Explanation Impacts Hypothesis Generation, but not Evaluation, During Learning

Erik Brockbank & Caren M. Walker

Department of Psychology, University of California, San Diego

In press at **Cognition**

Abstract

A large body of research has shown that engaging in self-explanation improves learning across a range of tasks. It has been proposed that the act of explaining draws attention and cognitive resources towards evidence that supports good explanations—information that is broad, abstract, and consistent with prior knowledge—which in turn aids discovery and promotes generalization. However, it remains unclear whether explanation impacts the learning process via improved hypothesis generation, increasing the probability that the most generalizable hypotheses are considered in the first place, or hypothesis evaluation, the appraisal of such hypotheses in light of observed evidence. In two experiments with adults, we address this question by separating hypothesis generation and evaluation in a novel category learning task and quantifying the effect of explaining on each process independently. We find that explanation supports learners' generation of broad and abstract hypotheses but does not impact their evaluation of them. These results provide a more precise account of the process by which explanation impacts learning and offer additional support for the claim that hypothesis generation and evaluation play distinct roles in problem solving.

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

Explanation Impacts Hypothesis Generation, but not Evaluation, During Learning

Though every student knows the fear of being asked to explain their answer in front of the class, the benefits of explaining for learning have been shown across a broad range of tasks and knowledge domains. These effects have also been observed across the lifespan: Children as young as three years of age are more likely to generalize on the basis of causal properties over salient perceptual features when prompted to explain (Walker et al., 2014; Legare & Lombrozo, 2014), five- and six-year-olds are better able to abstract the moral of a story when they are asked to explain key events (Walker & Lombrozo, 2017), and adolescents learning biology concepts construct better mental models and show improved abstraction when they self-explain during study (Chi et al., 1994). Explaining also supports learning among adults. In a category learning task with unfamiliar stimuli, adults prompted to explain why the evidence they observed was consistent with category distinctions were more likely to discover the underlying rule for category discrimination (Williams & Lombrozo, 2010; Williams & Lombrozo, 2013).

Why might explaining benefit learning across such a broad range of ages and domains? Some researchers have proposed that learners who are prompted to explain tend to privilege hypotheses that support "good explanations," focusing on simplicity, breadth, and consistency with prior knowledge (e.g., Bonawitz & Lombrozo, 2012; Lombrozo, 2016; Walker et al., 2017). In other words, the act of explaining may help guide learners towards hypotheses that best exhibit these *explanatory virtues* (Lipton, 2008; Walker et al., 2014). While this typically supports causal inference and abstract reasoning, in some contexts, explaining makes learners less attentive to counterevidence, biasing them too strongly in favor of broad generalizations and alignment with existing knowledge (e.g., Engle & Walker, 2021; Kuhn & Katz, 2009; Williams & Lombrozo, 2013; Williams et al., 2013; Walker et al., 2016). In other contexts, when the

available counterevidence is sufficiently strong, explaining can facilitate belief revision (Macris & Sobel, 2017; Walker et al., 2016) and encourage exploratory behavior aimed at forming new hypotheses (Legare, 2012). Taken together, this literature suggests that explaining prompts learners to pursue hypotheses that have the *broadest scope*, incorporating both their prior knowledge and their current observations (Walker et al., 2016; Williams & Lombrozo, 2013). Across studies, the act of constructing an explanation plays a selective role in the learning process by influencing which solutions the learner is most likely to entertain (Legare et al., 2010; Williams & Lombrozo, 2010; Schulz, 2012).

This ability to privilege certain kinds of hypotheses over others is central to commonsense reasoning. Despite the infinite space of solutions for everyday problems, learners tend to restrict their responses to those that best fit (Schulz, 2012, Lake et al., 2017). However, it remains unclear how people do this so effectively. Here, we explore this question by examining the particular effect of explanation on learning: How does the act of explaining lead learners to select certain hypotheses over others? If explaining ultimately supports learning by influencing the solutions that learners endorse, does it modify the set of hypotheses that they initially entertain, or does it change how they appraise the hypotheses under consideration? In other words, does explaining facilitate reasoning via hypothesis *generation* or hypothesis *evaluation*?

Generating hypotheses in novel situations is a central challenge for learners engaged in inductive inference (Kuhn, 1989; Mehle, 1982; Weber et al., 1993). Early theories of hypothesis generation proposed structured search processes in long-term memory (Gettys & Fisher, 1979) and stressed the foundational role of drawing analogies to other domains (Gick & Holyoak, 1980; Gentner, 1983). In settings that require generating hypotheses about decontextualized or unfamiliar stimuli, researchers have observed biases and limitations in the generation process,

such as the tendency to narrow existing hypotheses rather than generate them anew (Klayman & Ha, 1989; Goodman et al., 2008). In fact, hypothesis generation is often highly dependent on contextual factors. For example, early research exploring the role of schemas in problem solving found that, when faced with logically equivalent problems, people produce strikingly different hypotheses depending on the semantic content of those problems (e.g., evaluating the logical implications of $p \rightarrow q$ using rules about the legal drinking age vs. using abstract letter and number associations, Griggs & Cox, 1982; Cheng & Holyoak, 1985).

Indeed, a substantial body of subsequent work has provided evidence for the impact of environmental factors in determining which hypotheses are generated during a particular task, including working memory capacity, cognitive load, perceived likelihood, the number of alternatives available, and the design of the learning environment, among others (e.g., Dougherty & Hunter, 2003; Klein, 1993; Koehler, 1994; Schunn & Klahr, 1995; Walker et al., 2020). Recent results suggest that even young children are responsive to environmental features that restrict the hypothesis space, including how the data were sampled, who provided the evidence, and why (e.g., Bonawitz et al., 2011; Butler & Markman, 2012; Gergely et al., 2002; Walker et al., 2014). If the effectiveness of explanation for learning lies in directing the learner's attention and cognitive resources to hypotheses that are consistent with explanatory virtues, we might expect explaining to impact hypothesis generation in a similar way as other contextual changes.

However, it is possible that explaining might *also* affect hypothesis evaluation, by, for example, causing learners to overweight the data they observe in favor of hypotheses that are more consistent with explanatory virtues. Indeed, there is some evidence to support this possibility: Williams et al. (2013) found that when learners were given statistics problems in which the solutions violated their intuitions, participants who explained the evidence performed

better than controls, even when they had been provided with the correct procedure in advance. The authors conclude that since explanation still facilitates learning when participants had prior exposure to the rule, explaining must be helping them to *apply* this rule to the data they observe.

In the current study, we examine the role of explaining in hypothesis generation and evaluation by modifying methods used in prior work that was designed to pull these interrelated cognitive processes apart. Specifically, Bonawitz & Griffiths (2010) show that when participants were given a simple prime before performing a rule learning task, the prime impacted the proportion of participants who correctly inferred the rule but did *not* impact how likely participants rated the correct rule to be. In this way, priming can be interpreted as constraining hypothesis generation, but not evaluation. Building on these results, the current study asks whether there is a similar effect of engaging in explanation during learning.

The current study

To test this, we presented participants with a category learning task which required them to generate and evaluate hypotheses about which kinds of fishing lures were most likely to catch fish. All participants were presented with a series of events in which particular lure combinations did or did not catch fish. After each event, participants in an explanation condition were prompted to explain the evidence they observed, while control participants were asked to describe it. In Experiment 1, all participants were then presented with a hypothesis generation task, drawing on the "explicit report" method used in prior research on explanation (Williams & Lombrozo, 2010), as well as a related classification task to test their generalization. This was followed by a separate hypothesis evaluation task modeled after Bonawitz and Griffiths (2010). We then examined the effects of explanation on learning outcomes in each task.

Given prior findings that explanation recruits prior knowledge and encourages learners to search for abstract patterns, we designed the rule learning task so that it could be solved by capitalizing on these strategies. Specifically, each fishing lure was composed of two stacked shapes, and any lure combination with a triangle, diamond, or four-pointed star on the bottom would catch fish (see Figure 1). It is therefore possible to succeed on this task by attending to each lure's concrete features in pursuit of rule-like statistical patterns. Critically, however, this evidence was also consistent with an abstract rule: *lures with pointy shapes on the bottom catch fish*. This rule was chosen based on prior research suggesting that explainers are more likely to infer abstract hypotheses (i.e., *pointy*, rather than triangle, diamond, or star) (Williams & Lombrozo, 2010), and those that are more consistent with prior mechanistic knowledge (i.e., pointed objects are used to catch fish) (Williams & Lombrozo, 2013).[1]

We expected that participants would apply different cognitive strategies depending on whether they were prompted to explain or describe their observations. Specifically, while explainers may be more likely to recruit real world knowledge and search for broad patterns, describers may be more likely to attend to concrete features. Although both strategies can lead to success on the current task, the pursuit of an abstract rule is likely to increase the availability of the target hypothesis. Critically, our goal was *not* to demonstrate that explanation makes learners more likely to privilege this target hypothesis. We anticipated this outcome based on the prior work. Instead, by combining research on the effects of explanation during learning with investigations of hypothesis generation and evaluation, we aim to provide a more precise description of the expected impact that explaining has on the learning process.

[1] This was also confirmed in a pilot study.

Experiment 1

Participants

Participants were 86 undergraduate students at a major West Coast university who received course credit for their participation. Given that our study design was based on Bonawitz & Griffiths (2010), we conducted a power analysis using their free response results. This analysis suggests that 88 participants would be required to detect a similar effect size (Cohen's $w = 0.3$) as the priming intervention they report. For a closer comparison to the specific effects of explanation on category learning, we also analyzed the effect size reported in Williams & Lombrozo (2010), Experiment 1 (Cohen's $w = 0.33$, see Table 2 in Williams & Lombrozo, 2010). With 43 participants per condition, we had an estimated 86% power to detect a similar effect for hypothesis generation. Informed consent was obtained from all participants in accordance with the Institutional Review Board's approved protocol. Participants were randomly assigned to either *explain* or *describe* (control) conditions.

Procedure

Participants completed the experiment in a web browser on laboratory computers.[2] 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 (see Figure 1). Each of the top and bottom pieces were composed of one of six possible shapes, three of which were *rounded* (circle, oval, and teardrop) and three of which were *pointy* (triangle,

[2] All code for Experiment 1, as well as data and analysis code for the results presented, can be found at: `https://github.com/erik-brockbank/go_fish`.

diamond, and four-pointed star). Each shape in the fishing lure combination was one of four possible colors: red, blue, green, or yellow. In addition, each shape either did or did not have a purple dot. As noted previously, the fishing lure combinations that caught fish were determined by the following rule: *lures with pointy shapes on the bottom catch fish*.

The experiment 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 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 one of the elements of the earlier lure combination they observed. They were then asked to indicate whether they thought this new combination would catch fish or not. All participants saw the same prediction lure combinations on each trial, since these lures were designed to share a common element with

the previously presented lure combination (see Supplementary Materials for all prediction stimuli). Participants were not given feedback about their predictions.

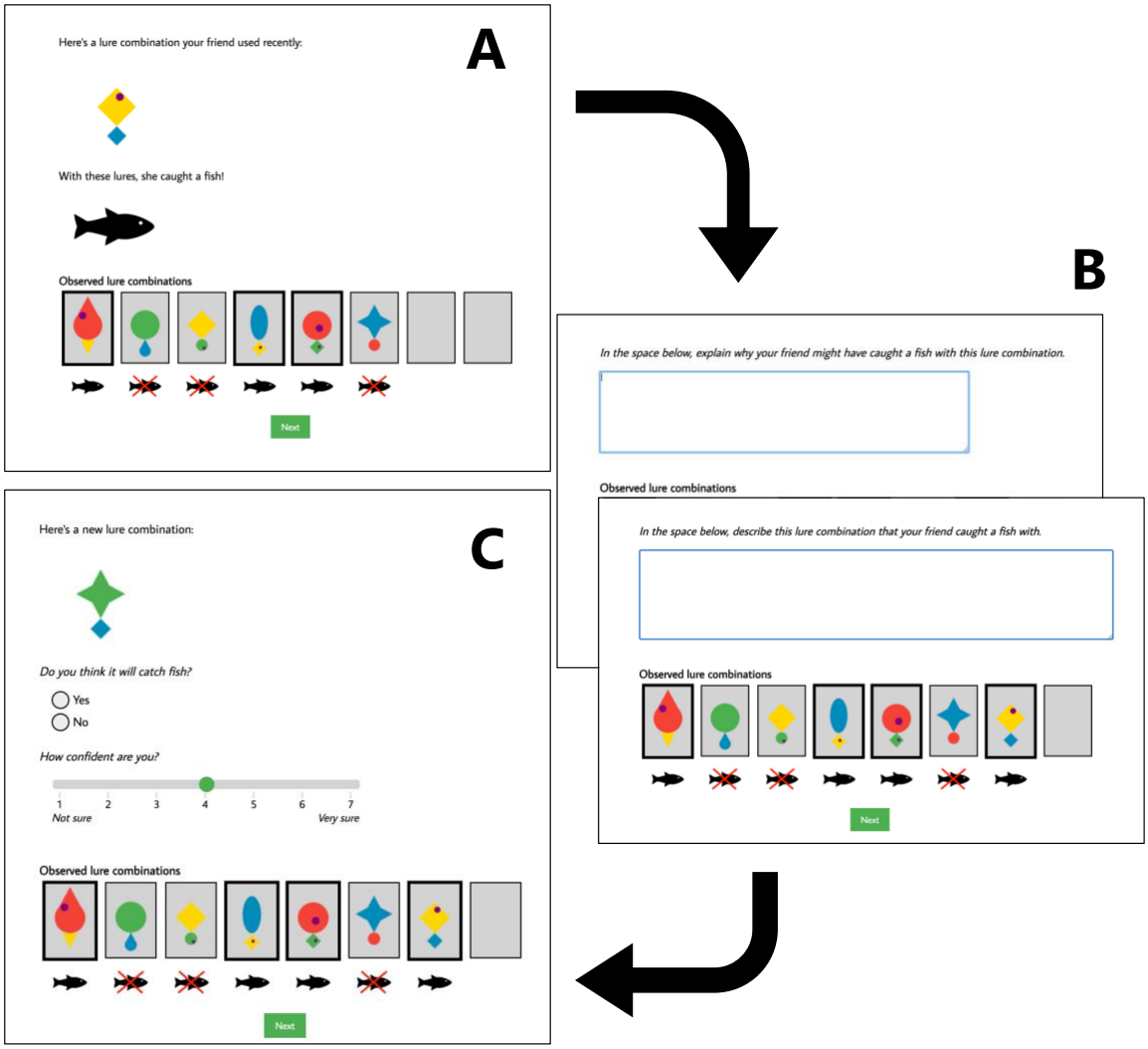

Figure 1. A sample trial from the trial phase of Experiments 1 and 2. (A) A sample evidence component in which participants see a lure combination that does or does not catch fish. (B) Response components for participants in the *explain* condition (top) and *describe* (control) condition (bottom). (C) A sample prediction component for a new lure.

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 components of the trials included four fishing lure combinations that did catch fish and four that did not. The fishing lures chosen and the order in which they appeared were identical across conditions (refer to Figures 1 and 2). 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.

Trial Response Coding. Following similar analyses used in past work (Williams & Lombrozo, 2010), each of the eight explanations and descriptions participants provided during the trial phase were coded for the number of *concrete* and *abstract* references to fishing lure features. Features of the fishing lure combinations were limited to their top and bottom shape, color, and the presence of a purple dot. A reference was coded as *concrete* if it was specific to that feature (e.g., "triangle," "yellow," "has a dot") and *abstract* if the reference was also true of other lures with different features (e.g., "pointy shape" refers to triangles, diamonds, and four-pointed stars, "bright color" refers to yellow and red, "has an eye" refers to lures with a dot) (Williams & Lombrozo, 2010). Critically, the number of abstract references do not by themselves indicate success on the task, since participants could detect the rule based on the use of an abstract strategy ("pointy shapes") *or* by attending to concrete patterns ("triangles, diamonds, or squares") (see below). However, examining the frequency of each type of reference *does* provide evidence for differences in participants' problem-solving approach.

In this vein, we compared the average number of abstract and concrete feature references made by participants in each condition to assess whether explainers generated a greater number of abstract hypotheses, relative to describers. We also coded explanations and descriptions for the number of references to an underlying *mechanism* to explain the data (e.g., *"Perhaps this teardrop piece is too round for the fish to latch onto", "Because the bottom blue part of the lure blended in with the water"*). In line with prior work, we predicted that explainers would be more likely to draw on prior knowledge and provide mechanism-based responses (Williams & Lombrozo, 2013). Again, although we anticipated that this tendency was likely to be beneficial for explainers, our primary aim was to examine whether these effects are associated with hypothesis generation, evaluation, or both.

A second coder who was blind to condition coded explanations and descriptions for reliability. A total of 552 explanations and descriptions belonging to 69 subjects were coded (this constitutes 80% of the complete set; the remaining 20% was used to train the second coder). Agreement ranged from 95.3% to 99.5% for shape, color, and purple dot references, with 94.6% agreement for references to mechanism. Disagreements were resolved through discussion among the two coders.

Hypothesis Generation Phase

After completing the eight evidence trials in the trial phase, participants were tested on hypothesis generation. First, they were given a free response prompt to assess whether they had inferred the target rule: "*Describe the single best rule you used for deciding whether or not each lure combination will catch fish.*" Next, they were given a classification task in which 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). This

classification task provided an indirect measure as a means of validating the hypothesis generation process alongside participants' free response answers. Critically, 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 itself.

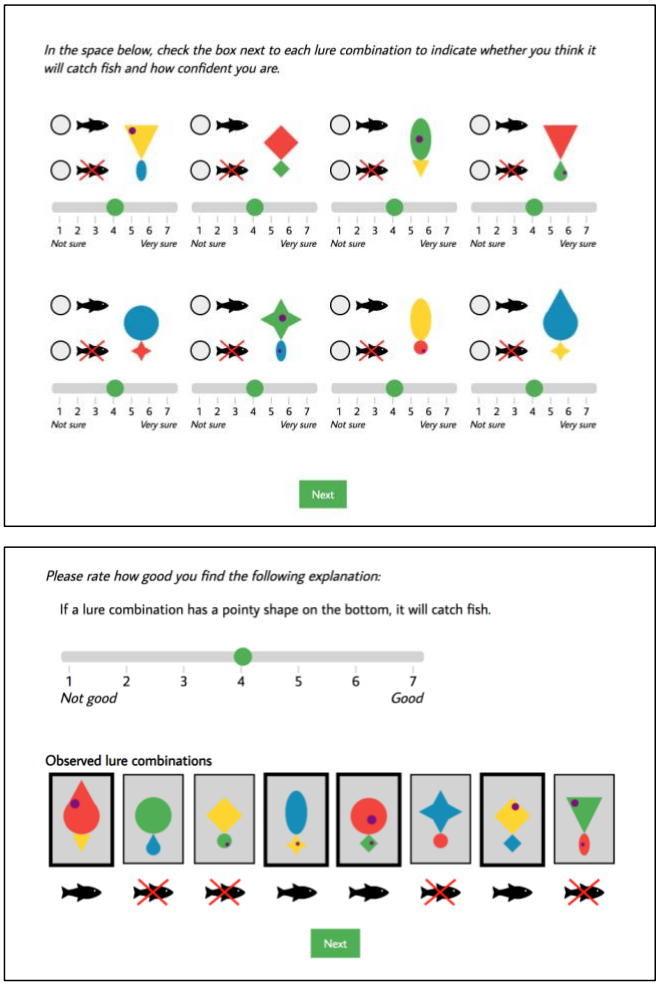

Figure 2. Top, the classification task used to test whether participants had generated a correct rule for categorizing fishing lure combinations in Experiment 1. Bottom, the hypothesis evaluation task for a sample rule in Experiment 1.

Generation Response Coding. 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, and agreement was 99%. Though it was possible to come up with a rule other than the target rule which was consistent with all of the evidence, no participant did so.

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 (see Table 1). The same six rules were presented to all participants in both conditions. 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 provided with a visual reminder of the outcome of each of the eight trials at the bottom of the screen. Following Bonawitz & Griffiths (2010), this was done to assess participants' appraisal of each hypothesis in light of the evidence. The decision to display the evidence during the evaluation phrase was critical. If the evidence had *not* been available, there is considerable risk that any variance observed in evaluation might have reflected variance in participants' memory of the training trials. Further, we might expect ratings of the target rule to be different depending on whether

participants had generated it during the earlier trials; those who did may be more likely to rate this rule highly.

Table 1. Rules presented in the hypothesis evaluation task in Experiment 1

Rule	Category	Consistency with Evidence
If a lure combination has a red shape or a blue shape, it will catch fish.	Misc.	62.5% (5/8)
If a lure combination has a diamond, it will catch fish.	Misc.	62.5% (5/8)
If a lure combination has a pointy shape on the bottom, it will catch fish.	**Target**	**100% (8/8)**
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% (8/8)*
If a lure combination has a purple dot on at least one of the lures, it will catch fish.	Misc.	75% (6/8)

Note: These six rules were presented in the fixed order above.

* See Footnote 3.

During the hypothesis evaluation task, the rules were presented in a fixed order for all participants: Any effects of order should therefore be stable across conditions. Of the six rules, the target rule and a "distractor" rule were both consistent with 100% of the evidence, but the distractor rule was considerably more complex ("*If a lure combination has a yellow shape or a*

diamond on the bottom, it will catch fish.").[3] If explaining 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 (i.e., simplicity, breadth, mechanism). The distractor rule was included to test whether explainers also *disfavored* rules that were consistent with all (or most) of the evidence but did not provide a "good" explanation. Three additional miscellaneous rules that were plausible but less consistent with the evidence (62.5% or 75%) were also included, as well as one rule suggesting that it was randomly determined which fishing lure combinations caught fish.

Memory Check

Finally, after completing the hypothesis evaluation task, participants were given a memory probe in which they were shown a set of eight fishing lure combinations, including four novel combinations and four that had previously appeared during the training phase. Participants all saw the same eight fishing lure combinations in the memory probe; the novel combinations were chosen from among a fairly limited set that participants had not previously seen on the evidence trials, predictions, or on the classification task. Participants were prompted to indicate whether they had seen each fishing lure combination at any point during the experiment. This was included to assess any differences in general attention between conditions. The memory probe also addressed the possibility that any condition differences observed on the hypothesis

[3] There is an alternative, pragmatically valid interpretation of the distractor rule (i.e., *if a lure combination has a yellow shape or a diamond on the bottom, it will catch fish* can be interpreted as either [1] *if a lure combination has a yellow shape anywhere or a diamond on the bottom,* or [2] *if a lure combination has a yellow shape on the bottom or a diamond on the bottom*). The latter interpretation is only consistent with 88% of the evidence (seven out of eight lure combinations). While the distractor remains appealing from an evidentiary standpoint in either case, it is possible that not all participants interpreted it as equivalent to the target rule [1]. Comparisons between distractor and target rule evaluations should therefore be interpreted with caution.

generation task were due to explainers having better memory for the evidence. Since participants were not provided with the evidence during the generation task, those learners who had not previously generated the target hypothesis but had improved memory for the stimuli might nonetheless have come up with the target hypothesis purely from memory. If so, we would expect better performance on the memory probe from participants in the *explain* condition.

Results

To understand the role of explaining on hypothesis generation and evaluation, we compare the *explain* and *describe* (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. For a summary of the evidence trial prediction results, refer to the Supplementary Materials.

Hypothesis Generation

We first examine the effect of explanation on hypothesis generation. Figure 3 shows accuracy on both hypothesis generation tasks. In line with our hypothesis, a significantly greater proportion of participants in the *explain* condition provided a correct hypothesis in their free response (51.2%) compared with describers (18.6%), $\chi^2(N = 86, 1) = 8.65$, $p = .003$. This difference is further borne out in participants' ability to apply the hypotheses they generated to novel stimuli; participants in the *explain* condition were better able to classify novel fishing lure combinations in the classification task compared with describers. First, we applied a similar strategy as the one used to analyze free responses above by coding participants who scored 100% on the classification task as having the correct hypothesis and all others as incorrect. The

proportion of participants meeting this criterion is significantly higher in the *explain* condition than in the *describe* condition (*explain*: .54; *describe*: .30; $\chi^2(N = 86, 1) = 3.87$, $p = .049$).[4]

However, this provides a rather coarse indication of the condition differences observed. To better account for the potential role of individual and item variation in classification accuracy across conditions, we fit a generalized linear mixed effects model (GLMM) to participants' response accuracy with condition as a fixed effect and random intercepts for participant and question item.[5] A likelihood ratio test revealed a significant effect of condition, with explainers more likely to produce correct classification responses, $\chi^2(1) = 8.31$, $p = .004$. Weighting participants' response accuracy by their confidence ratings—1 for correct answers, -1 for incorrect—produces similar results. Again, condition was a significant predictor of weighted accuracy judgments, $\chi^2(1) = 7.45$, $p = .006$. In sum, both the free response and classification measures indicate that participants in the *explain* condition were more likely to produce and apply a version of the target hypothesis, providing strong evidence that explanation plays a role in hypothesis generation.

---

[4] These results are robust to lower cutoff scores of 7/8 and 6/8 correct.
[5] All GLMMs and LMMs were fit in R using the 'lme4' package (Bates et al., 2015).

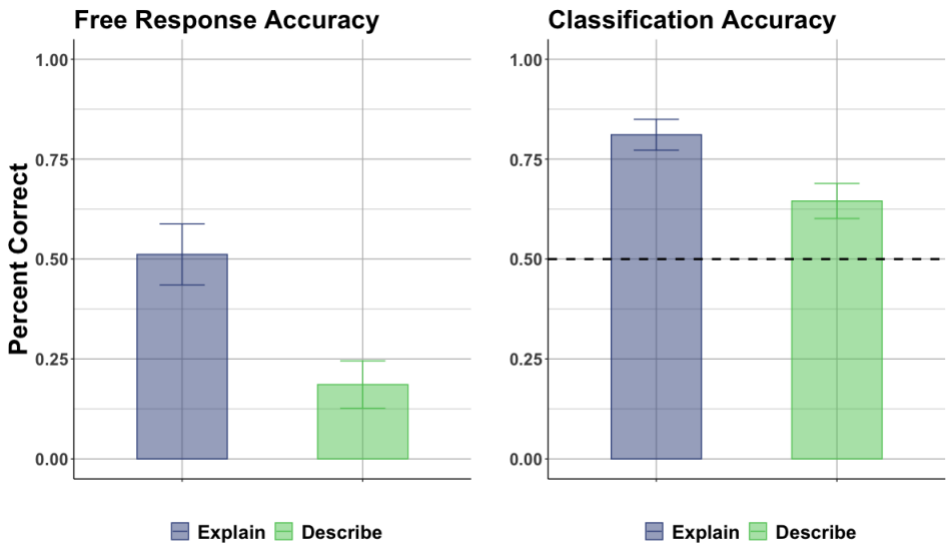

Figure 3. Accuracy on hypothesis generation tasks by condition in Experiment 1. Left: the proportion of responses coded as correct in the free response task. Right: the average classification accuracy per subject in the classification task. Error bars indicate one standard error of the mean (SEM).

Hypothesis Evaluation

Figure 4 shows average evaluation ratings for the target rule, the distractor rule, and the combined average ratings across all remaining rules.

Broadly, participants in both conditions rated the rules similarly: Evaluations of the target rule were near ceiling and higher than all other rules, including the distractor. We first analyzed whether the overall pattern of ratings differed across conditions and whether evaluations of the target and distractor rules were different in particular. To do this, we fit a linear mixed effects model to individual rule evaluations (1-7) with each rule as a fixed effect interacting with condition and a random intercept for participant. Model comparison via likelihood ratio test finds that including the main effect of condition does not improve model fit relative to a main effect of

rule alone, $\chi^2(1) = 0.35$, $p = .56$. This suggests that the general pattern of responding on each rule does not differ by condition. Further, including the interaction between condition and rule (as in the full model above) does not significantly improve model fit relative to the simple main effects of rule and condition, $\chi^2(5) = 9.53$, $p = .09$. Focusing on the target and distractor rule evaluations, estimated marginal means from the interaction model are not significantly different across conditions on the target or distractor rules (target: $p = .96$, distractor: $p = .85$). Thus, when accounting for individual variability in ratings, we do not find evidence of an effect of condition on hypothesis evaluation for the rules provided to participants.

When considering participant evaluations of each rule in isolation, we can compare the results above to traditional statistical methods based on individual ratings of a given rule. In a simple comparison of evaluation ratings across conditions, we find no significant difference in evaluations of the distractor rule, $t\ (84) = 1.53$, $p = .13$, but there *is* a significant difference between conditions in ratings of the target rule, $t\ (84) = -2.04$, $p = .045$. In paired *t*-tests comparing the target and distractor rule evaluations, we find that participants in both conditions rated the distractor rule significantly lower than the target rule (*explain*: $t\ (42) = -6.99$, $p < .001$; *describe*: $t\ (42) = -3.87$, $p < .001$). This may reflect a general prior preference for explanations which are not only consistent with the evidence, but also simple and easily generalizable (Williams et al., 2013). Recall, however, that some participants may have interpreted the distractor rule as consistent with seven out of eight, rather than all eight of the evidence trials. Although we must avoid drawing strong conclusions about participants' overall preference for the target rule, a 2 (*task*: explain, describe) by 2 (*rule*: target, distractor) analysis of variance (ANOVA) comparison of evaluation ratings finds a significant interaction between condition and rule type. This indicates that the difference between target and distractor rules was larger for

participants in the *explain* condition than in the *describe* condition, F (1, 82) = 5.60, p = .019. Explaining may therefore have led participants to treat the target and distractor rules as more distinct.

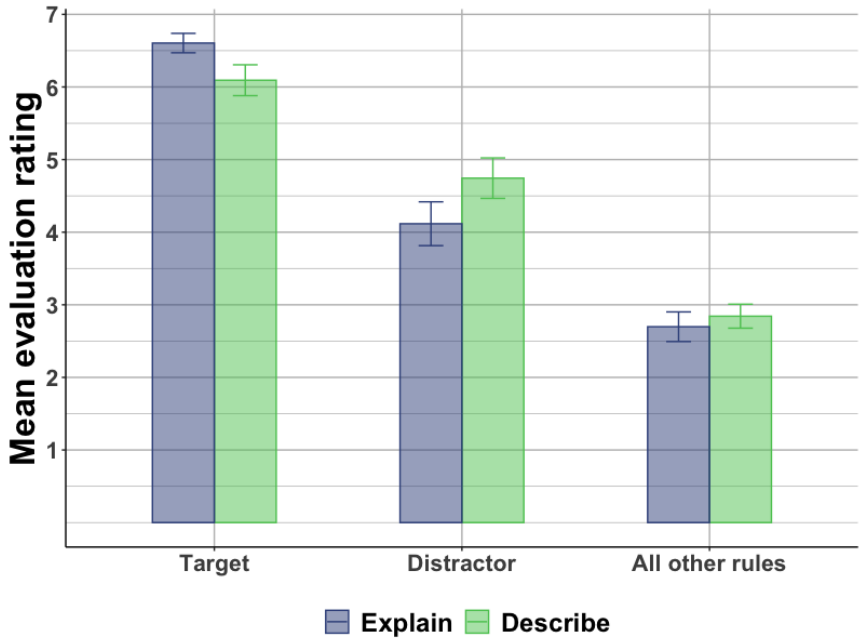

Figure 4. Experiment 1 evaluation ratings by condition for target rule, distractor, and all remaining rules aggregated. Error bars indicate one standard error of the mean (SEM).

In sum, our findings are consistent with the hypothesis that participants across conditions evaluate the target rule based on the available evidence, as well as its generality. While they do not conclusively suggest a role for explanation in hypothesis evaluation, these results also do not definitively rule out this possibility. Experiment 2 was designed to address this.

Memory

To assess whether the observed effects of explaining might be due to a general increase in attention or engagement that would be reflected in memory for task items, we tested condition

differences in memory for fishing lure combinations. To account for variation across individual responses and items, we fit a generalized linear mixed effects model to participants' response accuracy (binary) with a fixed effect of condition and random intercepts for subject and memory probe. A model comparison revealed no significant role of condition in explaining memory accuracy, $\chi^2(1) = 0.39$, $p = .535$. This suggests that the overall condition differences observed in hypothesis generation are not attributable to explainers' better memory of the stimuli.

We also investigated the potential role of attention or task engagement on hypothesis generation on an individual level, once again using data on the memory probe. We first ran a logistic regression of accuracy on the free response generation task as a function of condition and individual accuracy on the memory probe (percent correct). We find that memory probe accuracy is not a significant predictor of hypothesis generation behavior ($p = .563$), and that condition remains a significant predictor even after controlling for memory probe accuracy ($p = .002$). In line with our initial analysis of individual performance on the classification task, we fit a generalized linear mixed effects model to participant accuracy on each of the classification questions, this time with fixed effects of condition and individual memory probe accuracy (the random effects structure was the same as the previous analysis). Here, we find that including memory accuracy does not significantly improve fit over the random effects alone, $\chi^2(1) = 3.03$, $p = .082$, while the fixed effect of condition remains significant, even after including memory probe accuracy, $\chi^2(1) = 8.77$, $p = .003$. This further suggests that effects of attention or processing that may impact recall of task items cannot account for accuracy in hypothesis generation.

Explanation and Description Content

To assess whether explainers generate different *types* of hypotheses than describers (i.e., abstract, generalizable, mechanistic), or whether their improved performance is merely a result of generating *more* hypotheses during the process of explanation, we coded participants' responses during the trial phase for whether they addressed *abstract* or *concrete* features of the stimuli. In line with prior results (Williams & Lombrozo, 2010), we hypothesized that explainers would not reference *more* features of the lure combinations than describers, but would instead show a greater tendency to reference *abstract* features. We also coded participants' responses for whether they provided a mechanism in their explanation or description. We hypothesized that explainers would be more likely to provide a mechanism, though critically, there is nothing to prevent *describe* participants from providing this information in their descriptions as well. For example, one describer wrote: *"The main body of the lure is a circle and attached there is a teardrop shaped segment. Perhaps this teardrop piece is too round for the fish to latch onto."*

Figure 5 shows the average number of mechanisms and concrete and abstract feature references participants made in each condition, summed across the eight trials for each participant. Given the possibility of individual differences in participants' references to features and mechanisms over the eight trials, we analyze feature references using generalized linear mixed effect models for counts of each type of reference that factor in variance across subjects and trials (all GLMMs described in this section use Poisson regression unless otherwise noted). First, we model each participant's total reference counts on each trial (shape + color + purple dot) with random intercepts for subject and trial stimulus and a fixed effect of condition. A model comparison reveals that including the fixed effect of condition provides a significantly better fit to the data, $\chi^2(1) = 74.94$, $p < .001$, but this is because describers produce significantly *more*

feature references per trial than explainers (marginal mean estimates from the model above are 3.83 per trial for describers and 1.09 for explainers) (refer to Figure 5). Therefore, explainers' success in hypothesis generation is likely not a function of simply generating more hypotheses.

Instead, as Figure 5 suggests, explainers reference the same stimuli differently than describers. To better understand this, we model each participant's a) total mechanisms, b) total abstract references (shape + color + purple dot), and c) total concrete references (shape + color + purple dot) in each trial with a similar mixed effect structure to the one above, but now exploring the interaction between fixed effects of condition and reference "type" (mechanism, abstract, concrete). Model comparison using a likelihood ratio test reveals that the interaction of condition and reference type significantly improves model fit over main effects alone $\chi^2(2) = 682.88$, $p < .001$. Critically, estimated marginal mean number of mechanisms and abstract references per trial are significantly higher for explainers (mechanisms: *explain* = 0.48, *describe* = 0.06, $p < .001$; abstract references: *explain* = 0.52, describe = 0.09, $p < .001$), while concrete references per trial are significantly higher for describers (*explain* = 0.56, *describe* = 3.76, $p < .001$). These results echo previous findings reported by Williams & Lombrozo (2010) and suggest that explanation prompts learners to privilege certain *types* of hypotheses during generation. In particular, explainers were more likely to refer to the fishing lure combinations in abstract terms and provide mechanistic accounts of the ones that caught fish; in some cases, these accounts went well beyond the available evidence *("Because it looks like food that fish would like to eat and also have the smell the fish like"*).

An important question that arises from these results is whether explainers' success is a function of being prompted to explain (the *process* of explaining) or generating the right sort of explanation (the *product* of explaining) (e.g., Wilkenfeld & Lombrozo, 2015). To better

understand this, we ran a logistic regression with the free response data from participants in the *explain* condition to explore whether accuracy was predicted by the kinds of explanations provided in the earlier evidence trials. As predictors, we used: a) the total number of references (mechanism + abstract + concrete) across all eight evidence trials, b) the total number of mechanisms across all trials, and c) the *proportion* of abstract feature references out of all abstract and concrete references across all trials. This last metric was selected in place of total abstract or concrete references because we found that these references had a significant negative correlation (r = -0.36, p = .04). The proportion metric therefore allowed us to test whether the ratio of these features contributed separately from the overall number. We find a significant positive slope on proportion of abstract references only (p = .004).[6] To complement this, we ran a generalized linear mixed effects model comparison using individual responses on the *classification* task (rather than subjects' binary accuracy on the *free response* task); this analysis used a binomial link function for classification accuracy. A nested model comparison found that effects of mechanism and abstract reference proportion significantly improved model fit, while total references did not (*total references*: $\chi^2(1) = 1.80$, p = .18; *mechanisms*: $\chi^2(1) = 10.31$, p = .001; *abstract proportion*: $\chi^2(1) = 9.20$, p = .002).

Broadly, this suggests that an explainer's probability of generating a correct hypothesis likely increased with the proportion of abstract references and may also have increased by including mechanisms. Thus, our results suggest that the *kind* of explanation produced matters; participants who used more abstract references were also more likely to generate a correct hypothesis. However, given that nearly all explainers included these references in their

[6] This result and the one below are the same if we use total abstract references rather than the proportion.

explanations, these results cannot rule out the possibility that the *act* of explaining was itself epistemologically valuable, regardless of the explanation produced.

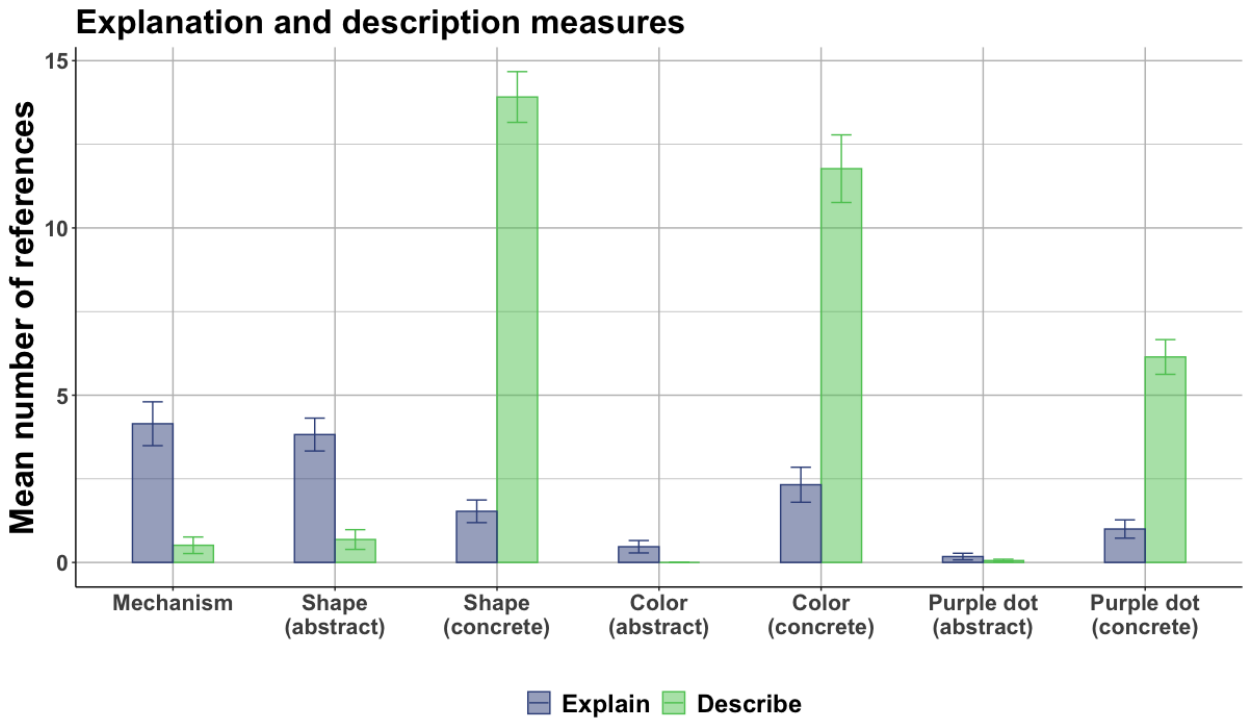

Figure 5. Average total number of mechanisms and concrete and abstract references to shape, color, and purple dot features in the explanations and descriptions provided by each participant in Experiment 1. Error bars indicate one standard error of the mean (SEM).

Discussion

In this experiment, we developed a novel category learning task to investigate the role that explaining plays in the processes of hypothesis generation and evaluation. We find that participants who are prompted to explain the evidence they observe are more likely to generate a correct rule for category membership than participants who were asked to describe the same evidence. This suggests that explaining may constrain the initial set of hypotheses generated by

the learner. By comparison, the effect of explaining on hypothesis evaluation is less clear. Although participants in both conditions rated the target rule significantly higher than all other rules, including the distractor rule, we did find a small but significant difference between conditions in their rating of the target rule. However, given that these ratings were near ceiling in both groups and that no such difference was observed when applying a mixed effects model, this finding is difficult to interpret. We address this issue in Experiment 2.

Experiment 2

Though the condition differences in hypothesis generation found in Experiment 1 were consistent with our initial predictions, the modest impact of explaining on hypothesis evaluation deserves further attention. One possibility is that the subtle difference we observed between conditions in their evaluations of the target rule was spurious. This seems likely, given that the target rule was rated highly across conditions and that our additional analysis accounting for individual subject variation in ratings did not find a significant effect of condition. In Experiment 2, we address this concern by increasing our sample size, modifying the hypothesis rating scale to reduce ceiling effects, and assessing participants' evaluation of a broader range of rules.

In addition to resolving this remaining uncertainty, we also aimed to address whether any effect of explaining on hypothesis evaluation was attenuated by the demands of the hypothesis *generation* task. First, the hypothesis generation task may itself involve some amount of tacit hypothesis evaluation. In particular, during the free response generation prompt, all participants were asked to provide the best rule for which lure combinations catch fish. This may have caused participants in both conditions to evaluate the goodness of the rule they were providing. In such a case, the hypothesis generation task might plausibly interfere with participants' subsequent evaluations, thereby masking effects of explaining on hypothesis evaluation.

A second possibility is that the free response hypothesis generation prompt served to reduce any differences between conditions on hypothesis evaluation by prompting *describe* participants to explain the evidence they saw. In other words, if the hypothesis generation prompt ("*Describe the single best rule you used for deciding whether or not each lure combination will catch fish*") led describers to seek a broad and generalizable hypothesis to apply to the data, they may have behaved more similarly to explainers in the subsequent hypothesis evaluation task. Concretely, producing "*the single best rule*" in the hypothesis generation task might have biased participants to evaluate abstract rules more favorably in both conditions. To address each of these concerns in Experiment 2, we removed the hypothesis generation tasks to assess condition effects on evaluation in isolation.

Method

Participants

Participants were 164 undergraduate students from a major West Coast university who received course credit for their participation. Unlike Experiment 1, which was completed on lab computers, Experiment 2 was administered to students online.[7] As in Experiment 1, participants were randomly assigned to either *explain* or *describe* (control) conditions. This sample size was chosen based on a power analysis indicating that we needed 82 participants in each condition to detect a difference in target rule evaluations with 80% power and an estimated effect size similar to Experiment 1.

An additional 9 participants were tested, but excluded, based on criteria established prior to data collection. Specifically, seven (*explain:* 4, *describe:* 3) were excluded for providing a

[7] All code for Experiment 2, as well as data and analysis code for the results presented, can be found at: `https://github.com/erik-brockbank/go_fish_v2`.

rating above 80 on a scale from 1 (“not good”) - 100 (“very good”) for a rule that was only consistent with 25% of the evidence observed (i.e., a very poor rule), and two (*explain:* 1, *describe:* 1) were excluded for total experiment completion times that were greater than five standard deviations above the group mean (i.e., over 8 hours). Note, however, that all reported results remain in the absence of one or both of these exclusions.

Procedure

The procedure for Experiment 2 was identical to Experiment 1, except for the following changes. First, we removed both hypothesis generation tasks. After completing the training phase, all participants proceeded directly to the hypothesis evaluation phase. Second, we modified the hypothesis evaluation task to include a set of eight rules (see Table 2). In addition to the target rule, distractor rule, and random (i.e., “no rule”) prompts from Experiment 1, we included two “virtuous” abstract rules that were consistent with 75% of the evidence (abstract shape rule: “*If a lure combination has a rounded top shape that resembles a fish’s body, it will catch fish*”; abstract color rule: “*If a lure combination has a top lure with bright colors that are more visible under water (red or yellow), it will catch fish*”). If explaining influences hypothesis evaluation, we predicted that explainers might rate these rules higher, despite their lack of parsimony (Williams & Lombrozo, 2013). Participants were also asked to evaluate three “miscellaneous” rules, which represented a broader range of consistency with the evidence.

Further, unlike in Experiment 1, in which the rule order was fixed, we randomized the rule order in Experiment 2 to avoid the possibility of order effects in either condition. Finally, participants evaluated all rules on a continuous 1-100 scale, from “not good” to “very good”, rather than a discrete 1-7 scale. As in Experiment 1, our primary dependent variable is the ratings that participants provided for each rule. However, once again, we also examine the *content* of the

explanations and descriptions provided during the trial phase to better understand *how* providing explanations may support the learning process.

Table 2. Rules presented in the hypothesis evaluation task in Experiment 2

Rule	Category	Consistency with Evidence
If a lure combination has a red shape on the bottom, it will catch fish.	Misc.	25% (2/8)
If a lure combination has a blue shape, it will catch fish.	Misc.	50% (4/8)
If a lure combination has a purple dot on at least one of the lures, it will catch fish.	Misc.	75% (6/8)
If a lure combination has a pointy shape on the bottom, it will catch fish.	**Target**	**100% (8/8)**
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% (8/8*)
If a lure combination has a rounded top shape that resembles a fish's body, it will catch fish.	Abstract (shape)	75% (6/8)
If a lure combination has a top lure with bright colors that are more visible under water (red or yellow), it will catch fish.	Abstract (color)	75% (6/8)

Note: The order of presentation for these eight rules was randomized.

*See footnote 8.

Results

To probe the role of explaining on hypothesis evaluation, we compare the *explain* and *describe* conditions on their evaluation of the eight candidate rules provided to all participants.

Hypothesis Evaluation

Figure 6 shows evaluation ratings for each rule: target, distractor, the two "abstract" rules, the "random" rule, and the "miscellaneous" rules, all indicated by their consistency with the evidence observed during the training phase (25%, 50%, 75%, or 100%).[8] The changes made in Experiment 2 removed the ceiling effects from Experiment 1, allowing for more meaningful analysis of target rule evaluations (*explain*: $M = 81.1$, $SD = 23.5$; *describe*: $M = 84.9$, $SD = 24.6$).

Paralleling our approach in Experiment 1, we begin with a mixed effects analysis with individual rule evaluations (1-100) modeled using interacting fixed effects of rule and condition and random intercepts for each subject. As in Experiment 1, the main effect of condition did not significantly improve model fit over a main effect of rule alone, $\chi^2(1) = 0.32$, $p = .57$, and the interaction between condition and rule did not improve model fit over the main effects $\chi^2(7) = 5.11$, $p = .65$. Considered alongside the previous findings, this suggests that explanation does *not* meaningfully intervene in hypothesis evaluation. Once again, we also evaluate the pairwise difference between rules across conditions using the full model described above. Here, a comparison of marginal means estimates finds no significant differences across conditions in their ratings of the individual rules.

We next turn to traditional statistics to complement these findings. Unlike in Experiment 1, a *t*-test of subject ratings on the target rule finds no significant difference in target rule evaluation (*explain*: $M = 81.1$, *describe*: $M = 84.9$), $t(162) = -1$, $p = .32$. We observe similar results in participant evaluations of the distractor rule (*explain*: $M = 63.8$; *describe*: $M = 61.6$, *t*

[8] As in Experiment 1, we note the possibility that the distractor rule, which was intended to be unambiguously consistent with 100% of the evidence, can also be interpreted in a way that is consistent with seven of the eight evidence trials.

(162) = 0.43, p = .67) as well as the "abstract" rules (abstract color rule; *explain*: M = 40.3; *describe*: M = 38.0, t (162) = 0.52, p = .60; abstract shape rule; *explain*: M = 44.5; *describe*: M = 39.4, t (162) = 1.15, p = .25). Taken together, findings of Experiment 2 provide no evidence that explanation impacts the process of hypothesis evaluation.

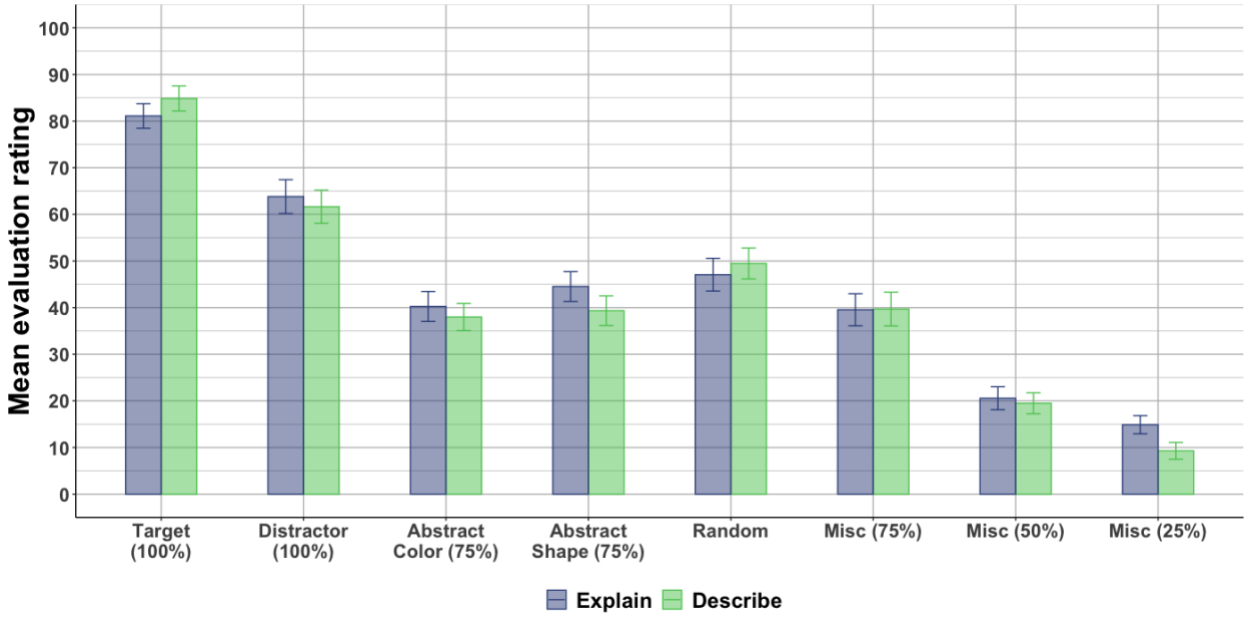

Figure 6. Evaluation results for Experiment 2. From left to right, ratings for the target rule, the distractor rule, the two abstract rules, the "random" rule, and three "miscellaneous" rules. Each label includes the percent of lure combinations (out of eight) that were consistent with the rule. Error bars indicate one standard error of the mean (SEM).

Instead, the current results suggest that hypothesis evaluation is sensitive to both the likelihood of the hypotheses (i.e., their consistency with the evidence), as well as information about their prior probabilities, signaled by their consistency with explanatory virtues. Further, results suggest that this sensitivity is not affected by explanation. First, as noted above, we

replicate the finding from Experiment 1 that, on average, the distractor rule is rated close to the midpoint of the scale, despite being consistent with all (or most) of the evidence. This indicates that prior knowledge likely plays an additional role in learner evaluations. Second, participant ratings of the "miscellaneous" rules suggest that learners incorporate likelihood information into their rule evaluations, with increases in mean ratings paralleling increases in consistency with the evidence (25%, 50%, and 75%). People's responsiveness to considerations of likelihood and prior knowledge appears to be equivalent for both *explain* and *describe* participants.

Explanation and Description Content

As in Experiment 1, we coded participant response data from the evidence phase of Experiment 2 to assess whether there were systematic differences in the kinds of hypotheses participants considered when viewing successful and unsuccessful lure combinations for the first time. For each participant's response on each of the eight evidence trials (explanation or description of the outcome), we count the number of abstract and concrete shape, color, and purple dot references, as well as the number of mechanisms provided in the response. The coding criteria were identical to those used in Experiment 1. As before, we hypothesized that explainers would show a greater tendency to reference abstract features, but not necessarily more features in aggregate (Williams & Lombrozo, 2010), and be more likely to provide a mechanistic account.

The trials were divided into two roughly equal sets and each set was coded by a pair of coders who were naïve to the subsequent analysis. The analysis reported here is based on 75% of the responses in each set (1,328 total), with the remaining responses used to train coders. Agreement between coders across the two sets averaged 95% for feature references and 88% for mechanisms. Disagreements were resolved by the experimenter.

Results from this analysis mirror those of Experiment 1. First, we model each participant's total reference counts on each trial (shape + color + purple dot) using a Poisson link function with random intercepts for subject and trial stimulus and a fixed effect of condition. As in Experiment 1, including the fixed effect of condition provides a significantly better fit to the data, $\chi^2(1) = 120.74$, $p < .001$, but this is because, once again, describers produced significantly *more* feature references per trial than explainers (marginal mean estimates are 3.8 per trial for describers and 0.8 for explainers, similar to the means in Experiment 1). Next, we model total a) mechanisms, b) abstract references (shape + color + purple dot), and c) concrete references (shape + color + purple dot) for each participant in each trial with the identical mixed effects structure used in Experiment 1. As in Experiment 1, we find that the interaction between condition and reference type significantly improves model fit, $\chi^2(2) = 1{,}220.79$, $p < .001$. Critically, estimated marginal mean number of mechanisms and abstract references per trial are significantly higher for explainers (mechanisms: *explain* = 0.42, *describe* = 0.04, $p < .001$; abstract references: *explain* = 0.41, *describe* = 0.07, $p < .001$), while concrete references per trial are significantly higher for describers (*explain*: 0.43, *describe*: 3.82, $p < .001$). Again, these are similar to the findings in Experiment 1. Together, findings suggest that explainers' success is due to generation of abstract and mechanistic hypotheses, not their generation of a greater number of hypotheses overall.

General Discussion

In two experiments, we examine whether explaining supports category learning by promoting *generation* of broad hypotheses, by leading learners to *evaluate* those hypotheses as more likely, or both. In Experiment 1, we found that participants who explained the evidence they observed were more likely to *generate* the target rule about which lure combinations catch

fish. However, we obtained mixed results with respect to the role of explanation in hypothesis *evaluation*. In Experiment 2, using a more diagnostic evaluation procedure and rule set, we find no evidence that explanation impacts hypothesis evaluation when examined in isolation. These findings provide strong evidence that explaining improves learning by intervening on the process of hypothesis generation, but not the evaluation of those same hypotheses.

There are several alternative explanations for the condition differences in hypothesis generation that are worth considering. First, it's possible that participants in the *explain* condition simply paid more attention to the evidence. Generating explanations is undoubtedly more challenging than simply describing that same evidence, so the increased attention required in this condition could have accounted for the results (e.g., Siegler, 2002). If this were the case, we might expect participants in the *explain* condition to have better memory for the fishing lure combinations. However, the results from the memory probe do not support this explanation. These same results also rule out a related interpretation of the observed condition differences, namely, that effects of explanation on generation were due to explainers' better recall of the training trials, since the evidence was not available for reference during the generation tasks.

Although these results rule out alternative proposals that the observed effects are due to increases in overall attention, it remains possible that explanation impacted performance by leading participants to generate *more* hypotheses than were generated in response to the describe prompt. Consistently re-sampling hypotheses over the eight evidence trials may have ultimately resulted in a greater proportion of explainers generating the target hypothesis. However, in that case, we would expect to find no differences in the *manner* in which fishing lure features were mentioned in explanations compared to descriptions; explainers might have provided candidate rules on the basis of features like shape and color (e.g., "lure combinations with yellow shapes

catch fish"), while describers might have simply described the same features (e.g., "this lure combination has a yellow shape on top"). Our analysis of the trial phase explanations and descriptions suggests that explainers were not merely sampling more rules about the same set of features but thinking about those features in fundamentally different ways. Specifically, although explainers provided fewer references to the fishing lure features overall, they were far more likely to make *abstract* references to those features (e.g., "round" rather than "circle"). Similar findings from the coded responses in our second experiment support these claims.

Finally, while we have suggested that it is broadly the *act* of explaining which produces the differences in hypothesis generation observed in our results, it is possible that producing the right *kind* of explanation (i.e., one that is sufficiently abstract and generalizable) facilitates success instead. Since nearly all explainers produced abstract or mechanistic references during the training trials, we are unable to evaluate whether the act of explaining supported hypothesis generation over and above the effect of explanation quality (e.g., see Wilkenfeld & Lombrozo, 2015). This represents a promising avenue for future work.

Future research might also explore the role of explaining in a wider range of generation contexts. Specifically, the constraints of the current task likely simplified hypothesis generation to a process of extrapolating from the available data and winnowing the set of possible hypotheses as evidence accumulates (Klayman & Ha, 1989; Goodman et al., 2008). However, this approach bypasses the more *constructive* process of hypothesis generation in everyday settings, in which the hypothesis space is initially less well-defined (Bramley et al., 2018; Gureckis & Markant, 2012).

Further, our finding that explaining does not impact hypothesis *evaluation* contrasts with at least one prior study (Williams et al., 2013) and raises additional questions for future inquiry

into the process of hypothesis evaluation. In Williams and colleagues, participants were asked to apply subtle statistical reasoning techniques to the evidence in order to evaluate hypotheses. In contrast, the current study provided participants with all the necessary evidence during evaluation, reducing the level of difficulty. It is therefore possible that explaining plays a greater role in the process of hypothesis evaluation in settings where evaluation itself is more cognitively demanding. It is also possible that explanation might impact hypothesis evaluation when the explanations are not generated by the participants themselves, e.g., in pedagogical settings where learners receive explanations from a teacher. In these situations, receiving possible explanations might affect the learner's evaluations of candidate hypotheses, independent of the evidence observed. Future work is needed to explore whether the effects of explanation in the current study generalize across learning contexts.

The present results provide several meaningful contributions to existing work on explanation and learning more broadly. First, they help to resolve a key question left unanswered by prior work on explanation: Though earlier results with children and adults showed that learners who explain tend to privilege hypotheses that are abstract and consistent with prior knowledge (Williams & Lombrozo, 2010, 2013; Walker et al., 2014, 2017), this might have been due to learners selectively generating these hypotheses, evaluating them differently, or both. Here, we show that explanation's primary function is to intervene on the process of hypothesis *generation*. This is consistent with prior literature on hypothesis testing, which has found that the set of hypotheses people entertain may be heavily dependent on contextual factors such as the framing of the task (Cheng & Holyoak, 1985) or the physical affordances of the problem (Walker et al., 2020). Prompting participants to explain can be viewed as a modification of the learning context which narrows the space of *candidate solutions* to the most broad and generalizable ones

(see, e.g., Ullman et al., 2016). This may provide additional insight into developmental results in which explaining has a dramatic and immediate effect on reasoning (Brockbank et al., in review; Walker, et al., 2017; Walker, et al., 2014 Experiment 1) and belief revision (Macris & Sobel, 2017). Further, this account may open the door to computational models of explanation, as well as hypothesis generation more broadly (e.g., Thomas et al., 2008).

More generally, the current work sheds light on some of the larger questions that lie at the heart of human learning and problem solving. First, our findings provide additional support for prior claims that hypothesis generation and evaluation are separable processes, and that different cognitive scaffolds may target learning in unique ways (Bonawitz & Griffiths, 2010). Second, these results add to a growing body of work examining the effects of learning context and goals in learners' ability to generate the right *type* of solution (Schulz, 2012, Lake et al., 2017; Walker et al., 2020). Despite the potentially infinite number of possible solutions to everyday problems, people are remarkably adept at selecting solutions that "make sense" (Ullman et al., 2016; Phillips, Morris & Cushman, 2019). Constraining the hypothesis space in this way remains a challenge for computational models of human inductive reasoning in many domains (Bonawitz & Griffiths, 2010; Lake et al., 2017). The current findings refine our understanding of *how* human learners accomplish this; the goal of producing good explanations constrains which hypotheses are initially generated when the learner is confronted with a novel problem.

Acknowledgements

This work was funded by a National Science Foundation CAREER award (SBE #2047581) and Jacobs Foundation Fellowship to C. Walker. We thank Alison Compton, Jake Truong, Jennifer Vu, and Amberley Stein for many hours of hard work coding participant

responses. In addition, we are thankful to Alex Rett and the members of the UCSD Early Learning and Cognition lab for their feedback at every stage of the project. We thank Elizabeth “Empirical Girl” Bonawitz and Tania Lombrozo for their comments on an earlier version of this manuscript and Edward Vul for assistance with statistical analyses.

References

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

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.

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, 322–330.

Bramley, N., Rothe, A., Tenenbaum, J., Xu, F., & Gureckis, T. (2018). Grounding compositional hypothesis generation in specific instances. In *Proceedings of the 40th annual conference of the cognitive science society*.

Brockbank, E., Lombrozo, T., Gopnik, A., & Walker, C. (in review). Ask me why, don't tell me why: Asking children for explanations facilitates relational thinking.

Butler, L. P., & Markman, E. M. (2012). Preschoolers use intentional and pedagogical cues to guide inductive inferences and exploration. Child Development, 83, 1416–1428.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive psychology*, *17*(4), 391-416.

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.

Engle, J., & Walker, C. M. (2021). Thinking Counterfactually Supports Children's Evidence Evaluation in Causal Learning. *Child Development*.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155-170.

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

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational behavior and human performance*, *24*(1), 93-110.

Gick, M., & Holyoak, K. (1980). Analogical problem solving. *Cognitive Psychology*, 12, 306-355.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1), 108-154.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British journal of psychology*, *73*(3), 407-420.

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

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, *94*(2), 211.

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.

Klein, G.A. (1993). A Recognition-Primed Decision (RPD) Model of Rapid Decision Making. In G. A. Klein, J. Orasanu, R. Calderwood, and C. Zsambok (Eds.), Decision Making in Action: Models and Methods. Norwood, NJ: Ablex Publishing Corp., 138–147.

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

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological review*, *96*(4), 674.

Kuhn, D., and Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology, 103,* 386–94.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and brain sciences, 40.

Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child development, 83*(1), 173-185.

Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child development, 81*(3), 929-944.

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

Lipton, P. (2008). *Inference to the best explanation*. London: Routledge.

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

Macris, D. M., & Sobel, D. M. (2017). The role of evidence diversity and explanation in 4-and 5-year-olds' resolution of counterevidence. *Journal of Cognition and Development*, *18*(3), 358-374.

Mehle, T. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychologica*, *52*(1-2), 87-106.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. Trends in cognitive sciences, 23(12), 1026-1040.

Schulz, L. (2012). Finding new facts; thinking new thoughts. In Advances in child development and behavior (Vol. 43, pp. 269-294). JAI.

Schunn, C.D., & Klahr, D. (1993) Self vs. Other-Generated Hypotheses in Scientific Discovery. In Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society.

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

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. *Psychological review*, *115*(1), 155.

Ullman, T., Siegel, M. H., Tenenbaum, J., & Gershman, S. (2016). Coalescing the Vapors of Human Experience into a Viable and Meaningful Comprehension. In CogSci.

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., Lombrozo, T., Williams, J.J., Rafferty, A., & Gopnik, A. (2016). Explaining constrains causal learning in childhood. *Child Development, 88*(1): 229-246.

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

Weber, E. U., Böckenholt, U., Hilton, D. J., & Wallace, B. (1993). Determinants of diagnostic hypothesis generation: effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1151.

Wilkenfeld, D. A., & Lombrozo, T. (2015). Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science & Education*, *24*(9), 1059-1077.

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.

Williams, J. J., Walker, C., Maldonado, S., & Lombrozo, T. (2013). Effects of Explaining Anomalies on the Generation and Evaluation of Hypotheses. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 35, No. 35).