2

Scientific Discovery and the Human Drive to Explain

Elizabeth Kon and Tania Lombrozo

Carl Hempel suggested that two human concerns provide the basic motivation for all scientific research (Hempel, 1962). The first is "man's persistent desire to improve his strategic position in the world by means of dependable methods for predicting and, whenever possible, controlling the events that occur in it." The second is "man's insatiable intellectual curiosity, his deep concern to *know* the world he lives in, and to *explain*, and thus to *understand*, the unending flow of phenomena it presents to him." Hempel isn't alone in highlighting a special role for explanations in science: others identify explanatory theories as the "crown of science" (Harre, 1985), with explanations as the "real payoff" from doing science (Pitt, 1988).

Why are explanations at the heart of science and scientific advance? In this chapter, we propose that explanations play a crucial role in scientific discovery, thereby advancing Hempel's first motivation for scientific research: the achievement of a better strategic position in the world through better prediction and control. The value of explanation is thus in large part instrumental (Lombrozo, 2011), with the quest for explanations driving scientific theory construction, and the generation of explanations linking theory to application.

The motivation for our proposal comes from recent work in cognitive psychology on the role of explanation in learning. This work suggests that the very process of seeking explanations motivates children and adults to go beyond the obvious in search of broad and simple patterns, thereby facilitating the discovery of such patterns, at least under some conditions (Lombrozo, 2016). Might the drive for scientific explanation play a similar role in science, prompting individuals and communities to search deeper and harder for broad and simple generalizations that characterize the natural world? Could features

of explanatory cognition themselves explain features of scientific practice and theorizing, such as the allure of exceptionless laws and simple theories?

To a large extent, these questions remain unanswered: there has been little empirical research on the role of explanation in actual scientific practice,[1] nor will we report such research here. Instead, our aim is to evaluate how research on everyday human cognition might extend to scientific contexts by reviewing prior psychological research, and by presenting new studies that explore the effects of explanation in a learning environment that is more representative of most scientific practice. Specifically, we explore a puzzle that arises from prior research. On the one hand, this research suggests that when people engage in explanation, they aim to achieve an explanatory ideal: obtaining explanations that are underwritten by *simple and exceptionless* patterns or generalizations. On the other hand, we know that in real scientific practice, such generalizations are rarely to be found. Could it be that the search for ideal explanations is beneficial in part because it facilitates the discovery of real but imperfect generalizations—for example, those that involve some exceptions? (For impatient readers, we offer a hint: the answer is "yes.") But before turning to our new studies, we briefly review relevant prior work.

The role of explanation in learning

Decades of research reveal that the process of explaining—even to oneself—can have a powerful effect on learning (e.g., Fonseca and Chi, 2011; Lombrozo, 2012; Chi et al., 1989). Several psychological processes contribute to this phenomenon. For example, attempting to explain something can help people appreciate what they do not know (Rozenblit and Keil, 2002), make them accommodate new information within the context of their prior beliefs (Chi et al., 1989; Williams and Lombrozo, 2013), and lead them to draw inferences to fill gaps in their knowledge (Chi, 2000). There is also evidence that when engaged in explanation, both children and adults seek explanations that are *satisfying*, where satisfying explanations are those that account for what is being explained by appealing to broad and simple rules or patterns (Lombrozo, 2016). For example, Williams and Lombrozo (2010, 2013) found that when presented with an array of items belonging to two categories, adults who were prompted to explain why each item belonged to its particular category (e.g., why robot A is a "glorp" and robot B is a "drent") were more likely than those in control conditions to discover a subtle classification rule that accounted for the category membership of all items on the basis of a single feature (see also Walker, Bonawitz, and Lombrozo, 2017). This was true whether participants in the control condition were prompted to describe the category exemplars, to think aloud as they studied them, or to simply engage in free study.

If explanation is so beneficial for learning, one might wonder why people don't explain more often. In other words, why don't children and adults engage in explanation spontaneously, even in the "control" conditions that are used as a baseline against which to compare the performance of participants who are explicitly prompted to explain? To some extent, people do explain without an explicit prompt: participants explain to varying degrees, even in control conditions, and this variation predicts what they ultimately learn (Edwards et al., 2019; Legare and Lombrozo, 2014). But there's more to the story than that. It's possible that people are frugal explainers not only because it is effortful to explain, but also because *explicitly engaging in explanation does not always yield superior performance.* Under some conditions, prompts to explain result in learning that is no different from that in a control condition (Kon and Lombrozo, in prep), suggesting that explanation is unnecessary. Under other conditions, prompts to explain can actually be detrimental (Williams, Lombrozo, and Rehder, 2013; see also Legare and Lombrozo, 2014; Walker et al., 2014). It's instructive to consider these cases in turn.

First, Kon and Lombrozo (in prep) identify conditions under which a prompt to explain is unnecessary in the sense that it does not lead to performance that exceeds that of control conditions. They find that when it comes to discovering a subtle, exceptionless pattern describing a set of observations, participants who are prompted to explain only surpass those in a control condition when there is a compelling but inferior pattern for those in the control condition to latch on to, such as a salient pattern that accounts for 75 percent of observed cases. For nonexplainers, this alternative is sufficiently compelling to limit the further expenditure of cognitive resources. But for explainers, a pattern that only accounts for a subset of cases, or that does so in a complicated way, isn't good enough; the expectation or hope of a more satisfying explanation spurs them on. Studies with young children similarly hint at the idea that explaining is a spur to go "beyond the obvious" to find a pattern that is more subtle, but also in some regards superior, to more salient possibilities (Walker et al., 2014; Walker, Bonawitz, and Lombrozo, 2017). In the case of science, this could mean that seeking explanations (and not, say, mere descriptions) is likely to spur additional discoveries when the discoveries go beyond salient regularities to capture generalizations over nonobvious properties.

Second, Williams, Lombrozo, and Rehder (2013) find that under some conditions, a prompt to explain observations can actually be detrimental. In their task, participants had to learn how to classify vehicles into two categories, or how to identify individuals as likely or unlikely to make charitable donations. In some cases, these tasks could be achieved by identifying a single theme or feature that characterized all members of one category and none of those in the other. But in other conditions, the only way to achieve perfect classification was to memorize the idiosyncratic properties of individual exemplars—for instance, that the cyan car was a "kez" or that *Janet* frequently donates to charities. Under these conditions, those participants who were prompted to explain took longer to learn how to accurately categorize all of the exemplars; they seemed to perseverate in looking for a broad and simple *pattern* before settling for a rote strategy based on individuals. Generalizing to science, we might expect the search for explanations to be detrimental when there is *no structure at all* to the observations being explained. This is probably an unusual situation for science, but it might arise when a set of observations is grouped according to a wholly inaccurate theory or when observations reflect noise rather than some underlying signal.

What is it about broad and simple patterns that satisfies the demands of explanation? Or conversely, what is it about patterns with exceptions or additional complexity that *fails* to satisfy the demands of explanation? Recent work by Kon and Lombrozo (in prep) contrasts two possibilities: that explainers favor exceptionless patterns because such patterns maximize predictive power, or that explainers favor exceptionless patterns because such patterns make for more virtuous explanations—that is, explanations that exhibit the explanatory virtues of simplicity and breadth. To differentiate these alternatives, they created learning tasks in which participants could achieve perfect predictive accuracy on the basis of two salient features of the stimuli (thus achieving breadth at the expense of simplicity), or potentially discover a more subtle pattern that also supported perfect predictive accuracy and did so on the basis of a single feature (thus achieving both breadth and simplicity, but at a cost of greater cognitive effort). Participants who were prompted to explain were significantly more likely than those in a control condition to discover the more subtle rule. This suggests that the salient, predictively perfect (but less virtuous) alternative was insufficient to satisfy their explanatory drive. This fits well with a familiar observation from science: the most predictive model isn't always the most explanatory. Explanation seems to require something more than successful prediction.

To sum up, prior work on the effects of explanation on learning suggests that when people are actively engaged in seeking explanations for particular observations, they're more likely to find simple, exceptionless patterns that underlie those observations, and that these patterns are compelling because they support virtuous (i.e., simple and broad) explanations. Explanation is not always beneficial (relative to control conditions), but it does appear to be beneficial when two conditions obtain: when there is a broad and simple pattern to be found, and when there is a more salient but inferior alternative for participants in the control condition to latch on to. It may not be accidental, then, that science focuses so heavily on explanation. The natural world does appear to be bursting with patterns, many of which can only be formulated over unobservable and otherwise nonobvious properties. Nature seems to reward those who not only consider the obvious but also go beyond it. These findings also resonate with a feature of how "idealized" physics is often portrayed: as a search for exceptionless laws and elegant theories. Even in domains that don't aspire to exceptionless laws, there seems to be value in minimizing and accounting for exceptions. The search for simple and broad generalizations thus seems to act as a powerful motivating force: it's not enough to find a pattern; it must be a pattern of the right sort.

Despite these synergies between our experimental studies and observations about science, an important puzzle remains. After all, scientists rarely succeed in identifying truly exceptionless laws. Especially within the social sciences, generalizations are invariably imperfect and riddled with exceptions. In some domains, accounting for even 75 percent of the variance in the manifestation of some property (such as personality) is a notable achievement. Could it be that engaging in explanation motivates everyday learners—and scientists—to search for simple, exceptionless patterns, but that in the course of doing so, *they're also more likely to discover other subtle but imperfect regularities that nonetheless constitute an advance*?

Evidence that this could be so comes from Experiment 3 of Kon and Lombrozo (in prep), in which participants were tasked with learning how to determine whether novel creatures eat flies or eat crabs. Half the participants were prompted to write down an explanation for each observation (i.e., for why a particular creature eats flies or crabs), and half (in the control condition) were prompted to write down their thoughts about that observation. The observations were designed to support two possible generalizations. First, participants could learn to predict the diet of all studied examples on the basis of two features of the stimuli, their habitat *and* age, which was a complex but exceptionless

pattern. Second, participants could learn to predict the diet of a majority of studied examples (75 percent) on the basis of a single feature—snout direction—which was a simple rule, but one with exceptions. Kon and Lombrozo found that participants who were prompted to explain were more likely than those in the control condition to discover each of these rules, presumably because they stumbled across them in their search for an ideal explanation: one that was *both* simple and exceptionless. This finding suggests that even if a simple, exceptionless pattern describes some explanatory ideal that is rarely realized, the pursuit of this ideal could spur meaningful discoveries.

So far we've considered the role of explanation in learning, and how prior work on explanation might shed light on the puzzle of whether and why seeking ideal explanations is beneficial, given that we inhabit a less-than-ideal world. In what follows, we describe a pair of novel experiments designed to test our core hypothesis in a more systematic fashion: that in pursuing an ideal explanation, explainers increase their odds of discovering some of the nonideal structure to be found. In Experiment 1, we thus present learners with a nonideal world (i.e., one that does not support a maximally simple and broad generalization) and investigate the effects of a prompt to explain.

Explaining in a nonideal world: Two novel experiments

Experiment 1: Is explaining beneficial when all generalizations involve exceptions?

Experiment 1 investigates whether in the absence of an ideal pattern (i.e., one that is both maximally simple and broad), engaging in explanation can nonetheless assist with the discovery of the best available alternatives. To test this, we designed a task in which participants learned to categorize items into one of two categories. As they studied twelve labeled exemplars (six from each category), they were prompted either to explain or to write down their thoughts about the category membership of the exemplars. Two rules could be used to categorize the items. One rule was fairly salient and therefore easy to discover, but only captured the category membership of eight of the twelve exemplars (it was thus a "66 percent rule"). Another rule was much subtler but captured the category membership of ten of the twelve exemplars (it was thus an "83 percent rule"). So while the latter rule still fell short of the ideal (i.e., a rule that captured all twelve items, a "100 percent rule"), it was superior to the initial rule along the dimension of breadth.

If explaining assists in the discovery of the best possible rule, even if it is imperfect, we would expect participants prompted to explain to be more likely than those in the control condition to discover the 83 percent rule. By contrast, if effects of explanation are restricted to the ideal case—an exceptionless rule[2]—then we would expect those participants who were prompted to explain to perform no better than those in the control condition.

In addition to the *no ideal rule* condition just described, we also considered an *ideal rule* condition, in which the more salient rule accounted for 83 percent of cases, and the more subtle rule accounted for 100 percent of cases. This more familiar situation is a replication of prior research, but with a larger number of training exemplars (twelve versus eight) to accommodate the intermediate percentages. We included it in Experiment 1 in part as an extension of prior research but also to serve as a basis for comparison against the *no ideal rule* condition. Thus we can ask not only whether a prompt to explain facilitates discovery of a "better" rule when the better rule is an 83 percent rule (versus a 66 percent worse rule), but also whether the magnitude of this effect is comparable to the effects of explanation when the "better" rule is a 100 percent rule (versus an 83 percent worse rule).

Method

Participants

The sample for Experiment 1 consisted of 1293 adults[3] (after exclusions)[4] recruited through Amazon Mechanical Turk and paid for their participation. In all studies, participation was restricted to adults with an IP address within the United States and with an approval rating of at least 95 percent on fifty or more previous tasks. Participants were also prevented from participating in more than one study from this paper.

Materials

The stimuli consisted of ten sets of twelve items. The twelve items in each set depicted flowers, containers, objects, simple robots, or complex robots. Throughout this chapter, we will use flowers as an illustrative example, but information concerning the other four stimulus types is available in Appendix A.

Each set contained items from two categories, with six items belonging to each category. For example, half of the flower items were SOMP flowers and half of the flower items were THONT flowers. For each set, participants could use

two possible rules to determine which category an item belonged to. One rule was always "better" in the sense that it could be used to correctly categorize more items than the "worse" rule. In the *ideal rule* condition, the better rule was a "100 percent rule" that perfectly accounted for the category membership of all twelve items, and the worse rule was an "83 percent rule" that correctly categorized only ten of the twelve items (see Figure 2.1). For the *no ideal rule* condition, the better rule was an "83 percent rule" that correctly categorized ten of the twelve items, and the worse rule was a "66 percent rule" that correctly categorized eight of the twelve items.

Procedure

The task consisted of a study phase followed by a reporting phase and a rule rating phase. At the start of the study phase, participants were randomly assigned to one of four conditions, which were created by crossing two prompt-types, *Explain* or *Write Thoughts*, with two pattern-types, *ideal rule* or *no ideal rule*. Participants were randomly assigned to see one of the stimulus sets.[5]

In the study phase, all participants were told to study the items, and that after the study phase they would be asked questions about how to determine which category each item belongs to. Participants were presented with a randomized array of the twelve items corresponding to their condition's pattern-type

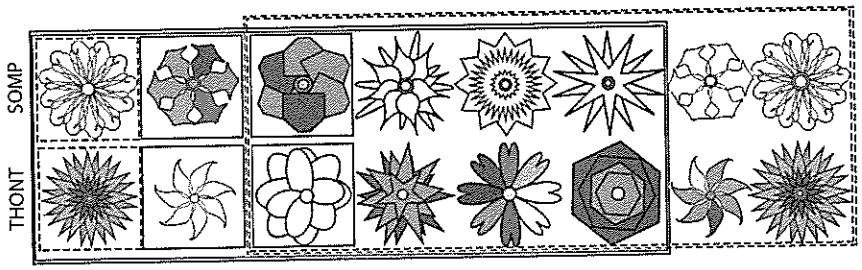

Figure 2.1 Flower stimuli. For these flower stimuli, the better rule (100 percent in the *ideal rule* condition and 83 percent in the *no ideal rule* condition) is that SOMP flowers have two concentric circles in their centers, whereas THONT flowers have one circle in their centers, and the worse rule (83 percent in the *ideal rule* condition and 66 percent in the *no ideal rule* condition) is that the petals of SOMP flowers are mostly one color (in the colored version, these are all different colors; in grayscale, they are all the same tone to increase the salience of the worse rule), while the petals of THONT flowers are mostly rainbow-colored (indicated in grayscale with higher-contrast leaves). Within this figure, the double solid outline contains the items in the *no ideal rule* condition, and the double dotted outline contains the items in the *ideal rule* condition. The exceptions to the worse rule are in solid boxes and the exceptions to the better rule are in dashed boxes.

(ideal rule or *no ideal rule).* They were then prompted to focus their attention on each item, individually, in a random order, with a prompt determined by the experimental condition to which they were randomly assigned. Participants in the *explain* conditions were told (for example) to "try to *explain why* flower A is a SOMP flower." Participants in the *write thoughts* conditions were told to "*Write out your thoughts* as you learn to categorize flower A as a SOMP flower." Participants were given 50 seconds to respond to each prompt by typing into a text box, at which time their responses were recorded and the prompt for the next item appeared.

In the reporting phase, participants were asked to report all patterns that they noticed that differentiated SOMPS and THONTS, even if the patterns were imperfect.[6] In addition to describing the rule they discovered in a free-response box, participants were asked how many of the twelve items they thought followed the rule.

After finishing the reporting phase, participants were again presented with all twelve items as well as four candidate rules, presented in a random order, purporting to explain "why flowers A–F are SOMPS (as opposed to THONTS)." They were forced to stay on the page for at least 15 seconds to ensure that they read the explanations. The candidate rules referenced the better rule, two versions of the worse rule (one indicating that it involved exceptions, one not), and one filler item that was a bad/untrue explanation (see Appendix B for complete set of stimuli). Samples for the flower items are included in Table 2.1. Ratings were collected on a 7-point scale with anchors at 1 ("Very Poor Explanation") and 7 ("Excellent Explanation").

Before concluding the experiment, participants completed an attention and memory check question that served as the basis for participant exclusion.[7] Finally, participants were asked to report their age and sex.

Results

Overall rule reporting. Participants reported finding an average of 1.23 patterns (*SD* = 1.23, min = 0, max = 9) that they reported accounted for an average of 8.18 exemplars (*SD* = 3.13, min = 0, max = 12). Reported patterns were coded for mention of the better rule and/or the worse rule.

Better rule reporting. To test whether explanation prompts affected discovery of the better rule (100 percent or 83 percent, depending on pattern-type), and whether effects differed across pattern-type (see Figure 2.2), we conducted a logistic regression predicting whether participants *discovered the better rule*

Table 2.1 Average rule ratings by condition

Rule Type	Flower Rule Text	Explainers-Ideal Rule Condition	Thinkers-Ideal Rule Condition	Explainers-No Ideal Rule Condition	Thinkers-No Ideal Rule Condition
Better	Because SOMP flowers have two circles in their centers, and THONT flowers have one circle in their centers.	6.07 (1.72)	5.95 (1.84)	4.41 (1.97)	4.32 (2.00)
Worse	Because SOMP flowers have petals which are all the same color, and THONT flowers have petals with many colors.	3.72 (2.19)	3.65 (2.05)	3.05 (1.80)	2.89 (1.73)
Worse + exception	Because SOMP flowers have petals which are all the same color, and THONT flowers have petals with many colors, though there are some exceptions.	4.74 (2.21)	5.03 (2.13)	4.26 (1.97)	4.26 (2.12)
Bad	Because SOMP flowers have black polka-dots on their petals, and THONT flowers have no black polka-dots on their petals.	1.49 (1.25)	1.43 (1.22)	1.70 (1.36)	1.54 (1.27)

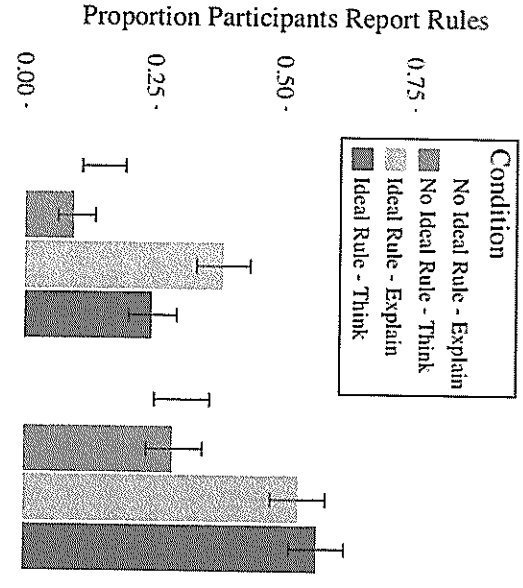

Figure 2.2 The proportion of participants reporting each rule in Experiment 1, as a function of rule type, condition, and prompt. Error bars correspond to 95 percent confidence intervals.

(yes vs. no) by *prompt-type* (explain vs. write thoughts) × *pattern-type* (*ideal rule* vs. *no ideal rule*) × *stimulus-type* (flowers vs. containers vs. objects vs. simple robots vs. complex robots). This revealed a significant effect of prompt-type on reporting the better rule ($\chi^2 = 17.23$, $p < 0.01$), with higher discovery rates for participants prompted to explain. There was also a significant main effect of pattern-type, with more participants reporting the better rule when it accounted for more items ($\chi^2 = 72.38$, $p < 0.01$). The interaction term between prompt-type and pattern-type was not significant ($\chi^2 = 0.63$, $p = 0.43$). The interaction term between prompt-type and stimulus-type was also not significant ($\chi^2 = 6.99$, $p = 0.14$).[8] These findings suggest that explaining did indeed facilitate discovery of the better rule, regardless of whether the better rule was ideal, and across a range of different stimulus types.

The results of this analysis are consistent with the hypothesis that when explaining, people seek simple and exceptionless rules, but that in the course of doing so, they are likely to discover "good" rules that may nonetheless fall short of this ideal. To further investigate this pattern of results, we ran additional logistic regressions for the *ideal rule* condition and *no ideal rule* condition separately.

We found that explainers reported the better rule significantly more often than those who wrote their thoughts within the *ideal rule* pattern-type condition ($\chi^2 = 15.53$, $p < 0.01$) and also within the *no ideal rule* condition ($\chi^2 = 3.94$, $p = 0.05$).[9] These results further support the claim that engaging in explanation can assist people in discovering the best available rule, even when it is imperfect.

Worse rule reporting. Previous studies have found that prompting participants to explain can sometimes decrease worse rule reporting relative to a control condition (e.g., Edwards, Williams, and Lombrozo, 2013; Williams and Lombrozo, 2010, 2013). To analyze worse rule reporting, we ran another logistic regression: *discovered the worse rule* (yes vs. no) by *prompt-type* (explain vs. write thoughts) × *pattern-type* (*ideal rule* vs. *no ideal rule*) × *stimulus-type* (flowers vs. containers vs. objects vs. simple robots vs. complex robots). The effect of prompt-type was not significant ($\chi^2 = 0.39$, $p = 0.53$). The effect of pattern-type was significant ($\chi^2 = 84.79$, $p < 0.01$): participants reported the 83 percent worse rule more often than the 66 percent worse rule. However, the interaction between prompt-type and pattern-type was not significant ($\chi^2 = 1.00$, $p = 0.32$).[10] These findings suggest that while explaining improved discovery of the better rule, it did so at no cost to discovery of the worse rule.

Rule Ratings. To analyze rule ratings (see Table 2.1), we first confirmed that participants were attentively engaged in the task by verifying that ratings for the bad rule were significantly lower than those for the other three options. Using *t*-tests comparing each of the three "good" options against the bad rule within each of the four conditions revealed a significant difference in each case, even using a Bonferroni correction for multiple comparisons.

To analyze ratings for the three good rules, we performed an ANOVA with prompt-type (2: explain, write thoughts) and pattern-type (2: *ideal rule, no ideal rule*) as between-subjects factors, and rule rated (3: better rule, worse rule, worse rule acknowledging exceptions) as a within-subjects factor. This analysis revealed no main effect of prompt-type, $F(1, 992) < 0.01$, $p = 0.97$, a significant main effect of pattern-type, $F(1, 992) = 133.21$, $p < 0.01$, and a significant effect of rule rated, $F(2, 1984) = 292.11$, $p < 0.01$. The main effects of pattern-type and rule rated were qualified by a significant interaction, $F(2, 1984) = 24.75$, $p < 0.01$.[11] Not surprisingly, the better rule was rated more highly when it accounted for 100 percent of cases than when it accounted for 83 percent of cases, $t(923) = -13.61$, $p < 0.01$, consistent with our assumption that explanatory evaluation favors patterns without exceptions. We also found that the worse rule was rated more highly when it accounted for more cases, whether the rule did, $t(982) = -4.68$, $p < 0.01$, or did not, $t(994) = -5.80$, $p < 0.01$, mention exceptions. This is consistent with our finding that explaining also favors the discovery of patterns that account for more cases, even when both fall short of 100 percent. However, the gap in ratings between the "better" 83 percent and 100 percent rules (1.64 points on a 7-point scale) was greater than that between the "worse" 66 percent and 83 percent rules, both when the rule did (0.63 points) or did not (0.71 points) mention exceptions, accounting for the significant interaction and also suggesting that there may be something special about a rule without any exceptions.

Discussion

The results of Experiment 1 both replicate and extend prior research. Consistent with prior research, we found that a prompt to explain facilitated discovery of a subtle, exceptionless rule. Going beyond prior research, we also found that a prompt to explain facilitated the discovery of a subtle rule that involved exceptions, albeit *fewer* exceptions than a more salient alternative. This helps resolve the puzzle with which we began. On the one hand, seeking explanations seems to push learners to achieve an explanatory ideal, which involves simple, exceptionless generalizations. (Indeed, our rule rating results suggest that such generalizations are highly valued.) But on the other hand, real-world domains rarely support the realization of this ideal. We find that even in a domain where the ideal cannot be attained, engaging in explanation may be useful because it pushes learners to go beyond the obvious in search of a "better"—albeit imperfect—regularity.

Experiment 2: Extension to an easier learning environment

In Experiment 2, we sought to replicate and extend the results of Experiment 1. Specifically, the experiment considers whether the effects observed in Experiment 1 will generalize to a context in which the task of discovery is simplified by making the defects of the "worse" rule, and the features that support the "better" rule, easier to identify. One possibility is that even in an easier learning task, the effects of explanation will continue to surpass those of our control condition. But another possibility is that by making it easier for participants in the control condition to "go beyond the obvious," we will boost their performance to a level comparable to that of participants prompted to explain.

How might the subtler pattern be made "more obvious" in a relevant way? Experiment 2 altered the difficulty of the learning task by using a different presentation format. In Experiment 2, items were presented in a more structured

array, and participants were asked to consider all items in a category together rather than studying each item independently. Previous work suggests that presentation format can have a significant effect on category learning (Meagher et al., 2017), and that increasing the ease with which (or the rate at which) category members are compared can also affect learning (Edwards et al., 2019; Lassaline and Murphy, 1998). There's also evidence that explanation may help, in part, by drawing attention to category-wide statistical properties, and by encouraging participants to focus on the contrast between one category and the other (Edwards et al., 2019; Chin-Parker and Bradner, 2017). One might therefore expect the presentation format in Experiment 2 to mimic some of the benefits of explanation by making category-wide properties and diagnostic features more salient, thereby rendering the inferior rule more obviously inferior, and the more subtle pattern less difficult to detect.

In sum, Experiment 2 contrasts two hypotheses: that effects of explanation (relative to a control condition) are fairly stable across variation in the difficulty of a learning task, and the alternative that the effects of explanation (relative to a control condition) are moderated by task difficulty. In an easier learning task, even participants in a control condition may readily go beyond the obvious, rendering a prompt to explain somewhat superfluous. Such a moderating effect could shed light on whether and why effects of explanation could vary as a function of the domain, data, and tools for analysis available to a scientist or everyday learner.

Method

Participants

The sample for Experiment 2 consisted of 1470 adults recruited as in Experiment 1.[12]

Materials

Stimuli were the same as those used in Experiment 1.

Procedure

The experiment consisted of a study phase and a rule reporting phase that were very similar to those of Experiment 1. However, there were two key differences. Rather than randomly arranging the twelve items into a 3 × 4 grid, items were grouped by category in two 3 × 2 groups. Additionally, rather than being asked about each item individually in a sequential manner, participants were given three minutes to study each of the two categories, with a prompt to either explain or write their thoughts about all of the items in a category together.

Results

Better rule reporting. To test whether explainers reported the better rule more frequently than thinkers (see Figure 2.3), we ran a logistic regression predicting whether participants *discovered the better rule* (yes vs. no) by *prompt-type* (explain vs. write thoughts) × *pattern-type* (*ideal rule* vs. *no ideal rule*) × *stimulus-type* (flowers vs. containers vs. objects vs. simple robots vs. complex robots). We found no effect of prompt-type ($\chi^2 = 0.30$, $p = 0.58$), suggesting that in this context, explaining does not facilitate the discovery of the better rule. There was also a significant effect of pattern-type ($\chi^2 = 73.79$, $p < 0.01$), with higher discovery rates when the better rule accounted for more cases.[13]

To further investigate why there was a significant effect of prompt-type on better rule discovery in Experiment 1 but not in Experiment 2, we ran a logistic regression predicting whether participants *discovered the better rule* (yes vs. no)

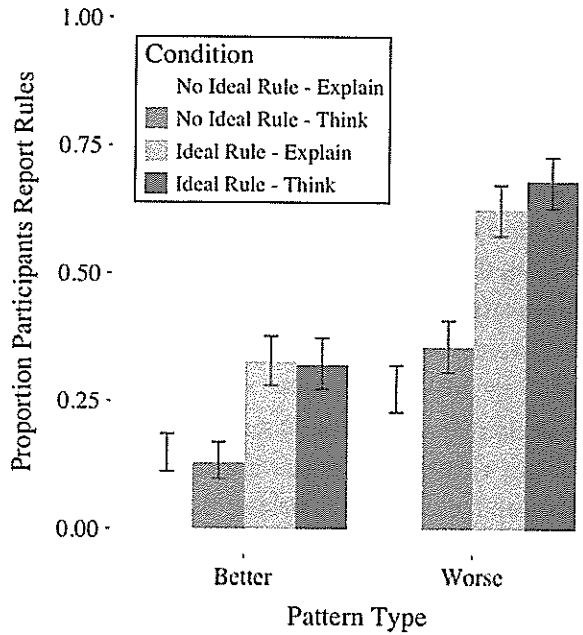

Figure 2.3 The proportion of participants reporting each rule in Experiment 2, as a function of rule type, condition, and prompt. Error bars correspond to 95 percent confidence intervals.

by *experiment number* (1 vs. 2) × *prompt-type* (explain vs. write thoughts) × *pattern-type* (*ideal rule* vs. *no ideal rule*) × *stimulus-type* (flowers vs. containers vs. objects vs. simple robots vs. complex robots). The goal of doing so was to identify whether changing the presentation format (as we did in Experiment 2) had a differential effect on explainers versus participants in the control group. We found a significant interaction between experiment number and prompt-type ($\chi^2 = 10.83$, $p < 0.01$). This interaction suggests that the change in presentation format did in fact impact explainers and control participants differently. To explore the nature of this difference, we ran additional logistic regressions for explainers and control participants separately. The change in presentation format between Experiments 1 and 2 only had a significant effect on better rule discovery for control participants ($\chi^2 = 7.09$, $p = 0.01$) and not for explainers ($\chi^2 = 1.76$, $p = 0.18$). It therefore appears that the changes to the presentation format in Experiment 2 (simultaneous presentation in separated categories) allowed control participants to perform more like explainers, perhaps by making the need to "go beyond the obvious" less effortful.

Worse rule reporting. To test whether prompt-type influenced the discovery of the worse rule, we ran another logistic regression predicting whether participants *discovered the worse rule* (yes vs. no) by *prompt-type* (explain vs. write thoughts) × *pattern-type* (*ideal rule* vs. *no ideal rule*) × *stimulus-type* (flowers vs. containers vs. objects vs. simple robots vs. complex robots). There was a significant effect of prompt-type ($\chi^2 = 6.60$, $p = 0.01$), suggesting that explaining *inhibited* discovery of the worse rules in this task (see Figure 2.3). There was also a significant effect of pattern-type ($\chi^2 = 173.59$, $p < 0.01$), with more frequent discovery of the worse pattern when it accounted for more cases.[14] Unlike Experiment 1, this suggests that explaining can result in lower detection or reporting of regularities that are superseded by better alternatives.

Discussion

Comparing the results of Experiment 1 to those of Experiment 2 suggests that when a learning environment makes it easier to recognize the flaws of a suboptimal pattern and to identify the features that support a better alternative, control participants (i.e., those who are *not* prompted to explain) receive a disproportionate benefit. By contrast, participants who are prompted to explain more often succeed in going beyond the obvious to find a better but more subtle pattern *whether or not* a learning environment makes it easy to do so. These findings reinforce a lesson from prior research: that the magnitude of the benefits of explanation (relative to control conditions) can be moderated by a variety of factors (Kon and Lombrozo, in prep; Williams, Lombrozo, and Rehder, 2013). But they also go beyond prior research in finding that task difficulty may be one of these moderating factors. Generalizing to science, we might expect the drive for explanation to have an especially pronounced effect on scientific discovery when the dominant ways of representing the relevant data or phenomena do not already support the alignment of relevant features or comparisons across relevant distinctions.

General discussion

Across two studies, we find support for a potential resolution to the puzzle with which we began. On the one hand, scientists are often driven to achieve an explanatory ideal with a prominent role for exceptionless generalizations and theories that support simple explanations. On the other hand, regularities in the natural world quite often have exceptions, and simple explanations are not always forthcoming. Our findings suggest that the process of seeking ideal explanations may be beneficial because it supports discovery, and that these beneficial effects on discovery are not restricted to the ideal case; explaining can facilitate the discovery of subtle patterns even when those patterns do not account for all cases. This finding is broadly consistent with the idea of "Explaining for the Best Inference" (EBI) introduced by Wilkenfeld and Lombrozo (2015); the *process* of seeking explanations can sometimes be beneficial because it has positive downstream consequences on what we learn and infer.

Needless to say, our artificial learning tasks are a poor match to real scientific practice, and our classification rules are a poor match to rich scientific explanations. The research we review and present here is no substitute for naturalistic studies of real scientific advance. That said, we expect the learning mechanisms documented here to apply quite broadly. For example, findings concerning the effects of explanation in artificial classification tasks (Williams and Lombrozo, 2010) have been replicated with property-generalization tasks that involve meaningful causal explanations (Kon and Lombrozo, in prep). The core phenomena found with adults have also been successfully replicated with preschool-aged children (Walker et al., 2014; Walker, Bonawitz and Lombrozo, 2017). These findings suggest that effects of engaging in explanation are fairly widespread and baked into our explanatory activities from a young age. While

science undoubtedly involves a refinement of these widespread explanatory tendencies, we expect a great deal of continuity to be maintained nonetheless.

Beyond the lack of scientific realism, other limitations of these studies should be acknowledged. Our participant pool was restricted to online participants within the United States, our learning tasks occurred over a short time scale, and participants were almost certainly more motivated to receive their pay than to uncover the structure of our artificial worlds. Moving forward, it will be important to pursue research that preserves the experimental control of the studies we present here while simultaneously overcoming these limitations.

Zooming out, our findings support a functionalist approach to scientific explanation (Lombrozo, 2011). On this view, explanation is crucial to science because it serves an instrumental role. By pursuing explanations of the natural world, we're more likely to generate discoveries and develop theories that in turn improve our strategic position in the world, satisfying Hempel's first motivation for science by pursuing the second.

Appendix A: Better and worse rules for all stimulus types

- Flowers
  - 100 percent ideal & 83 percent no ideal—SOMP flowers have two concentric circles in their centers; THONT flowers have one circle in their center.
  - 83 percent ideal & 66 percent no ideal—The petals of SOMP flowers are mostly one color; the petals of THONT flowers are mostly rainbow-colored.

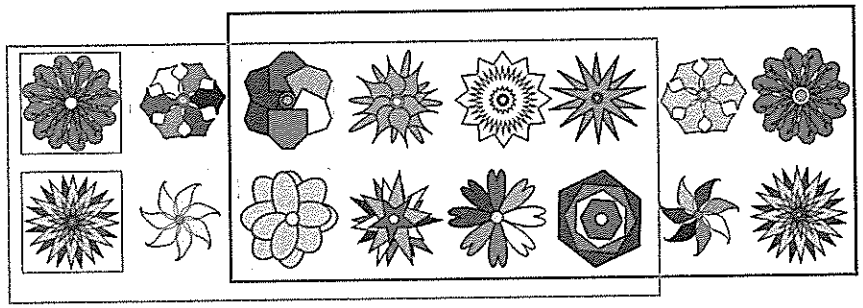

Figure 2.4 Flowers.

- Containers
  - 100 percent ideal & 83 percent no ideal—ANDRAK containers rest on platforms that are larger than their openings; ORDEEP containers rest on platforms that are smaller than their openings.
  - 83 percent ideal & 66 percent no ideal—ANDRAK containers were mostly tall and narrow; ORDEEP containers were mostly short and wide.

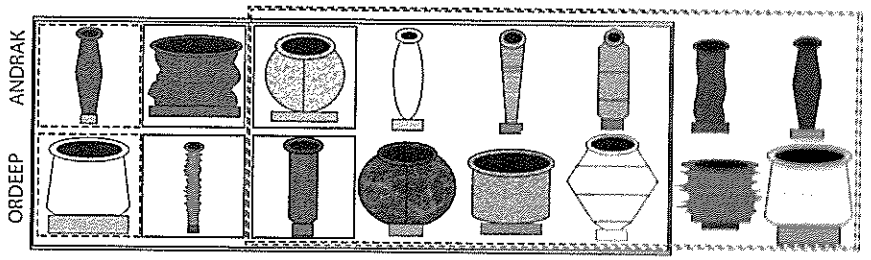

Figure 2.5 Containers.

- Objects
  - 100 percent ideal & 83 percent no ideal—TRING objects have their larger portion on the top; KRAND objects have their larger portion on the bottom.
  - 83 percent ideal & 66 percent no ideal—TRING objects mostly have vertical lines dividing their sections with and without dots; KRAND objects mostly have diagonal lines dividing their sections with and without dots.

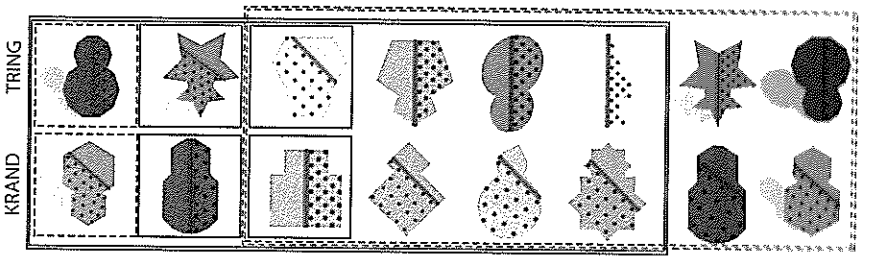

Figure 2.6 Objects.

- Simple robots
  - 100 percent ideal & 83 percent no ideal—the bottom of the DRENT robots' feet were flat; the bottom of the GLORP robots' feet were pointed.
  - 83 percent ideal & 66 percent no ideal—the bodies of most DRENT robots were round; the bodies of most GLORP robots were square.

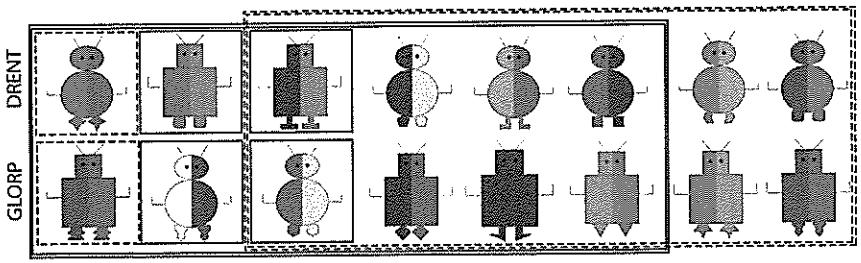

Figure 2.7 Simple robots.

- Complex robots
  - 100 percent ideal & 83 percent no ideal—the bottom of the DRENT robots' feet were flat; the bottom of the GLORP robots' feet were pointed.
  - 83 percent ideal & 66 percent no ideal—
    1) The bodies of most DRENT robots are round, and the bodies of most GLORP robots are square.
    2) The antennae of most DRENT robots are curled, and the antennae of most GLORP robots are straight.
    3) Most DRENT robots have stars on their hands, and most GLORP robots have nothing at the ends of their arms.
    4) Most DRENT robots have no dot on their chest, and most GLORP robots have a chest dot.
    5) Most DRENT robots have their colored sections split evenly down the middle, and most GLORP robots have a checkered pattern.

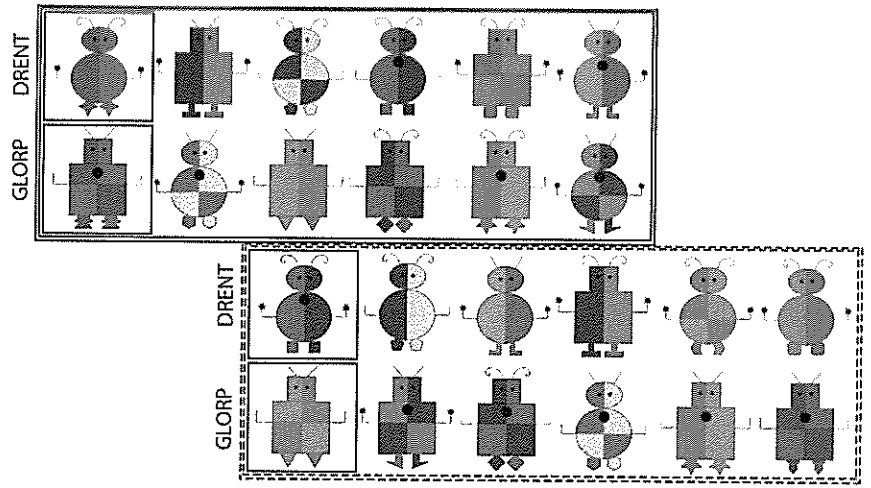

Figure 2.8 Complex robots.

Appendix B: Rated rules

- Better rule:
  - Because SOMP flowers have two circles in their centers, and THONT flowers have one circle in their centers.
  - Because ANDRAK containers rest on platforms that are larger than their openings, and ORDEEP containers rest on platforms that are smaller than their openings.
  - Because TRING objects have a larger shape on top of a smaller shape, and KRAND objects have a smaller shape on top of a larger shape.
  - Because GLORP robots have feet that are pointy along the bottom, and DRENT robots have feet that are flat along the bottom.
- Worse rule:
  - Because SOMP flowers have petals which are all the same color, and THONT flowers have petals with many colors.
  - Because ANDRAK containers have thin bodies, and ORDEEP containers have wide bodies.

  - Because TRING objects have a vertical line dividing their sections with and without dots, and KRAND objects have a diagonal line dividing their sections with and without dots.
  - Because GLORP robots have square-shaped bodies, and DRENT robots have circular bodies.
- Worse rule + exception:
  - Because SOMP flowers have petals which are all the same color, and THONT flowers have petals with many colors, though there are some exceptions.
  - Because ANDRAK containers have thin bodies, and ORDEEP containers have wide bodies, though there are some exceptions.
  - Because TRING objects have a vertical line dividing their sections with and without dots, and KRAND objects have a diagonal line dividing their sections with and without dots.
  - Because GLORP robots have square-shaped bodies, and DRENT robots have circular bodies, though there are some exceptions.

Notes

1 Some historical analyses do aim to chart psychological aspects of scientific discovery (e.g., Gentner et al., 1997), use observational/ethnographic methods with qualitative and quantitative analyses to better understand scientific practice (e.g., Dunbar, 1997), or consider how science works from a cognitive scientific perspective (e.g., Proctor and Capaldi, 2012; Thagard, 2012). To our knowledge, however, this research has not focused on how psychological features of our drive for explanations affect scientific advance.

2 It's worth clarifying a feature of our nomenclature regarding classification rules. In labeling a rule as 66 percent, 83 percent, or 100 percent, we are highlighting the percentage of training items captured by an unqualified generalization (e.g., "Thont flowers have a single circle in their centers"); we do not intend the percentage to be built-in to form a probabilistic classification rule (e.g., "Thont flowers have a single circle in their centers with 66 percent probability"), in which case the "exception" items would arguably still fall under the generalization.

3 The mean age of participants was 34 (SD = 11, min = 18, max = 74); 510 participants identified as male and 857 as female. Initially, we collected a sample of 1309 participants (before exclusions); however, when we analyzed these responses, the results were inconclusive, as we explain in footnote 9. We therefore collected data from additional participants. Analyses correspond to the full sample, but the patterns were the same within each subsample.

4 An additional 1007 participants failed attention or memory checks (see footnote 6) and were therefore excluded from analyses. We indicate any cases in which these exclusions affect the statistical significance of results.

5 Data on the complex robot stimuli were collected separately from the other four stimulus-types, and the stimuli and procedure varied slightly. Specifically, the complex robot stimuli contained five equally good "worse" rules (rather than only one) in addition to one better rule, and participants did not complete the rule rating phase. We combine the data here because the experimental questions and results were the same.

6 Specifically, participants were told "we're interested in any patterns that you noticed that might help differentiate SOMPS and THONTS. For example, did most or all of the SOMPS you studied tend to have one property, and most or all of the THONTS you studied have another property? We're going to ask you to list all of the patterns (differences between SOMPS and THONTS) that you noticed, one at a time. PLEASE REPORT ANY PATTERNS THAT YOU NOTICED, EVEN IF THEY WEREN'T PERFECT AND EVEN IF YOU DON'T THINK THEY'RE IMPORTANT." This language, adapted from Edwards, Williams, and Lombrozo (2013), was employed to encourage participants to report the worse rule (83 percent in the *ideal rule* condition, and 66 percent in the *no ideal rule* condition) even if they thought it was incidental or superseded by the better rule (100 percent in the *ideal rule* condition, and 83 percent in the *no ideal rule* condition).

7 This consisted of a fairly long passage that asked them to select "None of these objects look familiar" and to write in the category of the item they recognized. Specifically, it said "look at the following images and select the one that you have studied in previous questions. In the text box next to that image, please also type in whether you think that it is a [category 1] or a [category 2]. It is important for us to know whether our participants are paying attention and are reading all of the instructions, so if you are reading this, what we actually want you to do is to select 'None of these objects look familiar,' and in the corresponding text box to write in whether the image you recognize from the other options is a [category 1] or a [category 2]." By selecting the instructed button, participants indicated they had been reading instructions, and by correctly reporting the category of the item they recognized, participants indicated that they attended to the stimuli in the primary task.

8 There was also a significant main effect of stimulus-type ($\chi^2 = 112.98$, $p < 0.01$), and a significant interaction between pattern-type and stimulus-type ($\chi^2 = 13.72$, $p < 0.01$).

9 We initially collected a smaller sample size, but the statistical analyses were inconclusive. Specifically, we found the expected effect of explanation (with more participants reporting the better rule when prompted to explain), but we also (a) failed to find an interaction between pattern-type and prompt-type,

suggesting that the effects of explanation were comparable across the *ideal* and *no ideal* rule conditions, and (b) failed to find a significant effect of the explanation prompt when restricting analysis to the *no ideal rule* condition, suggesting that explanation did *not* have an effect under these conditions. Because (a) and (b) supported different conclusions, we decided to collect additional data. It is worth noting that while increasing the sample size did change the statistical significance of the effect of explanation within the *no ideal rule* condition, the proportions of participants reporting the rules remained fairly unchanged by the increased sample size (approximately 15 percent of the explainers reported the imperfect better rule in both the initial and increased sample, approximately 10 percent of control participants reported the imperfect better rule in the initial sample, and approximately 9 percent reported it in the increased sample). This suggests that the initial sample was simply underpowered.

10 The effect of stimulus-type was also significant ($\chi^2 = 57.08$, $p < 0.01$); no interactions were significant (without exclusion criteria, the interaction between pattern-type and stimulus-type was significant ($\chi^2 = 13.79$, $p = 0.01$)).

11 Without exclusion criteria, there is also a significant interaction between prompt-type and rule rated ($\chi^2 = 4.00$, $p = 0.02$)

12 An additional 1048 participants failed attention or memory checks and were therefore excluded from analyses. The statistical significance of results are unchanged unless noted when these participants are included. The mean age of participants was 35 ($SD = 11$, min = 18, max = 79); 1020 participants identified as male and 1487 as female. As in Experiment 1, initially a smaller sample was collected (949), but this was increased to keep approximately the same number of participants across the two experiments.

13 There was also a significant effect of stimulus-type ($\chi^2 = 104.18$, $p < 0.01$). No interaction was significant (without exclusion criteria, there was a significant interaction between pattern-type and stimulus-type ($\chi^2 = 10.24$, $p = 0.04$)).

14 There was also a significant effect of stimulus-type ($\chi^2 = 59.44$, $p < 0.01$). No interaction was significant (without exclusion criteria, there was also a significant interaction between prompt-type and stimulus-type ($\chi^2 = 13.60$, $p = 0.01$)).

References

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in Instructional Psychology*, 5: 161–238.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., and Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13: 145–82.

Chin-Parker, S. and Bradner, A. (2017). A contrastive account of explanation generation. *Psychonomic Bulletin and Review*, 24(5): 1387–97.

Dunbar, K. (1997). How scientists think: Online creativity and conceptual change in science. In T. B. Ward, S. M. Smith and S. Vaid (Eds.), *Conceptual Structures and Processes: Emergence Discovery and Change* (pp. 461–93). Washington: American Psychological Association.

Edwards, B. J., Williams, J. J., and Lombrozo, T. (2013). Effects of explanation and comparison on category learning. In M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 406–11). Austin: Cognitive Science Society.

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

Fonseca, B. and Chi, M. T. H. (2011). Instruction based on self-explanation. In R. Mayer and R. Alexander (Eds.), *Handbook of Research on Learning and Instruction*. New York: Routledge.

Gentner, D., Brem, S., Ferguson, R. W., Markman, A. B., Levidow, B. B., Wolff, P., and Forbus, K. D. (1997). Analogical reasoning and conceptual change: A case study of Johannes Kepler. *The Journal of the Learning Sciences*, 6(1): 3–40.

Harré, R. (1985). *The Philosophies of Science*. Oxford: Oxford University Press.

Hempel, C. (1962). Explanation in science and history. In R. G. Colodny (Ed.), *Frontiers of Science and Philosophy* (pp. 7–33). London: Allen & Unwin.

Kon, E. and Lombrozo, T. (in prep). Why explainers take exception to exceptions.

Lassaline, M. E. and Murphy, G. L. (1998). Alignment and category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(1): 144–60.

Legare, C. H. and Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, 126: 198–212.

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass*, 6: 539–51.

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak and R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning*. Oxford: Oxford University Press.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Science*, 20(10): 748–59.

Meagher, B. J., Carvalho, P. F., Goldstone, R. L., and Nosofsky, R. M. (2017). Organized simultaneous displays facilitate learning of complex natural science categories. *Psychonomic Bulletin and Review*, 24(6): 1987–94.

Pitt, J. C. (1988). *Theories of Explanation*. Oxford: Oxford University Press.

Proctor, R. W. and Capaldi, E. J. (Eds.) (2012). *Psychology of Science: Implicit and Explicit Processes*. Oxford: Oxford University Press.

Rozenblit, L. and Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5): 521–62.

Thagard, P. (2012). *The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change*. Cambridge: MIT Press.

Walker, C. M., Bonawitz, E., and Lombrozo, T. (2017). Effects of explaining on children's preference for simpler hypotheses. *Psychonomic Bulletin and Review*, 24(5): 1538–47.

Walker, C.M., Lombrozo, T., Legare, C., and Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, 133: 343–57.

Wilkenfeld, D. A. and Lombrozo, T. (2015). Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science and Education*, 24(9–10): 1059–77.

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

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

Williams, J. J., Lombrozo, T., and Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, 142(4): 1006–14.