broader questions about how learners generate useful hypotheses about the world.

References

Bonawitz, E. B., & Griffiths, T. L. (2010). Deconfounding hypothesis generation and evaluation in Bayesian models. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 32, No. 32).

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental psychology*, *48*(4), 1156.

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

Dougherty, M. R., & Hunter, J. E. (2003). Hypothesis generation, probability judgment, and individual differences in working memory capacity. Acta psychologica, 113(3), 263-282.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(2), 461.

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of experimental child psychology*, *126*, 198-212.

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

Siegler, R. S. (2002). Microgenetic studies of self-explanation. Microdevelopment: Transition processes in development and learning, 31-58.

Walker, C. M., Bonawitz, E., & Lombrozo, T. (2017). Effects of explaining on children's preference for simpler hypotheses. *Psychonomic bulletin & review*, *24*(5), 1538-1547.

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., Rett, A., & Bonawitz, E. (2020). Design drives discovery in causal learning. *Psychological Science*, 0956797619898134.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive science*, *34*(5), 776-806.

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

Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, *142*(4), 1006.