Running head: EXPLANATION AND PRIOR KNOWLEDGE

Explanation and Prior Knowledge Interact to Guide Learning

Joseph J. Williams

Tania Lombrozo

University of California at Berkeley

Author Note

Joseph J. Williams, Department of Psychology, University of California at Berkeley

Tania Lombrozo, Department of Psychology, University of California at Berkeley

Correspondence concerning this article should be addressed to Joseph Williams, Department of Psychology, University of California at Berkeley, 3210 Tolman Hall, Berkeley, CA, 94709.

Contact: joseph_williams@berkeley.edu

Word count: 15 503

Phone: 510-501-7884

Fax: 510-642-5293

Abstract

How do explaining and prior knowledge contribute to learning? Four experiments explored the relationship between explanation and prior knowledge in category learning. The experiments independently manipulated whether participants were prompted to explain the category membership of study observations and whether category labels were informative in allowing participants to relate prior knowledge to patterns underlying category membership. The experiments revealed a superadditive interaction between explanation and informative labels, with explainers who received informative labels most likely to discover (Experiments 1 & 2) and generalize (Experiments 3 & 4) a pattern consistent with prior knowledge. However, explainers were no more likely than controls to discover multiple patterns (Experiments 1 & 2), indicating that effects of explanation are relatively targeted. We suggest that explanation recruits prior knowledge to assess whether candidate patterns are likely to have broad scope (i.e., to generalize within and beyond study observations). This interpretation is supported by the finding that effects of explanation on prior knowledge were attenuated when learners believed prior knowledge was irrelevant to generalizing category membership (Experiment 4). This research provides evidence that explanation can serve as a mechanism for deploying prior knowledge to assess the scope of observed patterns.

Keywords: explanation, self-explanation, prior knowledge, learning, generalization, category learning

Explanation and Prior Knowledge Interact to Guide Learning

1. Introduction

Children, adults, and students of all ages face the common challenge of discovering useful information and then generalizing it to novel contexts. While learning and generalization engage a variety of cognitive processes, researchers across several fields have recognized an important role for explanation (Lombrozo, 2012). For example, prompting young children to explain observations that challenge their intuitive theories can accelerate conceptual development (e.g., Amsterlaw & Wellman, 2006; Siegler, 1995), and prompting students to explain why a fact is true or why a solution to a problem is correct can improve both learning and transfer to novel problems (e.g., Chi et al., 1994). How and why does explaining have these effects? In particular, how does explaining guide discovery and generalization?

We propose that explaining recruits a set of criteria for what constitutes a good explanation, and that these criteria in turn act as constraints on learning and generalization (Lombrozo, 2012). For example, explanations are typically judged better if they are simple (Lombrozo, 2007; Read & Marcus-Newhall, 1993) and have what we refer to as broad *scope* – appealing to features, principles, or patterns that accurately apply to numerous instances across a range of contexts (Pennington & Hastie, 1992; Preston & Epley, 1995; Read & Marcus-Newhall, 1993). In this paper we focus on scope to consider whether the act of generating explanations makes learners more likely to discover and generalize patterns with broad scope. For example, in trying to explain why peafowl at the zoo vary in color, one might discover that males (peacocks) tend to be colorful while females (peahens) tend to be drab. This discovery and the reasoning behind it could in turn support inferences about unobserved peafowl, such as the generalization that all male and female peafowl are likely to conform to this pattern, and not just the particular

species observed at the zoo.

The idea that explaining makes learners more sensitive to scope predicts that explaining should increase the extent to which learners consult prior knowledge.[1]

Learning poses a challenging inductive problem, and prior knowledge can serve as an important cue to which patterns are likely to have broad scope. For example, an explanation for variation in peafowl coloration that appeals to a generalization over sex (males versus females) could be preferred over one formulated over size (larger versus smaller) because prior knowledge favors the former as more likely to generalize beyond the peahen sample observed. So if explaining changes the criteria that learners adopt in generating or evaluating hypotheses by leading them to privilege patterns with broad scope, then explaining should recruit prior knowledge in evaluating the scope of candidate patterns**.** In addition to testing this prediction, we consider whether such an effect (if found) results from a special relationship between explanation and prior knowledge or instead from a more general effect, such as a global increase in how much information explainers discover and retain.

By focusing on the relationship between explanation and prior knowledge, we gain unique leverage in addressing two important questions in cognitive science: how explanation impacts learning and generalization, and when and how prior knowledge is brought to bear on learning. In addition to bridging research on explanation and prior knowledge, we bridge two research traditions by examining questions about explanation and learning (typically studied by educational psychologists) in the context of artificial category learning (typically studied by cognitive psychologists). In the remainder of the introduction we briefly review past work from each of these traditions before presenting the key theory, questions, and predictions that motivate

[1] Throughout the paper we use the term "prior knowledge" to indicate a learner's beliefs or commitments, whether or not they are true. That is, our use of the term "knowledge" is non-factive.

the four experiments that follow.

1.1. General and Selective Effects of Explanation on Learning

Research in education has investigated the role of explanation in learning in the context of the "self-explanation effect": the phenomenon whereby explaining, even to oneself, can improve learning. Effects of self-explanation have been documented in domains from biology to mathematics, from elementary school through university, and under a variety of methods for eliciting explanations (e.g., Aleven & Koedinger, 2002; Chi et al., 1989; Chi et al., 1994; Crowley & Siegler, 1998; Graesser et al., 1994; Nokes et al., in press; Renkl, 1997; Rittle-Johnson, 2006; Siegler, 2002). This diversity is matched by a wide range of proposals concerning how explanation affects learning. For example, a prompt to explain could encourage the generation of inferences and invention of procedures (e.g., Chi et al., 1994; Renkl, 1997; Rittle-Johnson, 2006), boost metacognitive monitoring and help identify gaps in comprehension (e.g., Chi et al, 1989; Nokes et al, in press; Palinscar & Brown, 1984), and/or promote the revision of beliefs and strategies (e.g., Chi et al., 1994; Chi, 2000; Legare, Gelman, & Wellman, 2010; Siegler, 2002; Rittle-Johnson, 2006).

Many of these accounts are compatible with the idea that explaining effectively increases the same kind of cognitive processing that occurs in the absence of explanation. For example, some effects of explanation are attributed to an increase in learners' attention, motivation, or processing time (e.g., Siegler, 2002), and one recent review of research on self-explanation proposes that explanation improves learning because it is a constructive activity, and that equivalently constructive activities have comparable effects (Chi, 2010). While explaining could be especially well-suited to increasing attention, engagement, or some other cognitive resource,

the outcome of such an increase is likely to be "general" in the sense that it extends to many kinds of learning and is not selectively tuned to properties of explanation.

A complementary approach is to focus on effects of explanation that are more "selective" in the sense that they derive from particular properties of explanation and have more targeted consequences. For example, research suggests that explaining encourages young children to focus on causal mechanisms at the expense of memory for color (Legare & Lombrozo, under review), and asking middle-school children to explain leads them to privilege causal hypotheses at the expense of observed covariation (Kuhn & Katz, 2009). Studies with adults additionally find that explaining worked examples can foster detailed verbal elaboration of concepts at the expense of procedural knowledge (Berthold et al., 2011) and promote insight problem solving at the expense of memory for what was studied (Needham & Begg, 1991). These examples indicate that explanation is not merely neutral with respect to some kinds of learning, such as memory for observed examples, but can even be harmful.

Of course, explaining is likely to have both relatively general and more selective effects, and the difference is potentially one of degree rather than kind. Nonetheless, the distinction is useful in motivating a set of questions and analyses that allow us to more precisely specify how and why explanation is selective in the way that it is. For example, explaining could improve students' learning by increasing general engagement, but in particular engage learners in searching for underlying patterns. More generally, selective effects can clarify how and why explaining helps learning by identifying *what* people are more engaged in, *which* beliefs are revised, what *kinds* of inferences are generated, and so on. Our goal in this paper is to more precisely specify what the effects of explanation are and why it is that explaining, in particular, produces those effects. Identifying selective effects of explanation – cases in which explanation

impacts some kinds of learning but not others – is a useful strategy for doing so. In the experiments that follow, we therefore include more than one measure of learning, where we predict effects of explanation for some measures but not for others.

1.2. Prior Knowledge and Explanation in Learning

Only a few studies in educational settings have directly investigated the relationship between explanation, prior knowledge, and learning. These studies have examined how the efficacy of explanation prompts is influenced by a learner's level of prior knowledge about the topic being learned. However, findings have been mixed (e.g., Best, Ozuru, & McNamara, 2004; Chi et al., 1994; Chi & VanLehn, 1991; McNamara, 2004; Renkl et al., 1998; Wong, Lawson & Keeves, 2002). One challenge for interpreting these inconsistent findings is the variation in how different studies assess and operationalize prior knowledge, explanation, and learning. Moreover, they rely on existing variation in learners' knowledge, rather than using experimental manipulations that can more clearly isolate causal relationships between prior knowledge and learning.

Taking a complementary approach to education research, a sizeable literature in cognitive psychology has investigated effects of prior knowledge on learning by experimentally manipulating a learner's prior knowledge concerning artificial categories that are learned in the context of well-controlled laboratory tasks (e.g., Heit, 2001; for reviews see Murphy, 2002; Ross, Taylor, Middleton, & Nokes, 2008; Wattenmaker et al., 1986; Wisniewski, 1995). Within this tradition, prior knowledge has typically been shown to facilitate learning (although see Murphy & Wisniewski, 1991), increase the rate at which novel categories are learned (e.g., Kaplan & Murphy, 1999), decrease prediction errors during learning (e.g., Heit & Bott, 2001), and make it possible for learners to acquire categories with a complex relational structure

(Rehder & Ross, 2001). For example, Murphy and Allopenna (1994) had participants learn novel categories that either grouped features relevant to being a "space building" or an "underwater building" or scrambled these features across categories. Participants in the former condition learned the categories more quickly and were more accurate in reporting the frequency with which different features appeared in each category.

How might explanation affect whether and how prior knowledge influences category learning? Prominent theories of conceptual representation accord a central role to "explanatory beliefs" (Carey, 1985; Murphy & Medin, 1985), a phrase that is often used synonymously with a learner's prior knowledge (see also Ahn, 1998; Lombrozo, 2009; Rehder, 2003; Rips, 1989). However, research in these traditions has overwhelmingly focused on explanations as the outcome of learning, and not on the process of explaining as itself a mechanism for concept acquisition and revision. In fact, only one study (to our knowledge) has experimentally manipulated whether participants explained during category learning (Chin-Parker, Hernandez, and Matens, 2006). The study found that participants who explained were more successful than those who did not in learning diagnostic features of category membership that could be related to prior knowledge, but additionally learned arbitrary diagnostic features – consistent with the idea that explanation recruits prior knowledge through mechanisms with either general or selective effects. No studies (to our knowledge) have manipulated both whether learners explain and the extent or nature of their prior knowledge to directly investigate how explanation and prior knowledge interact.

1.3. Explanation and Prior Knowledge: A Subsumptive Constraints Account

We propose a *subsumptive constraints* account of the relationship between explanation and prior knowledge in learning and test this account using the experimental methods of research

on category learning. Our predictions follow from a commitment to what constitutes an explanation: To be explanatory, explanations must explicitly or implicitly appeal to a pattern or generalization of which the explanandum (what is being explained) is an instance. This idea is motivated by "subsumption" and "unification" theories of explanation in philosophy of science, according to which explanations subsume the explanandum under a law or explanatory pattern, and in so doing ideally unify disparate observations or phenomena under that law or pattern (Friedman, 1974; Kitcher, 1981; 1989; see Woodward, 2010 for review). In the context of everyday judgments, subsuming patterns can take the form of rules, causal relationships, or principles, among others. For example, explaining an object's membership in one category rather than another could appeal to a rule concerning membership (e.g., "avocados are fruits rather than vegetables, because fruits contain the seed of their plant while vegetables do not"), explaining why someone has a particular characteristic could appeal to a causal regularity (e.g., "Anna is politically savvy because she comes from a family of activists"), and explaining the solution to a problem could appeal to a general principle (e.g., "The desired angle must be 30 degrees, because the sum of angles in a triangle is 180"). As a consequence, explaining will drive learners to seek underlying patterns, which then serve to guide learning and generalization. For example, in explaining why your friend Anna is so politically informed, you might note that she comes from a family of activists, and induce the general pattern that people who are raised by activists tend to be politically informed.

According to this account, explanations should be better to the extent that the patterns they invoke unify or subsume a large number of cases and are violated by few exceptions. Explaining should accordingly drive learners to seek patterns that match the greatest proportion of cases to which they can be applied. We refer to the number of (observed and unobserved)

cases to which a pattern successfully applies as its "scope." Because a pattern's scope is rarely directly available, it must be inferred on the basis of several cues, including how many of the currently observed cases fall under the pattern, the proportion of cases from past experience to which it has successfully applied, and more generally, any prior knowledge that can inform inferences about the pattern's likely extension. If the subsumptive constraints account is correct, then explaining should not only make learners more likely to discover patterns, but also influence *which* patterns are discovered, with prior knowledge especially likely to be consulted as explainers evaluate the scope of candidate patterns. This generates the prediction that explaining will interact with prior knowledge relevant to assessing scope to guide discovery and generalization. Specifically, learners who are prompted to explain should consult prior knowledge to a greater degree than those who learn without explaining, and prompts to explain should accordingly have a targeted impact on measures of learning that track prior knowledge and scope, but not necessarily other measures of learning, such as the total number of patterns discovered or recalled. In contrast, if explanation's primary effects are instead to increase attention, motivation, or even the overall search for underlying patterns, the effects of explanation and prior knowledge could be independent, and also generate more widespread consequences for learning.

Williams and Lombrozo (2010) first proposed the *subsumptive constraints* account and reported evidence consistent with the idea that explanation drives learners towards patterns with broader scope. Participants learned about two categories of robots and were prompted to either explain the category membership of eight labeled examples or to engage in a control task, such as description or thinking aloud. Across three experiments, explaining promoted the discovery of a subtle pattern relating foot shape to category membership (i.e., that "Glorp" robots have pointy

feet and "Drent" robots have flat feet), which accounted for the membership of every study observation. In the control conditions participants tended to discover a more salient pattern concerning body shape (i.e., that "Glorp" robots are typically square and "Drent" robots are typically round) that had lower scope (i.e., it only accounted for six of the eight examples) or to encode specific properties of the examples, such as their color. These findings provide initial evidence that seeking explanations promotes the discovery of patterns, and is consistent with the prediction that explaining favors patterns that account for a larger proportion of cases – in these experiments, eight out of eight observations as opposed to six out of eight. However, the experiments were not designed to test the broader issues of interest here concerning the role of prior knowledge in learning or the selectivity of explanation's effects.

In the four experiments reported below, we test the broader implications of the subsumptive constraints account. Specifically, we aim to address the following key questions. First, does explaining make learners more likely to consult prior knowledge in learning, and therefore to discover and generalize patterns consistent with prior knowledge? If so, is this the result of a general effect (e.g., boosting attention or the discovery of all kinds of patterns) or a selective effect (e.g., a constraint on *which* patterns are discovered)? And second, does explanation's selectivity in part derive from the evaluation of the scope of candidate patterns, as our account implies?

2. Overview of experiments

To investigate whether and how explanation and prior knowledge interact to guide learning and generalization, we presented participants with a category learning task in which we manipulated both the extent to which learners explained and their ability to recruit relevant prior knowledge. We accomplished the former by prompting some participants to explain the category

membership of category exemplars and others to engage in a control task (either free study or writing their thoughts during study). We accomplished the latter by providing category labels that were either "blank" (i.e., nonsense words) or meaningful and potentially relevant to particular category features.

While most research on knowledge effects in category learning has manipulated prior knowledge through the features that make up novel categories (e.g., Murphy & Allopenna, 1994) or with explicit hints about relevant prior knowledge (e.g., Pazzani, 1991; Wattenmaker et al., 1986), the relatively subtle manipulation of category labels has been shown to influence the prior knowledge learners can recruit in learning (e.g., Barsalou, 1985, Wisniewski & Medin, 1994). For example, Wisniewski and Medin (1994) gave participants a set of drawings from children identified as coming from a "creative" versus a "noncreative" group, or from "group 1" versus "group 2," and found that participants constructed different features to discriminate the categories across these conditions. Our experiments used a similar manipulation to influence whether participants could recruit prior knowledge relevant to the learning task.

In the task we employed, participants were presented with category exemplars consistent with multiple patterns, only some of which were knowledge-relevant. For example, participants in all experiments were presented with sample robots from two categories, where those in one category had feet that were flat on the bottom and those in the other had feet that were pointy. The robots also varied across categories in other ways, including (in some experiments) the length of their antennae. When the robots received meaningful labels, such as "indoor robots" versus "outdoor robots," the feature of foot shape was "label-relevant" in that a learner could plausibly relate flat versus pointy feet to use on different indoor versus outdoor surfaces, while a feature such as length of antennae was "label-irrelevant."

With this simple experimental design and appropriate category structures, we examined whether and how explanation and prior-knowledge interacted in the discovery and generalization of patterns underlying category membership. In Experiments 1 and 2, we tested the prediction that prior knowledge is more likely to be recruited to guide discovery when learners engage in explanation. Specifically, we examined how discovery of the label-relevant pattern was influenced by informative labels in the absence of a prompt to explain (control + blank labels versus control + informative labels), and compared this effect to that obtained when learners were prompted to explain (explain + blank labels versus explain + informative labels).

Experiments 1 and 2 additionally considered the mechanisms by which explanation influenced discovery. If explaining increases pattern discovery through a general effect – such as a boost in attention, engagement, or motivation – then effects of explanation would likely extend to multiple measures of learning. In contrast, if explaining influences discovery through a more selective effect, then a prompt to explain could have more targeted consequences. To test the generality of explanation's effects, we examined how a prompt to explain and the provision of informative labels influenced discovery of *more than one* pattern underlying category membership.

Experiments 3 and 4 moved away from discovery to focus on generalization. First, when multiple patterns have been discovered, does explaining make a further contribution in guiding generalization? We predicted the same interaction for pattern generalization as for discovery, with explanation increasing the extent to which learners recruited prior knowledge to guide judgments. In addition, in Experiment 4 we more directly tested our claim that explanation recruits prior knowledge because it informs the assessment of scope.

In sum, the four experiments we present below considered the ways in which explanation and prior knowledge interact to guide learning and generalization. In particular, we considered how both general and selective effects of explanation are influenced by a learner's prior knowledge to better understand the role of explanation in learning and the relationship between explanation and prior knowledge.

3. Experiment 1

Experiment 1 investigated the effect of constructing explanations (task: explain vs. free study) and possessing prior knowledge (label type: blank vs. informative) on discovery of label-relevant and label-irrelevant patterns underlying the category membership of study observations. Participants learned about two categories of alien robots by studying the eight observations shown in Figure 1. After study, novel robots were presented for classification in order to ascertain whether category membership was extended on the basis of the label-relevant pattern, the label-irrelevant pattern, or similarity to a studied observation.

The design independently manipulated task (explain vs. free study) and prior knowledge (blank vs. informative labels) in order to examine the independent and joint effects of explanation and prior knowledge on: (1) the discovery of label-relevant and label-irrelevant patterns; (2) the number of patterns discovered; (3) the relationship between discovering the label-relevant and label-irrelevant pattern; and (4) the use of particular patterns in categorizing novel items. With these varied measures we could evaluate the selectivity of explanation's effects.

3.1. Methods

3.1.1. Participants. Four-hundred-and-seven UC Berkeley undergraduate students participated for course credit or monetary reimbursement.[2]

3.1.2. Materials.

3.1.2.1. Study observations. Participants learned about eight alien robots from two categories, shown in Figure 1a. In the blank labels conditions, the first category was labeled "Glorp robots" and the second "Drent robots," while in the informative labels conditions the first category was labeled "Outdoor robots" and the second "Indoor robots."

The category membership of these eight robots followed two patterns, identified as the label-relevant pattern and the label-irrelevant pattern. The label-relevant pattern was that all four Outdoor (Glorp) robots had pointy feet while all four Indoor (Drent) robots had flat feet. These features were chosen with the assumption that participants could utilize prior knowledge to relate pointy versus flat feet to properties of Outdoor versus Indoor robots.[3] The label-irrelevant pattern was that all four Outdoor (Glorp) robots had a shorter left antenna and all four Indoor (Drent) robots had a shorter right antenna; we expected that participants' prior knowledge would less readily relate relative antenna length to properties of Outdoor versus Indoor robots. Each robot also varied in body shape and in left and right colors, but these features were not diagnostic of category membership as they occurred equally often in each category.

3.1.2.2. Categorization probes. To assess which features participants used in generalizing category membership from the study observations to novel robots, participants classified fifteen

[2] Experiments using related images were previously conducted with this participant pool, so after the study we asked participants if they might have seen the robots before, and excluded an additional 124 participants who responded affirmatively.

[3] In order to verify that participants associated the informative labels with these features, we presented a separate group of participants from the same pool with the individual features of robots from Experiment 2 (see Figure 1b), which contained the features used in all four experiments. Ratings of how important the features were to which category a robot belonged to verified our assumptions: Foot shapes were rated as most important for robots labeled Outdoor/Indoor and antenna shapes as most important for robots labeled Receiver/Transmitter (these labels are used in Exp. 3 & Exp. 4).

unlabeled robots. Participants could categorize these robots in at least three ways. First, participants could discover the label-relevant pattern about feet (pointy vs. flat feet) and categorize new robots based on foot shape. Second, participants could discover the label-irrelevant pattern about antennae (shorter left vs. shorter right antenna) and categorize based on antenna height. Finally, instead of using a pattern, participants could categorize new items on the basis of their similarity to individual study items, where similarity was measured by tallying the number of shared features across items.[4] We refer to these bases for generalizing category membership as "label-relevant pattern," "label-irrelevant pattern," and "item similarity," respectively.

Ten of these novel robots pitted one basis for categorization against the other two and were constructed by taking novel combinations of features from study observations. Specifically, four label-relevant pattern probes yielded one classification according to the label-relevant pattern and another according to both the label-irrelevant pattern and item similarity, with three label-relevant pattern probes and three item similarity probes that likewise isolated a single basis for categorization. Four addition label-relevant transfer probes also pitted the label-relevant pattern against the other two bases for generalization, but used previously unseen foot shapes that conformed to the pointy/flat pattern. Finally, there was one item for which all three bases yielded the same classification. As described later, participants' bases for generalization were inferred from patterns of classifications across these fifteen probes.

3.1.3. Procedure.

[4] We have verified in previous work (Williams & Lombrozo, 2010) that this measure tracks participants' similarity judgments for stimulus materials like those employed in the current experiment.

3.1.3.1. Learning phase. Participants in both the explain and free study conditions were instructed that they would be looking at two types of robots on the planet Zarn and that they would later be tested on their ability to remember and categorize robots.

The eight study observations were shown onscreen for two minutes. The robots were presented in a scrambled order, with category membership and identifying number (1 through 8) clearly indicated for each robot. Participants in the free study conditions were told, "Robots 1, 2, 3 & 4 are Outdoor (Glorp) robots, and robots 5, 6, 7 & 8 are Indoor (Drent) robots." Participants in the explain conditions were told "Explain why robots 1, 2, 3 & 4 might be Outdoor (Glorp) robots, and explain why robots 5, 6, 7 & 8 might be Indoor (Drent) robots." Participants typed their explanations into a box onscreen.

3.1.3.2. Test phase.

3.1.3.2.1. Pattern discovery. For both the label-relevant (foot) pattern and the label-irrelevant (antenna) pattern, participants were asked if they could tell whether a robot was Outdoor (Glorp) or Indoor (Drent) by looking at its feet (antennae), and if they could, to state the difference(s) between categories.

3.1.3.2.2. Basis for categorization. The categorization probes were presented in random order, with participants categorizing each robot as Outdoor (Glorp) or Indoor (Drent).

3.1.3.2.3. Explanation self-report. To examine effects of spontaneous explanation, all participants were asked if they were trying to explain category membership while viewing the eight robots, and responded "Yes," "Maybe," or "No."

3.1.3.2.4. Additional measures. To examine whether being prompted to explain changed participants' assumptions about the likely presence of a pattern, they were asked, "What do you think the chances are that there is one single feature that underlies whether a robot is Outdoor

(Glorp) or Indoor (Drent) - a single feature that could be used to classify ALL robots?" Participants responded on a scale from 0 to 100.

Participants were also asked to report any differences they noticed across categories and used in classification, and to rank the relative importance of each feature (feet, antennae, body, and color) in categorization. These questions were included in case participants reported unanticipated differences between categories, but as this very rarely happened the responses were redundant with the pattern discovery questions, and are not discussed further.

Participants encountered the test measures in the following order: categorization probes, probability of pattern, category differences, discovery of label-irrelevant antenna pattern, explanation self-report, discovery of label-relevant foot pattern.

3.2 Results

3.2.1 Discovery of patterns. On the pattern discovery questions, participants were credited with discovery of the label-relevant (foot) pattern and label-irrelevant (antenna) pattern if they accurately cited the corresponding diagnostic features. The primary coder's reliability was confirmed by agreement of 98% with a second coder's classification of 25% of the responses. Figure 2a reports discovery of the label-relevant and label-irrelevant patterns as a function of task and label type, and illustrates that discovery rates were higher for participants who explained, with the pattern most likely to be discovered dependent on the presence of informative labels.

The effects of task and label type on discovery of the *label-relevant* pattern were explored using a log-linear analysis on task (explain, free study), label type (blank labels, informative labels), and discovery of the label-relevant pattern (discovered, not discovered). This revealed an interaction between task and discovery, $\chi^2(1, N = 407) = 11.65$, $p < 0.01$, with higher

discovery rates for participants who explained, as well as an interaction between label type and discovery, $\chi^2(1, N = 407) = 11.61$, $p < 0.01$, with higher discovery rates for participants who received informative labels. However, these interactions were superseded by a three-way interaction between task, label type, and discovery, $\chi^2(1, N = 407) = 3.98$, $p < 0.05$: Discovery was highest among participants who explained *and* received informative labels. In fact, discovery of the label-relevant pattern was not significantly improved by explaining when blank labels were provided, $\chi^2(1, N = 207) = 1.43$, $p = 0.15$, nor by providing informative labels in free study conditions, $\chi^2(1, N = 200) = 0.55$, $p = 0.52$.

A parallel analysis on discovery of the *label-irrelevant* pattern also revealed a three-way interaction with task and label type, $\chi^2(1, N = 407) = 5.48$, $p < 0.05$, superseding interactions between task and discovery, $\chi^2(1, N = 407) = 17.39$, $p < 0.001$, and label type and discovery, $\chi^2(1, N = 407) = 11.47$, $p < 0.001$. However, this interaction was driven by elevated discovery of the label-irrelevant pattern by participants who explained with blank labels. In fact, explaining with informative labels led to *lower* discovery of the label-irrelevant (antenna) pattern than explaining with blank labels, $\chi^2(1, N = 207) = 17.98$, $p < 0.01$.

These findings suggest that explaining boosts the discovery of patterns underlying category membership, with prior knowledge influencing *which* pattern is discovered. When informative labels were provided, explaining boosted discovery of the label-relevant pattern. When blank labels were provided, explaining boosted discovery of the label-irrelevant pattern.

3.2.2 Number of patterns discovered. Figure 2b indicates the proportion of participants who discovered neither pattern, exactly one pattern, or both the label-relevant and label-irrelevant patterns, and illustrates that participants in the free study conditions overwhelmingly

discovered zero patterns, while those in the explain condition most often discovered exactly one, irrespective of label type.

A log-linear analysis on task (explain, free study), label type (blank, informative), and number of patterns discovered (zero, one, two) revealed interactions between number of patterns discovered and task, $\chi^2(2, N = 407) = 80.97$, $p < 0.001$, as well as between number and label type, $\chi^2(2, N = 407) = 8.53$, $p < 0.05$. We therefore performed three separate log-linear analyses on whether or not a participant had discovered zero, one, or two patterns. Participants prompted to explain were less likely than participants in the free study conditions to discover zero patterns, $\chi^2(1, N = 407) = 71.52$, $p < 0.001$, but more likely to discover exactly one, $\chi^2(1, N = 407) = 74.86$, $p < 0.001$, which was also more likely among participants receiving blank labels, $\chi^2(1, N = 407) = 7.64$, $p < 0.01$. There was no effect of explanation on discovering two patterns, although there was a marginal effect of label type, $\chi^2(1, N = 407) = 3.50$, $p = 0.062$, with informative labels increasing discovery of two patterns.

These results confirm the importance of explaining in pattern discovery, but it is notable that explaining did not boost the discovery of *multiple* patterns, instead driving participants to discover *a* pattern.

3.2.3 Conditional pattern discovery. We additionally examined the discovery rate for one pattern given discovery of the other, which we call "conditional discovery" (see Figure 2c). Log-linear analyses were performed with task and label type crossed against (1) discovery of the label-irrelevant pattern given discovery of the label-relevant pattern (i.e., discovered label-relevant pattern, discovered both patterns) and (2) discovery of the label-relevant pattern given discovery of the label-irrelevant pattern (i.e., discovered label-irrelevant pattern, discovered both patterns).

Among participants who discovered the label-relevant pattern, the probability of *also* discovering the label-irrelevant pattern was *lower* in the explain than free study conditions, as revealed by a task by discovery interaction, $\chi^2(1, N = 42) = 7.10$, $p < 0.01$. And among those who discovered the label-irrelevant pattern, those in the explain conditions were less likely to have *also* discovered the label-relevant pattern, $\chi^2(1, N = 69) = 6.73$, $p < 0.01$. In other words, relative to free study, participants in the explain conditions who discovered either pattern were *less likely* to discover a second pattern. In addition, those in the informative labels conditions who discovered the label-irrelevant pattern were more likely to have also discovered the label-relevant pattern, $\chi^2(1, N = 200) = 11.88$, $p < 0.01$), which was driven primarily by the free study-informative labels condition. No other effects were significant.

These findings reinforce the idea that explaining has selective effects, and even suggest that explaining can *hinder* discovery under some conditions.

3.2.4 Basis for categorization. Participants' basis for generalizing category membership to new robots was inferred from classification of the categorization probes – specifically, whether there were more judgments consistent with use of the label-relevant pattern, the label-irrelevant pattern, or item similarity, with ties coded as 'other.' Table 1 reports the proportion of participants classified as using each basis for categorization.

Effects of explanation were first analyzed with a log-linear test with three factors: task (explain, free study), label type (informative, blank), and basis for categorization (label-relevant pattern, label-irrelevant pattern, item similarity). This analysis revealed interactions between task and basis, $\chi^2(3, N = 407) = 92.02$, $p < 0.0001$, as well as between label type and basis, $\chi^2(3, N = 407) = 17.34$, $p < 0.001$, with a marginal three-way interaction, $\chi^2(3, N = 407) = 6.89$, $p = 0.07$. To interpret these effects we performed log-linear analyses on task, label type, and each

individual basis for categorization (target basis vs. all others). Overall, the results paralleled those for discovery. Explaining interacted with the provision of informative labels to promote use of the label-relevant pattern, $\chi^2(1, N = 407) = 7.27$, $p < 0.01$, superseding the effects of explanation, $\chi^2(1, N = 407) = 4.98$, $p < 0.05$, and prior knowledge, $\chi^2(1, N = 407) = 4.21$, $p < 0.05$. Task and label type also interacted with use of the label-irrelevant pattern, $\chi^2(1, N = 407) = 7.18$, $p < 0.05$, with significant effects of task, $\chi^2(1, N = 407) = 17.39$, $p < 0.001$, and label type, $\chi^2(1, N = 407) = 11.47$, $p < 0.01$. One additional finding of note was that participants in the free study conditions were significantly more likely to generalize category membership by item similarity, $\chi^2(1, N = 407) = 3.90$, $p < 0.05$. No other effects were significant.

These findings mirror those for pattern discovery very closely, and could thus simply reflect the consequences of discovery. Alternatively, they could reflect independent effects of explanation and prior knowledge on how patterns are generalized. Effects of generalization that were *not* attributable to the consequences of discovery could in principle be detected by restricting analyses to just those participants who discovered both patterns. However, discovery of both patterns was sufficiently low to preclude a statistically reliable analysis (log-linear analysis typically requires that there be no fewer than five observations per cell). We revisit this question in Experiments 3 and 4, where we examine effect of explanation on generalization more directly.

3.2.5 Self-reported explanation. Participants were credited with explaining if they answered "yes" to the *explanation self-report* question, resulting in the following rates of self-reported explanation: 65% for free study/blank labels, 88% for explain/blank labels, 58% for free study/informative labels, and 82% for explain/informative labels. A significantly higher proportion of participants reported self-explaining after receiving explain than free study

prompts, $\chi^2(1, N = 407) = 26.79$, $p < 0.001$, although self-reported explanation was still considerable in free study. Label type did not impact self-reported explanation, $\chi^2(1, N = 407) = 1.21$, $p = 0.162$.

To examine the relationship between spontaneous explanation, pattern discovery, and generalization, we replicated the previous analyses, examining only the free study conditions and replacing the variable of "task" with "self-reported explanation." Table 2 reports the data relevant to this analysis. Overall, the pattern of results for self-reported explanation paralleled the previous findings and suggest that spontaneous explanation in the free study condition had similar effects to prompted explanation. Specifically, all two-way interactions from the analyses above (sections 3.2.1 and 3.2.4) reached significance (ps < .01), but the three-way interactions did not.[5] In particular, the key interaction between explanation, label type, and discovery of the label-relevant pattern was not significant ($p = .15$), and that for explanation, label type, and use of the label-relevant pattern as a basis for categorization was marginal ($p = .06$). This could be due to the smaller number of participants and reduced statistical power in these analyses.

3.2.6 Probability of pattern. Judgments of the probability that there was a single pattern underlying the category membership of all robots was (as expected) higher for participants who

[5] Self-reported explanation was related to both discovering the label-relevant pattern, $\chi^2(1, N = 407) = 8.64$, $p < 0.01$, and using it as a basis for categorization; $\chi^2(1, N = 407) = 8.05$, $p < 0.01$. Informative labels similarly increased discovery, $\chi^2(1, N = 407) = 14.05$, $p < 0.01$, and use, $\chi^2(1, N = 407) = 7.10$, $p < 0.01$, of the label-relevant pattern. However, the interaction between self-reported explanation, prior knowledge, and discovery of the label-relevant pattern did not reach significance as it did for the previous analysis of explanation, $\chi^2(1, N = 407) = 2.12$, $p = 0.15$, nor did the interaction for basis use, $\chi^2(1, N = 407) = 3.67$, $p = 0.06$.

The analysis for the label-irrelevant pattern found that self-reported explaining was associated with higher discovery, $\chi^2(1, N = 407) = 16.63$, $p < 0.001$, and use in categorization, $\chi^2(1, N = 407) = 7.84$, $p < 0.01$, and when informative labels were provided both discovery, $\chi^2(1, N = 407) = 10.46$, $p < 0.01$, and use in categorization, $\chi^2(1, N = 407) = 12.52$, $p < 0.01$, were lower. However, the interactions of explanation and informative labels with discovery and use were not significant (discovery: $\chi^2(1, N = 407) = 3.75$, $p = 0.06$; use in generalization: $\chi^2(1, N = 407) = 0.024$, $p = 0.88$).

A third analysis involving the use of item-similarity in generalizing category membership revealed that reliance on item-similarity was *lower* when participants self-reported explaining, $\chi^2(1, N = 407) = 30.71$, $p < 0.01$, replicating the previous findings

discovered a pattern (75%) than those who did not (36%), $t(405) = 12.60$, $p < 0.001$, $d = 1.29$. For participants who did not discover a pattern, a task by label type ANOVA with probability judgments as a dependent variable did not reveal significant effects of label type (blank labels: $M = 32\%$, $SD = 28\%$; informative labels: $M = 41\%$, $SD = 30\%$; $F(1, 150) = 2.68$, $p > 0.10$), or of task (explain: $M = 45\%$, $SD = 31\%$; free study: $M = 34\%$, $SD = 28\%$; $F(1, 150) = 3.40$, $p = 0.07$), suggesting that effects of task on discovery were driven by engaging in explanation, and were not merely the result of task demands, such as inferences about the category structure resulting from the instruction to explain.

3.2.7. Summary. Experiment 1 found that generating explanations interacted with the provision of informative labels to promote discovery of the label-relevant pattern. When blank labels were provided, explaining again interacted with label type, but in promoting discovery of the label-*irrelevant* pattern. In other words, explaining increased the rate at which participants discovered a pattern underlying category membership, but *which* pattern was discovered depended on the kinds of labels presented and their relationship to prior knowledge. These findings were closely mirrored by those concerning participants' bases for generalizing category membership to novel items, with suggestive evidence that spontaneous explanation in the free study conditions produced similar effects.

These findings not only suggest that explaining increases the extent to which participants recruit prior knowledge to guide discovery, but additionally bear on the selectivity of explanation's effects. While explaining increased the rate at which participants discovered one pattern, it had no beneficial effect – and in fact may have hindered – the discovery of a second pattern.

4. Experiment 2

Experiment 2 extended the findings from Experiment 1 in two important ways. First, the experiment compared a prompt to explain to a more demanding control condition: Participants were prompted to type their thoughts onscreen as they studied category members in the learning phase. This tests an alternative interpretation of the findings from Experiment 1: that effects of a prompt to explain resulted from greater engagement, the need to articulate thoughts in language, or some other consequence of generating written text during learning.

Second, to provide a more stringent test of whether explaining in fact fails to influence or even impairs additional discovery beyond a single pattern, we increased the number of additional patterns from one to three. In addition to a label-relevant pattern and a label-irrelevant pattern that accounted for all observations (100% patterns), the study materials included two patterns that accounted for six out of eight observations (75% patterns).

4.1 Methods.

4.1.1 Participants.

Five-hundred-and-fifty-four members of the Amazon Mechanical Turk workplace participated online for monetary compensation. Participation was restricted to users from the United States.

4.1.2 Materials & Procedure.

4.1.2.1. Study observations. Study observations were modified from those in Experiment 1 (see Figure 1) so that body shape (round vs. square) and antenna length were each partially diagnostic of category membership. Each feature accounted for six of eight study observations (75%), generating a 75% body pattern and a 75% antenna pattern, respectively. Foot shape served as a label-relevant pattern that accounted for all observations (100% foot pattern), with arm configuration as a new label-irrelevant pattern for all eight robots (100% arm pattern). The

arms were either matching (both pointing up or down at the same angle) or mismatching (one pointing up and one pointing down).

4.1.2.2. Learning phase. As in Experiment 1, participants studied the image of all eight robots for exactly two minutes, with one group prompted to explain why robots 1-4 might be Outdoor (Glorp) robots and robots 5-8 might be Indoor (Drent) robots, as in Experiment 1. However, in the *write thoughts* control condition, participants received the following prompt: "Write out your thoughts as you study and learn to categorize robots 1, 2, 3, 4 as Outdoor (Glorp) robots and robots 5, 6, 7, 8 as Indoor (Drent) robots." In both conditions participants then typed responses onscreen.

4.1.2.3. Test phase. After study participants were asked whether they could tell which category a robot belonged to by looking at its antennae, arms, body, and/or feet, responding "Yes," "Maybe," or "No." If they indicated "Yes" or "Maybe," they were asked to state how the categories differed.

4.2 Results & Discussion.

4.2.1 Discovery of patterns. Figure 2d indicates the proportion of participants who discovered each of the four patterns as determined by a response of "Yes" or "Maybe" as to whether the corresponding features differed across categories. A log-linear analysis on task (explain, write thoughts), label type (blank, informative) and discovery of the label-relevant pattern (discovered, not discovered) revealed a three-way interaction, $\chi^2(1, N = 554) = 5.31$, $p < 0.05$, which superseded the effects of task, $\chi^2(1, N = 554) = 7.00$, $p < 0.01$, and label type, $\chi^2(1, N = 554) = 8.64$, $p < 0.01$. As in Experiment 1, discovery of the label-relevant pattern was highest when participants explained *and* were provided with informative labels.

Similar log-linear analyses involving task and label type were carried out for the label-irrelevant pattern, the antenna pattern, and the body shape pattern. Blank labels led to greater discovery of the label-irrelevant pattern than informative labels, $\chi^2(1, N = 554) = 5.02$, $p < 0.05$. In addition, discovery of the body shape pattern was higher in the write thoughts than explain conditions, $\chi^2(1, N = 554) = 5.97$, $p < 0.05$. No other effects were significant.

Despite a more demanding control condition, these results replicate the key finding from Experiment 1 that explanation and prior knowledge interact to guide discovery of a label-relevant pattern.

4.2.2 Number of patterns discovered. Figure 2e indicates the proportion of participants who did not discover any patterns, who discovered exactly one pattern, or who discovered multiple patterns (two or more). A log-linear analysis on task (explain, write thoughts), label type (informative, blank), and number of patterns discovered (none, one, multiple) revealed effects of task, $\chi^2(1, N = 554) = 16.22$, $p < 0.01$, and label type, $\chi^2(1, N = 554) = 13.44$, $p < 0.01$, on how many patterns were discovered. The effect of task and label type on each discovery outcome was therefore examined using three further log-linear analyses. Participants in the write thoughts conditions were more likely to fail to discover any patterns, $\chi^2(1, N = 554) = 3.88$, $p < 0.05$, while those in the explain conditions were more likely to discovery exactly one pattern, $\chi^2(1, N = 554) = 9.30$, $p < 0.01$. However, engaging in explanation and writing thoughts did not differ significantly in promoting discovery of multiple patterns, $\chi^2(1, N = 554) = 2.76$, $p = 0.10$. There were no additional significant effects.

4.2.3 Conditional pattern discovery. Figure 2f indicates the probability of having discovered *another* pattern given that the label-relevant pattern or the label-irrelevant pattern was discovered. Given discovery of the label-relevant (foot) pattern, participants in the explain

conditions were less likely to discover additional patterns than those in the control conditions, $\chi^2(1, N = 88) = 6.05$, $p < 0.05$. Similarly, given discovery of the label-irrelevant (arm) pattern, participants in the explain conditions were less likely than control participants to discover additional patterns, $\chi^2(1, N = 203) = 4.56$, $p < 0.05$. There were no other significant effects (all ps > 0.10).

These findings again mirror Experiment 1: A prompt to explain did not boost discovery of additional patterns, and in fact *lowered* the probability that participants would discover another pattern given that either the label-relevant or label-irrelevant pattern was discovered.

4.2.4. Written responses. Because all participants in Experiment 2 were prompted for written responses, we could compare these to see whether the explain and write thoughts conditions were effectively matched in terms of overall engagement and attention to category labels, which should roughly be tracked by response length and mention of category labels, respectively. Some participants left responses blank and are not included in these analyses; The proportion of participants who left items blank did not differ significantly across the explain (15.9%) and the write thoughts conditions (22.2%), $\chi^2(1, N = 554) = 3.55$, $p = 0.06$.

A task by label type ANOVA on the number of words per response revealed that response length did not differ significantly between the explain conditions ($M = 18.1$ words, $SD = 11.0$) and the write thoughts conditions ($M = 19.5$ words, $SD = 12.5$), $F(1, 443) = 1.51$, $p = 0.22$. However, participants wrote more when provided with informative labels ($M = 20.0$ words, $SD = 12.6$) than with blank labels ($M = 17.4$ words, $SD = 10.1$), $F(1, 443) = 5.27$, $p < 0.05$. There were no other significant results.

A log-linear analysis found that the proportion of participants who mentioned one or more category labels was not significantly influenced by explaining versus writing out thoughts

(explain: 64%; write thoughts: 58%; $\chi^2(1, N = 447) = 1.14$, $p = 0.29$). However, for participants in both study conditions, informative labels were mentioned more frequently than blank labels (informative: 67%; blank 55%; $\chi^2(1, N = 447) = 6.78$, $p < 0.01$). These findings make it unlikely that the effects of explanation documented above can be attributed to verbalization, greater engagement with the task, or greater attention to category labels.

4.2.5 Summary. Experiment 2 replicated the key findings from Experiment 1 with a more demanding control condition ("write thoughts") that was well matched in terms of engagement and attention to category labels, and with a more complex category structure involving additional patterns. The findings nonetheless support the claim that explanation increases the extent to which learners consult prior knowledge in learning, and that explanation has relatively selective effects rather than producing a global or all-purpose boost to learning.

5. Experiment 3

Experiments 1 and 2 provide evidence that explaining magnifies the role of prior knowledge in pattern discovery, with additional effects (in Experiment 1) on how patterns are generalized to novel category members. However, this raises the question of whether explanation's role in generalization is simply a consequence of its role in discovery. Does explaining guide generalization directly, even when it confers no advantage for discovery? To address this question we modified the study materials to increase the rate of discovery and to directly evaluate effects of explanation and prior knowledge on generalization when multiple patterns are discovered.

Experiment 3 also went beyond the preceding experiments in three notable ways. First, to more directly assess whether explanation changes the role of prior knowledge in assessing a candidate pattern's scope, the experiment included additional measures of generalization that

corresponded more closely to how broadly a pattern was extended. Participants still classified novel items that pitted patterns against each other, thus tracking the diagnosticity of different features. But Experiment 3 also asked participants how frequently each pattern-related feature occurred in members of each category – a measure of category validity, or the probability of a feature given category membership. This provides an additional and potentially more direct measure of beliefs concerning a pattern's scope than binary classifications. Second, Experiment 3 counterbalanced whether feet or antennae featured in the label-relevant pattern (and therefore what the informative labels were), ensuring that our findings did not result from a unique property of the foot pattern or the Indoor/Outdoor labels. And finally, the study observations were modified to create uncertainty about whether the label-relevant pattern subsumed all of the observed cases, allowing us to assess whether explaining recruits prior knowledge in generalization even when prior knowledge conflicts with an alternative cue to scope: the number of explained examples to which a pattern is known to apply.

5. 1 Methods

5.1.1 Participants. Two-hundred-fifty-eight UC Berkeley undergraduates participated in the lab for course credit and two-hundred-eighty-five members of the Amazon Mechanical Turk workplace from the United States participated online for monetary compensation, yielding a total of 543 participants.

5.1.2 Materials. The adapted robots are shown in Figure 3, and were modified from Experiment 1 to facilitate discovery of the antenna and foot patterns: All members of a given category were given the same feet and antennae shapes, the size of these features was increased to make them more salient, and the features were changed to solid black. To manipulate uncertainty concerning the patterns' scope, the features for one of the patterns (which in the

informative labels condition would always be the label-relevant pattern) were only shown for three of the four robots in each category, with the feature for the fourth item in each category hidden behind a box labeled "unknown." As a result the label-irrelevant pattern subsumed eight out of eight observations (100%), while the label-relevant pattern only applied to six out of eight observations (75%) with certainty. We counterbalanced across two sets of materials: (1) the informative labels were "Indoor/Outdoor" and feet figured in the label-relevant pattern (Fig. 3a), or (2) the informative labels were "Receiver/Transmitter" and antennae figured in the label-relevant pattern (Fig. 3b).

"Glorp/Drent" labels were used in all blank labels conditions. Although the labels were not informative with respect to either pattern, we counterbalanced materials to match the informative labels conditions. This means that in the blank labels condition the "label-relevant pattern" refers to the pattern with potentially narrower scope (two relevant features "unknown") and "label-irrelevant" to the pattern that applied to all study examples.

5.1.3 Procedure. The learning phase was identical to Experiments 1 and 2, except that participants were informed before study that information that was not known about the robots would be indicated with an "unknown" box, and the robots were displayed by category to facilitate pattern discovery (exactly as in Fig. 3). After the learning phase participants were informed that the robots they had seen were just eight of the thousands on planet ZARN and made the following judgments. The order of these blocks was randomly chosen and did not have any effect in later analyses.

5.1.3.1. Pattern discovery. Participants responded "Yes," "Maybe," or "No" as to whether there were differences in the feet, antennae, and colors of robots in each category. They

also reported these differences and indicated how many of the eight study robots exhibited these differences.

5.1.3.2 Basis for categorization. The original image with the study observations was reproduced on screen during classification to eliminate memory demands. Participants classified two novel robots for which the label-irrelevant and label-relevant patterns generated opposite classifications. For example, one item involved pointy feet (associated with Outdoor/Receiver/Glorp) paired with a shorter left antenna (associated with Indoor/Transmitter/Drent). The robot's face and body were concealed by an "unknown" box such that only the antennae and feet were visible. Confidence ratings on a scale from 1 (not at all confident) to 7 (extremely confident) were also collected.

5.1.3.3. Beliefs about pattern scope. The original image with the study observations was reproduced on screen and a robot that was identified as novel was presented behind an "unknown" box such that a single feature was visible. For each of the features (a pair of antennae with a shorter left side, a pair of antennae with a shorter right side, triangle feet, or square feet) participants were shown a corresponding robot and asked: (1) "Out of every 100 Outdoor (Receiver/Glorp) robots on ZARN, how many do you think have antennae (feet) like the robot above?" (2) "Out of every 100 Indoor (Transmitter/Drent) robots on ZARN, how many do you think have antennae (feet) like the robot above?" Responses were made on a scale from 0 to 100. An identical block of transfer questions included four features that were novel antennae and feet following the same abstract patterns: shorter right/left antenna and pointy/flat feet.

5.2 Results & Discussion

5.2.1 Pattern discovery. The majority of participants discovered both patterns: Only 11% of participants reported that there were no feature differences across categories. Task and

label type had no significant effects on whether participants reported that they did not detect any differences (all *p*s > 0.10, free study/blank labels, 12%; explain/blank labels, 11%; free-study/informative labels, 13%; explain/informative labels, 8%). These participants are included in subsequent analyses, as excluding them did not change the results.

A majority of participants reported differences in color (80%), with no effect of condition. Participants noticed that the label-relevant pattern applied to six observations and the label-irrelevant pattern to eight (these were the modal responses), with no significant effects of condition (all *p*s > 0.10).

5.2.2 Basis for categorization. High rates of discovery made it possible to examine the effects of explanation on the selection of patterns as a basis for categorization. Figure 4a indicates the proportion of novel robots (out of two) classified by using the label-relevant pattern as opposed to the competing label-irrelevant pattern. An ANOVA with this proportion as a dependent measure and task (explain, free study) and label type (informative, blank) as between subjects factors revealed a significant interaction between task and label type, $F(1, 539) = 3.92$, $p < 0.05$, which superseded main effects of task, $F(1, 539) = 6.05$, $p < 0.05$, and label type, $F(1, 539) = 7.51$, $p < 0.01$. Participants who explained with informative labels privileged the label-relevant pattern to a greater degree than those in any other condition (the explain/blank labels condition, $t(262) = 3.30$, $p < 0.01$, $d = 0.41$, the free study/informative labels condition, $t(260) = 2.98$, $p < 0.01$, $d = 0.37$, and the free study/blank labels condition, $t(267) = 3.77$, $p < 0.001$, $d = 0.46$).

While there were additional effects of population and materials, neither factor interacted with the variables of interest, nor did including them in analyses change the significance of

reported results.[6] This indicates that explanation's effects depended on whether the labels favored one pattern over the other, not the particular labels and materials used in the previous studies.

5.2.3 Inferred and relative pattern scope. To represent participants' inferences about how broadly a pattern in study observations would extend to the entire category, we computed an aggregate measure of *inferred pattern scope* from participants' judgments about the prevalence of the foot and antenna features in each category. Each response about how many unobserved category members (out of 100) would have a particular feature serves as an intuitive estimate of a feature's *category validity* – the probability that a member of the category has the feature. To create an aggregate across these judgments, we added the number of estimated pattern-consistent robots and subtracted the number of estimated pattern-inconsistent robots. So, for example, suppose a participant reported that 90 out of 100 Outdoor robots have triangular feet and 90 out of 100 Indoor robots have square feet, consistent with the study pattern, but that 5 out of 100 robots of each type have the opposite type of feet, violating the study pattern. The average pattern-inconsistent judgment (5) would be subtracted from the average pattern-consistent judgment (90) to create a composite score of 85 for this participant.[7]

Inferred pattern scope is presented in Table 3 for the label-relevant and label-irrelevant patterns. Additionally, Table 3 reports a conversion of these judgments into *relative pattern*

[6] The effect of population was as follows: Lab participants tended to generalize the label-relevant pattern more than online participants, $t(541) = 2.70$, $p < 0.01$, $d = 0.23$. There was also an effect of materials: The label-relevant pattern was more likely to be generalized when the pattern and labels concerned feet than when they concerned antennae, $t(541) = -2.77$, $p < 0.01$, $d = -0.24$. However, including *population* and *materials* as factors in the reported analysis did not alter the statistical conclusions or reveal any interactions with task or label type.

[7] While we could have converted participants' judgments into an estimate for the *probability* of a pattern-relevant feature given category membership, doing so required division and multiplication, so estimates of zero posed a problem. However, the aggregate measure we employed produced the same pattern of results as calculating category validities by dropping zero scores or replacing them with 0.5.

scope, which is calculated as the inferred pattern scope for the label-relevant pattern minus inferred pattern scope for the label-irrelevant pattern.

Mirroring our analysis of basis for categorization, a task (explain, free study) by label type (blank, informative) ANOVA was performed on relative pattern scope. Overall, participants believed that the label-irrelevant pattern (which applied to all eight study observations) had broader scope than the label-relevant pattern (for which the status of two observations was uncertain), as relative pattern scope was significantly less than zero, $F(1, 539) = 84.79$, $p < 0.01$. However, there was one additional significant effect: an interaction between task and label type, Participants who were prompted to explain and received informative labels penalized the label-relevant pattern (relative to the label-irrelevant pattern) *less* than those in other conditions, $t(262) = 2.70$, $p < 0.01$, $d = 0.33$, presumably because prior knowledge played a larger role in informing their judgments. Interestingly, in the blank labels conditions there was a marginal trend for explaining to have the opposite effect, $t(259) = -1.67$, $p = 0.097$, $d = -0.21$, more strongly favoring the label-irrelevant pattern, which accounted for more observed cases with certainty. Such an effect would be consistent with the idea that explaining increases reliance on all cues to scope.

Finally, recall that the experiment additionally asked participants how many robots would have novel "transfer" features. However, the majority of participants, 55%, reported that none of the transfer features would be present in *any* unobserved category member, and so we do not analyze this measure further.

5.2.4. Summary. Experiment 3 examined which of two discovered patterns was utilized in classifying novel category members and believed to generalize to unobserved category members. Classification judgments revealed an interaction between task (explain vs. free study)

and label condition (blank vs. informative), with participants who explained with informative labels using the label-relevant pattern more often than participants in any other condition, and doing so to a degree that exceeded the summed, independent effects of explanation and label type. This impact of explaining with informative labels was mirrored by participants' beliefs about whether more category members – observed *and* unobserved – conformed to the label-relevant or label-irrelevant pattern. These findings mirror those from Experiments 1 and 2, with generalization driven by a parallel interaction between explanation and prior knowledge. Unlike Experiment 1, however, we can be confident that effects on generalization were not merely a consequence of discovery, as most participants discovered both patterns.

6. Experiment 4

Experiments 1, 2, and 3 found that explaining can influence discovery and generalization by recruiting the knowledge cued by informative category labels. We proposed a subsumptive constraints account of explanation as the basis for predicting and interpreting these effects. Specifically, we suggested that explanations are better to the extent that they invoke patterns with broad scope, and that prior knowledge is recruited to infer the scope of candidate patterns.

Experiment 4 provided a more direct test of the idea that prior knowledge is recruited in explanation as a cue to the scope of candidate patterns. We accomplished this by creating a situation in which participants possessed semantically-relevant prior knowledge that was *not* in fact a reliable cue to scope. If prior knowledge is not a reliable cue to scope, then participants prompted to explain should be no more likely than participants in control conditions to rely on prior knowledge. To create this situation, participants in a *random labels* condition were presented with study examples with informative labels (e.g., Indoor, Outdoor) that could be related to particular features of the examples (e.g., foot shape), but – crucially – they were told

that the labels were assigned based on the outcome of a random coin flip. As a result, the features of observed category members should not be correlated with category membership, making prior knowledge an unreliable cue to whether patterns that effectively differentiate study items generalize to the robot population. In this situation, explaining should not lead to greater reliance on prior knowledge as a cue to scope.

In addition to the random labels condition, we also included a *representative labels* condition, which matched previous experiments: Participants were not told how labels were assigned to examples, but could reasonably assume that study observations were representative of their respective categories. Including both the random and representative labels conditions also introduced a second cue to the scope of diagnostic patterns, roughly "method of label assignment," since diagnostic patterns across study observations (whether or not they relate to prior knowledge) should only generalize to the population in the representative labels condition. If explanation heightens people's sensitivity to all cues to scope – and not just to prior knowledge – then participants in the explain condition should be more responsive to this manipulation than those in the control condition.

Experiment 4 also aimed to replicate the key findings from Experiment 3 while addressing two potential concerns. First, the task differences found in Experiment 3 are subject to the same concern as Experiment 1, namely that the control task was less demanding than explanation in some relevant respect. Experiment 4 introduced the stronger control condition used in Experiment 2, requiring participants to write out their thoughts during study and therefore matching the explain condition along more dimensions. Second, the manipulation of label type in Experiment 3 was confounded with the presence of "unknown" features, which were always involved in the label-relevant pattern. The interaction between explanation and label

type could therefore have been produced by the presence of the "unknown" features, with a prompt to explain encouraging participants to focus on and draw inferences concerning these features. Experiment 4 avoided this concern by testing whether the interaction between explanation and label type occurred even when all features were visible.

Finally, Experiment 4 provided two additional extensions to previous experiments. The comparison of informative and blank labels in Experiments 1-3 provided one way of examining the effects of prior knowledge, namely by *increasing* the knowledge available to some participants. Experiment 4 instead manipulated the *content* of available prior knowledge by comparing two sets of informative labels: Outdoor/Indoor versus Receiver/Transmitter.[8] We predicted that explanation and label pair would interact to determine the extent to which category membership was generalized on the basis of the foot versus antenna pattern. The second extension in Experiment 4 was to evaluate whether the previous findings would generalize to learning contexts with extremely sparse observations. Instead of four examples from each category, Experiment 4 presented participants with only one. Forming generalizations from such limited information is a valuable inductive capacity, and one for which explanation and prior knowledge could be especially critical (Ahn, Brewer, Mooney, 1991).

6.1. Methods

6.1.1 Participants. Six-hundred-and-eighty-two members of the Amazon Mechanical Turk workplace from the United States participated online for monetary compensation.

6.1.2. Materials & procedure. Participants studied just two robots, one from each category (robots 1 and 8 in Fig. 3), and no features were hidden with "unknown" boxes. The learning phase was adapted from Experiment 3 with the following changes. First, we

[8] This comparison across informative label pairs was technically possible in Experiment 3, which likewise employed both sets of labels, but would be problematic to interpret given that a pattern's label-relevance was confounded with its inclusion of an "unknown" feature.

manipulated *learning task* through prompts to *explain* versus *write thoughts*, as in Experiment 2. Second, we used only the two *label pairs* from the informative labels conditions of Experiment 3 (Outdoor/Indoor or Receiver/Transmitter). And finally, we added an additional factor, *label assignment,* by changing the cover story about how labels were assigned to produce *representative labels* or *random labels*.

For all participants, the cover story mentioned that the robots were created by the aliens living on the planet, and included information about their function that was appropriate to the label pair, either "Outdoor robots work on outdoor terrain and Indoor robots work inside houses," or "Receiver robots receive messages and Transmitter robots send messages."

In the *representative labels* conditions, participants received no additional information. In the *random labels* conditions, participants were additionally told: "The aliens decide which robots are Outdoor (Receiver) robots and which robots are Indoor (Transmitter) robots when they are manufactured. When a robot comes off the assembly line at the robot factory, a coin is flipped. If the coin lands heads, the robot is declared an Outdoor (Receiver) robot. If the coin lands tails, the robot is declared an Indoor (Transmitter) robot."

As in Experiment 3, participants classified robots and answered questions about the prevalence of features, as detailed below. These two tasks occurred in randomized order after the learning phase.

6.1.2.1. Basis for categorization. Participants classified six different robots, making their ratings on a six-point scale from "Definitely an Indoor (Transmitter) robot" to "Definitely an Outdoor (Receiver) robot." Two robots looked exactly like the original study items, two robots involved the same features but introduced a conflict between the two patterns (i.e., the feet from one category but the antennae from the other), and the final two presented the same conflict with

novel "transfer" features (i.e., novel feet that were pointy versus flat, and novel antennae that were longer on the right or left).

6.1.2.2. Inferred pattern scope. Participants answered 16 questions (8 judgments for each category), which all asked how likely it was that a randomly selected Outdoor/Indoor robot (or Receiver/Transmitter) would have a particular feature, a picture of which was shown. The eight features were: the two foot shapes observed at study, the two observed antenna configurations observed at study, two previously unseen transfer foot shapes following the foot pattern, and two previously unseen transfer antenna configurations following the antenna pattern. Responses to these questions were used to calculate inferred pattern scope, as in Experiment 3.

6.2 Results

We first examine the effects of explanation and label assignment on categorization and inferred scope of the label-relevant and label-irrelevant patterns, collapsing across the two label sets. We then consider individual effects of the Outdoor/Indoor versus Receiver/Transmitter label pairs and characteristics of participants' written responses.

6.2.1. Basis for categorization. Figure 4b reports the average ratings for the categorization task, with responses coded such that higher numbers correspond to judgments consistent with the label-relevant pattern. This measure was analyzed in an ANOVA with task (write thoughts, explain) and label assignment (random, representative) as independent variables. The critical finding was a task by label assignment interaction, $F(1, 678) = 5.40$, $p < 0.05$, which superseded a main effect of label assignment, $F(1, 678) = 27.51$, $p < 0.001$. Relative to the write thoughts condition, explaining promoted categorization consistent with the label-relevant pattern in the *representative labels* condition, $t(341) = 2.35$, $p < 0.05$, $d = 0.25$, but had no effect in the *random labels* condition, $t(337) = 0.97$, $p = 0.33$, $d = 0.11$. Moreover, the effect of label

assignment was greater when participants engaged in explanation, $t(332) = 5.21$, $p < 0.001$, $d = 0.58$, than when they wrote their thoughts, $t(357) = 2.28$, $p = 0 < 0.05$, $d = 0.24$. These results were not changed by including label pair (Outdoor/Indoor, Receiver/Transmitter) as a between-subjects factor and *categorization item* (original observations, conflict items pitting patterns against each other, conflict items with novel features) as a within-subjects factor in the analysis..

These findings are consistent with the prediction that explanation does not recruit prior knowledge as a basis for judgment when it is an unreliable cue to scope (i.e., in the random labels condition), and also the prediction that explanation heightens participants' sensitivity to additional cues to scope – in this case, the method of label assignment.

6.2.2 Inferred pattern scope. Table 4 reports participants' beliefs about the scope of the label-relevant and label-irrelevant patterns. These were calculated using the same procedure as Experiment 3 in order to reflect participants' implicit beliefs about how likely the patterns in study observations would be to apply to the entire category.

We analyzed inferred pattern scope as the dependent measure in a mixed ANOVA, treating *pattern type* (label-relevant, label-irrelevant) as a within-subjects factor, and task (write thoughts, explain) and label assignment (random, representative) as between-subjects factors. There was a main effect of label assignment, $F(1, 678) = 22.93$, $p < 0.001$, with higher ratings of pattern scope in the representative than random labels conditions, and a main effect of pattern type, $F(1, 678) = 92.26$, $p < 0.001$, with higher ratings for the label-relevant pattern. However, these effects were qualified by three two-way interactions. First, as predicted, there was an interaction between task and label assignment, $F(1, 678) = 9.49$, $p < 0.01$, with participants prompted to explain more sensitive to the manipulation of label assignment than those in the control condition: In the explain condition, representative labels led to judgments of greater

pattern scope than random labels, $t(322) = 5.83$, $p < 0.001$, $d = 0.65$, with no effect of labels in the write thoughts condition, $t(356) = 1.04$, $p = 0.30$, $d = 0.11$ (see "pooled pattern scope" in Table 4). Second, there was an interaction between task and pattern type, $F(1, 678) = 15.02$, $p < 0.001$, with participants who explained more strongly differentiating the scope of the label-relevant and label-irrelevant patterns. Finally, label assignment also interacted with pattern type, $F(1, 678) = 7.10$, $p < 0.01$, with the two patterns more strongly differentiated in the representative labels conditions than in the random labels conditions. Including kind of feature (original, transfer) and label pair in analyses did not change these results.

These findings again support the prediction that explanation increases participants' sensitivity to a novel cue to scope: method of label assignment. The interaction between task and pattern type is also consistent with our previous results in that participants who explained were more sensitive to prior knowledge than those who wrote thoughts. However, we did not find that explanation's effects on prior knowledge were eliminated with random labels (which would have been reflected in a three-way interaction between task, pattern type, and label assignment) to mirror the predictions and findings for categorization Instead, participants inferred a broader scope for the label-relevant pattern than the label-irrelevant pattern for both explain conditions.

6.2.3. Effects of label pair. The representative labels conditions in Experiment 4 varied from the preceding experiments in using two different label pairs in otherwise identical conditions. These conditions allow us to assess whether explanation and prior knowledge interact when the *content* rather than *amount* of prior knowledge is manipulated.

Average categorization ratings were therefore analyzed with a task (write thoughts, explain) by label pair (Indoor/Outdoor, Receiver/Transmitter) ANOVA, but restricted to the representative labels conditions and with ratings coded such that higher numbers indicated

consistency with the foot pattern. This analysis revealed main effects of task, $F(1, 678) = 5.47$, $p < 0.05$, and label pair, $F(1, 678) = 100.16$, $p < 0.001$, and a task by label pair interaction, $F(1, 678) = 4.09$, $p < 0.05$. Average categorization ratings were higher (more consistent with the foot pattern) for the two Outdoor/Indoor labels conditions (write thoughts: $M = 4.6$, $SD = .5$, explain: $M = 4.6$, $SD = .6$), and lower for the Receiver/Transmitter labels (write thoughts: $M = 4.2$, $SD = .6$, explain: $M = 3.9$, $SD = .5$). Although labels affected categorization judgments for participants in *both* groups (write thoughts: $t(356) = 5.95$, $p < 0.001$, $d = 0.63$, explain: $t(322) = 8.07$, $p < 0.001$, $d = 0.90$), the effect was still more pronounced for those prompted to explain.

Analyses of inferred pattern scope mirrored these findings. Table 5 reports inferred pattern scope for the foot and antenna patterns in the representative labels conditions, as well as relative pattern scope, the difference between them, with positive numbers corresponding to higher relative scope for the foot pattern. A task x label pair ANOVA on relative pattern scope found a main effect of label pair, $F(1, 678) = 53.79$, $p < 0.001$, and a task by label pair interaction, $F(1, 678) = 16.00$, $p < 0.001$. Participants in both study conditions inferred a broader scope for feet than for antennae with the Indoor/Outdoor labels, and the reverse pattern held true with Receiver/Transmitted labels, but the magnitude of the difference across label pairs was greater for participants prompted to explain. Nonetheless, the effect of label pair was still independently significant in the write thoughts condition, $t(356) = 2.45$, $p < 0.05$, $d = 0.26$.

6.2.4. Written responses. Analyses of written responses were restricted to participants who did not leave responses blank; the proportion of participants who did so did not differ significantly across conditions (all ps > 0.10) and was less than 1%. An ANOVA on response length with task and label assignment as between-subjects factors revealed that typed responses were longer when participants were asked to write thoughts than to explain (explain: $M =$ 28.2,

SD = 15.5; write thoughts: *M* = 33.4, *SD* = 18.5; $F(1, 678) = 15.6$, $p < 0.001$), substantiating the trend observed in Experiment 2. Responses were also longer in the representative labels conditions (*M* = 32.21, *SD* = 18.1) than the random labels conditions (*M* = 29.7, *SD* = 16.5), $F(1, 678) = 3.84$, $p < 0.05$.

The proportion of participants who mentioned label was influenced by a task by label assignment interaction, $\chi^2(1, N = 682) = 4.82$, $p < 0.05$. When the labels were randomly assigned, participants in the explain condition mentioned them *less frequently* than participants who wrote out thoughts (explain: 34%; write thoughts: 47%, $\chi^2(1, N = 339) = 5.80$, $p < 0.05$), while no such difference existed for representative labels (explain: 41%; write thoughts: 37%; $\chi^2(1, N = 343) = 0.47$, $p = 0.51$).

These findings suggest that the effects of explanation on generalization reported above are unlikely to derive from differences in general engagement or attention to labels across conditions.

6.2.5. Summary. Experiment 4 went considerably beyond the previous experiments in manipulating a novel cue to the scope of patterns across study observations: whether observed category members had features that could be assumed to correlate with category membership or were assigned labels at random. When labels were assigned at random, such that prior knowledge was no longer a reliable cue to the scope of diagnostic patterns, prior knowledge differences between the explain and write thoughts conditions were eliminated when it came to categorization. The manipulation of label assignment also interacted with explanation analogously to the previous manipulations of prior knowledge: Participants prompted to explain were more sensitive to this cue to pattern scope, with greater differentiation of the representative and random conditions for both the classification of novel robots and the extension of observed

features to unobserved category members. The fact that explanation had a comparable impact on a quite distinct cue to scope bolsters our interpretation that effects of informative labels in the preceding experiments are best understood as a consequence of the fact that explaining directs learners to assess patterns' scope, where the number of current observations consistent with a pattern, prior knowledge, and how categories are formed (i.e., method of label assignment) are all cues to scope.

Finally, Experiment 4 also addresses potential concerns about the preceding results. First, key findings from Experiment 3 replicated without "unknown" features, with a stronger control condition, and with sparser data, showing that explaining can promote the recruitment of prior knowledge to guide generalization with just one or two category observations. Second, Experiment 4 found that label type had a significant effect on participants in control conditions. This finding helps address a concern with the previous experiments – that superadditive effects of explanation and labels are restricted to conditions under which participants do not spontaneously consult labels in the absence of explanation. This alternative explanation is less plausible, since explanation and label type had superadditive effects even when label type had significant effects independently.

7. General Discussion

Four experiments examined how generating explanations and possessing prior knowledge (cued by informative category labels) influenced what participants learned and inferred about novel categories from examples. Experiments 1 and 2 found that explaining increased the extent to which participants relied on prior knowledge in learning, leading to elevated discovery of a pattern related to the informative labels when they were provided. However, the effects of explaining were selective: Explaining increased the rate at which participants discovered one

pattern without increasing the discovery of additional patterns. In fact, when just those participants who had discovered at least one pattern were considered, discovery of additional patterns was *lower* for participants prompted to explain than for those in control conditions. These results were replicated in Experiment 2 despite the inclusion of four patterns and a more demanding control condition that required participants to write their thoughts during study.

Experiments 3 and 4 examined whether explaining could directly impact which patterns were generalized beyond study observation. Although seeking explanations had no impact on pattern discovery (which was near ceiling), participants prompted to explain with informative labels were more likely to categorize novel items using the label-relevant pattern, and more likely to believe that the label-relevant pattern applied to unobserved category members. Experiment 4 additionally found that explaining increased sensitivity to an additional cue to the scope of patterns across observed category members: whether category members were drawn from randomly assembled populations.

Jointly, the results from Experiments 1-4 provide strong support for the idea that explaining can increase the extent to which learners consult prior knowledge to guide discovery and generalization. The findings also shed light on the mechanisms by which explaining generates these effects. First, several results challenge the idea that explaining affects learning through a general increase in attention or engagement, or even through a global increase in the extent to which people seek patterns. Instead, effects of explanation were quite selective (Experiments 1-2), and extended to cases in which multiple patterns were available to learners and needed to be preferentially applied to new cases (Experiment 3-4). Second, the results support our proposal that explaining increases learners' consultation of prior knowledge as a cue to patterns' scope. Explaining magnified the role of informative labels on estimates of a label-

relevant pattern's scope in Experiments 3 and 4, with a parallel impact on a completely different cue to scope (random versus representative category labels) in Experiment 4.

We interpret these findings in terms of the *subsumptive constraints* account. To briefly review, the account maintains that people prefer explanations that appeal to patterns with broad scope, with the result that explaining constrains learners to identify patterns and make use of cues to patterns' scope. Our experiments manipulated two distinct cues to scope, prior knowledge (through informative labels) and method of label assignment (random versus representative), finding the predicted effects of explanation in each case. Combined with previous work (Williams & Lombrozo, 2010) demonstrating comparable effects of explanation on a third cue to scope – the number of explained observations to which a pattern applies – there is good reason to think that explanation's effects are truly tracking cues to scope, and not alternative features of each manipulation.

While an important relationship between explanation and prior knowledge is often endorsed (for discussion see Lombrozo, 2006), little empirical work has tried to characterize *which* knowledge is consulted and *why* it is brought to bear through explanation. One reason may be the challenge posed in relating explanation to the range of beliefs that count as "prior knowledge." The current work suggests that explaining will invoke knowledge relevant to evaluating whether an observed pattern extends to novel cases and contexts. But explaining should play a smaller role in deploying other kinds of knowledge, such as idiosyncratic facts about examples or information that serves a purely mnemonic purpose. The present account also predicts that the influence of prior knowledge must trade-off against other cues to scope, which suggests that when alternative cues to scope are very strong, explaining could actually *decrease* the role of prior knowledge in learning. This paradoxical prediction can make sense of an

otherwise puzzling feature of explanation: that explaining an anomalous observation can sometimes lead to "explaining away" and the preservation of current beliefs (Chinn & Brewer, 1993; see also Bott & Murphy, 2007; Hayes, Foster, & Gadd, 2003), but at other times presage deep conceptual change (e.g., Amsterlaw & Wellman, 2006).

7.1 Alternative explanations

Our experiments were designed to assess and rule out a few alternative explanations for the results. First, effects of explanation could potentially be attributed to task demands if explanation prompts somehow communicated to participants that the experimenter intended for them to find a pattern or take category labels seriously. Counter to this view, however, spontaneous explanation in the control condition from Experiment 1 had comparable effects to prompted explanation, and explain and free study participants who did not discover a pattern were equally likely to believe one existed.

More generally, while each individual experiment is prone to alternative interpretations, these are rendered less plausible by the systematic effects of explanation and prior knowledge across four experiments that differed in various ways. For example, in Experiments 1 and 3, participants in the free study condition could have been less engaged and paid too little attention to the labels to benefit from prior knowledge, with explaining simply increasing attention or engagement past some threshold where prior knowledge could have an effect. But the key results from these experiments were replicated when using the more engaging control condition of typing thoughts (Exp. 2 & 4), where we found that participants in the explain and control conditions were equally likely to mention informative category labels, and when simplified stimuli in Experiments 3 and 4 reduced the attentional resources required to notice patterns and labels.

We do acknowledge that the experiments can only provide indirect evidence that explaining recruited prior knowledge in the service of assessing patterns' scope. However, it is notable that explaining had comparable effects on multiple cues to scope: the availability of prior knowledge (Experiments 1, 2, and 3), the content of prior knowledge (Experiment 4), whether category labels were randomly assigned (Experiment 4), and the number of study observations conforming to a pattern (Williams & Lombrozo, 2010). This convergence supports our appeal to scope. In other words, we take the broad scope of our scope explanation as evidence in its favor.

7.2 Implications for Category Learning

The current findings shed light on how explaining could play a distinctive role in category learning, much as classification and inference learning each do (Chin-Parker et al., 2006; Markman & Ross, 2003). In particular, our account predicts that explaining should encourage learners to focus on patterns underlying category membership that are expected to have broad scope. When scope is assessed only in terms of the examples encountered in training, then explaining should result in the reduction of classification error on examples, a core mechanism underlying category learning (Kruschke, 2008). In fact, the findings from Williams & Lombrozo (2010) are consistent with the idea that explanation can have this effect, and prior knowledge likely does influence learning by reducing training error (Rehder & Murphy, 2003). However, the number and proportion of study items accommodated by a given pattern is only one cue to scope. The consequences of explaining category membership could therefore diverge from error-driven learning when learners have access to additional cues to scope, such as prior knowledge. Along these lines, we have found that prompting 5-year-olds to explain can actually make them less likely than children in a control condition to favor a pattern that accounts for all observations, but is inconsistent with prior knowledge (Walker, Williams, Lombrozo, & Gopnik,

2012, under review). The findings from Experiment 3 have a similar flavor: Explaining led adults to less strongly favor a pattern that accounted for all observations with certainty over an alternative that accounted for only 75%, but was more congruent with informative category labels.

It is also possible that *spontaneous* explanation during learning can help explain characteristics of learning in prior research. In particular, explaining could be a cause or consequence of the learning mode employed, shifting learners towards a rule-based system (Ashby & Maddox, 2004; Nosofsky & Palmeri, 1994), or to prototypical rather than exemplar-based representations (Griffiths et al., 2007; Smith & Minda, 1998; Vanpaemel & Storms, 2008). More broadly, explaining could constrain learning to be more explicit (Maddox & Ing, 2005; Matthews et al., 1989), intentional (Love, 2002) and reliant on language and abstract construals (Lupyan & Rakison, 2007; Trope & Liberman, 2010). Undocumented effects of spontaneous explanation are especially plausible in cases where learning is sensitive to prior knowledge and cannot be fully explained through the reduction of classification error (for examples, see Bott, Hoffman & Murphy, 2007; Kim & Rehder, 2011). Our demonstration of the powerful role of explanation in category learning indicates the value of not only experimentally manipulating explanation, but also tracking spontaneous explanation through verbal protocols or post-test questions (as in the explanation self-report measure from Experiment 1).

Spontaneous explanation might also play a role in cases where categorical judgments diverge from statistical learning. For example, Murphy and Spalding (1999) found that participants who learned knowledge-consistent ("integrated") categories were less sensitive to the frequencies of category features than those who learned arbitrary ("nonintegrated") categories when it came to judgments of typicality (see also Murphy & Allopenna, 1994;

Wisniewski, 1995), consistent with the effect in Experiment 3, where participants who explained with informative labels were least sensitive to the difference in frequency between the features that appeared in the 75% and 100% patterns when it came to inferring pattern scope. However, Spalding and Murphy also found that participants who learned knowledge-consistent categories were *more accurate* in their estimates of feature frequencies when they were simply asked to report them. One speculative possibility is that judgments that require people to relate features to each other or to category membership, such as categorization and typicality ratings, are more likely to trigger spontaneous explanation than judgments that involve descriptive reporting, such as feature frequency estimates (see also Murphy & Medin, 1985; Rips, 1989). Spontaneous explanation could also play a larger role in more open-ended and constructive categorization tasks, such as Wisniewski and Medin's (1994) paradigm, which required participants to construct novel features and rules to differentiate complex stimuli, and to explain while they did so.

Although explanation likely contributes to previous findings concerning the role of prior knowledge in category learning, our findings also provide suggestive evidence that explaining and prior knowledge can play quite different roles when it comes to learning material that is knowledge-irrelevant. Many studies have found – perhaps surprisingly – that learning a category that is only partially consistent with prior knowledge does not hinder learning of knowledge-irrelevant features, and may even generate improvements relative to learning categories that are not related to prior knowledge (Heit, Briggs, & Bott, 2004; Kaplan & Murphy, 2000; see Murphy, 2002 for discussion). In our own data, there was a marginal effect (in Experiment 1, p = .062) for participants who received informative labels to be more likely than those who received blank labels to discover more than one pattern, and a significant effect where those who discovered the label-irrelevant pattern were more likely to have also discovered the label-

relevant pattern, consistent with the idea that prior knowledge facilitates learning of knowledge-relevant *and knowledge-irrelevant* patterns. In contrast, explaining did not increase the rate at which participants discovered more than one pattern, and in fact decreased the probability that a second pattern was discovered given discovery of an initial pattern (Experiments 1 and 2). These findings suggest that explanation and prior knowledge might impose unique constraints on learning. Where explaining recruits constraints that privilege patterns with broad scope (potentially at the expense of other patterns or kinds of structure), prior knowledge could have mnemonic or other processing benefits that extend to knowledge-irrelevant features.

Of course, these are empirical hypotheses in need of further support. An additional dimension worth exploring concerns the nature of the subsumptive relationship between an explanation and category membership. Here we have considered cases in which patterns are better or broader if they account for the category membership of more items. An alternative sense of scope, however, concerns the number of features of individual members that can be explained by appeal to category membership. For example, one pattern (pointy versus flat feet) could successfully differentiate many robots, while another pattern (features relevant to working in space versus underwater) could apply to fewer robots, but explain a larger number of features for those robots (e.g., why they are a particular color, made of a particular material, *and* of a particular size). Research on knowledge effects in category learning has varied both the number of items and the number of features to which themes apply; Similar variation would be fruitful to examine within our paradigm, especially as a way to understand whether and how these two factors trade-off when learners explain. It is also likely that not all items or features are equal when it comes to assessing scope. For example, explaining could favor patterns that account for

more diverse cases (Kim & Keil, 2003) or more ideal cases (Barsalou, 1985), even when doing so does not account for the largest number of items or features.

We also expect that the content of explanation prompts (i.e., what it is that people actually explain) should influence which patterns are relevant, and therefore which patterns are discovered and evaluated for scope. In our experiments, participants explained why an object belonged to one category (as opposed to another). Successful explanations therefore invoked patterns that were "diagnostic" in the sense that they identified features that differentiated members from the two categories, and our experiments correspondingly assessed whether diagnostic patterns were discovered and generalized. However, categories can involve additional patterns that could be targeted by other explanation prompts. For example, having participants explain why two features might co-occur in members of a given category should instead affect the discovery and generalization of "co-occurrence" patterns, with the relevant sense of scope concerning which co-occurrences are likely to generalize beyond observed cases to unobserved category members.

Nonetheless, explanation might not have comparable effects for all kinds of scope. People prefer explanations with broader scope in the sense that the explanation can account for more actual phenomena (e.g., Preston & Epley, 2005) or actual observations (e.g., Read & Marcus-Newhall, 1993), but people prefer explanations with *narrow* "latent scope" – that is, that are committed to fewer potential observations that have not been made (Khemlani, Sussman, & Oppenheimer, 2011). An important question for future research is whether and when this preference for narrow latent scope manifests in effects of explanation on learning. One possibility is that a benefit for patterns with broad scope will be tempered when those patterns involve a commitment to entirely new kinds of observations (e.g., a novel kind of feature that has

not been observed, such as hats on robots) as opposed to new instances of features that have already been observed (e.g., pointy feet on unobserved robots).

Finally, our account generates the counterintuitive prediction that under some conditions, explaining will hinder category learning. In particular, explaining could divert effective learning when categories lack underlying patterns or involve "unexplainable" exceptions. Under these conditions, explaining could reinforce broad patterns that make sense in light of prior knowledge at the expense of effectively tracking the world. Our ongoing research supports this prediction (Williams, Lombrozo, & Rehder, 2010, 2011, under review), and helps explain why participants in control conditions may not have always explained spontaneously or engaged in equivalent processing: it is not always beneficial to do so (see also Berthold et al., 2011; Kuhn & Katz, 2009).

7.3 Implications for Education

In the introduction we identified several proposals concerning the effects of explanation on learning, including the ideas that explaining can increase a learner's attention or motivation (e.g., Siegler, 2002) or help identify gaps in understanding (e.g., Chi et al., 1989; Nokes et al., 2011), among others. The current work was not designed to directly challenge these account or arbitrate between them. In fact, we see our findings as importantly complementary. If we are correct that explaining imposes a set of criteria for what constitutes a good explanation, and that these criteria constrain discovery and generalization, then the factors we identify should inform how learners direct their attention, what they are motivated to discover, which gaps in understanding are most problematic, what kinds of inferences must be drawn, and so on. An important direction for future research is thus to combine the richness of past research on explanation and learning from education with the kind of experimental control afforded by

artificial category learning, allowing the selectivity of explanation's effects to be studied in more complex and real-world environments.

Our account can also shed new light on past findings from self-explanation. For example, previous research has noted that one consequence of self-explanation is increased awareness of principles and laws, whether learning about physics (Chi et al., 1989), probability (Renkl, 1997), or arithmetic (Rittle-Johnson, 2006). A subsumptive constraints account helps explain why this is the case: Constructing successful explanations for a fact or problem solution should direct learners towards broad patterns, and principles and laws are prime examples of such patterns. However, our account also predicts pedagogically relevant conditions under which subsumptive constraints on explanation can *impair* learning. For example, for students without the requisite background to generate accurate generalizations, or who have not encountered enough counterevidence to erroneous beliefs for such observations to trump prior knowledge as a cue to scope, a prompt to explain could *reinforce* existing misconceptions (see also Walker, Williams, Lombrozo, & Gopnik, 2012, under review; Williams, Lombrozo, & Rehder, 2010, 2011, under review). Our findings can therefore inform future research aimed at testing the conditions under which explanation is most beneficial for learning in educational contexts.

8. Conclusion

Four experiments on learning categories provided evidence that explanation and prior knowledge interact in promoting the discovery and generalization of patterns underlying category membership. The findings support a subsumptive constraints account of explanation and learning, according to which explaining drives learners to seek underlying patterns and to consult prior knowledge in assessing the scope of such patterns – that is, how broadly the patterns apply within and beyond study observations. Our findings and account provide insight

into how constraints on explanation influence the role of observations and prior knowledge in guiding learning and generalization, and suggest that explaining can act as a mechanism for bringing prior knowledge to bear in learning.

Acknowledgments

This research was partially supported by the James S. McDonnell Foundation as well as NSF grant (DRL-1056712) awarded to the second author. We would like to thank Sam Maldonado for help programming experiments and analyzing data; Sam Maldonado, Caren Walker and Mike Pacer for providing feedback on previous versions of the manuscript, Lucie Vosicka for feedback on and assistance editing the manuscript, Norielle Adricula, Dhruba Banerjee, Vanessa Ing, Evan Kim, Adam Krause, Sean Trott, Jing Wang, and Kelly Whiteford for assistance collecting data, and members of the Concepts and Cognition lab for feedback on this research. JJW was supported by an NSERC post-graduate fellowship.

References

Ahn, W. (1998). Why are different features central for natural kinds and artifacts? *Cognition, 69,* 135-178.

Ahn, W., & Kalish, C.W. (2000). The role of mechanism beliefs in causal reasoning. In F. Keil & R. A. Wilson (Eds.), Explanation and cognition (pp. 199-226). Cambridge, MA: MIT Press.

Ahn, W.-K., Brewer, W. J., Mooney, R. J., University of Illinois at Urbana-Champaign. Beckman Institute, Cognitive Science. (1991). *Schema acquisition from a single example*.

Aleven, V., & Koedinger, K. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26,* 147-179.

Amsterlaw, J., & Wellman, H. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development, 7*, 139-172.

Ashby, F. G., & Maddox, W. T. (2004). Human category learning. *Annual Review of Psychology, 56,* 149-78.

Barsalou, L.W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 629-654.

Berthold, K., & Renkl, A. (2010). How to foster active processing of explanations in Instructional Communication. *Educational Psychology Review, 22,* 25-40.

Berthold, K., Roder, H., Knorzer, D., Kessler, W., & Renkl, A. (2011). The double-edged effects of explanation prompts. *Computers in Human Behavior, 27,* 69-75.

Best, R., Ozuru, Y., & McNamara, D.S. (2004). Self-explaining science texts: Strategies,

knowledge, and reading skill. In Y.B. Kafai, W. A. Sandoval, N. Enyedy, A. S. Nixon, & F. Herrera (Eds.), *Proceedings of the Sixth International Conference of the Learning Sciences: Embracing Diversity in the Learning Sciences* (pp. 89-96). Mahwah, NJ: Erlbaum.

Bott, Lewis; Hoffman, Aaron B.; Murphy, Gregory L. (2007). Blocking in category learning. *Journal of Experimental Psychology: General, 136*, 685-699.

Bott, L., & Murphy, G. L. (2007). Subtyping as a knowledge preservation strategy in category learning. *Memory & Cognition, 35*, 432-443.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General*, 127, 355-376.

Chi, M.T.H. (2009) Active-contructive-interactive: a conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1,* 73-105

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

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

Chi, M. T., VanLehn, K. A. (1991). The content of physics self-explanations. *Journal of the Learning Sciences, 1,* 69-105.

Chinn, C. A., & Brewer, W. F. (1993). The Role of Anomalous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction. *Review of Educational Research*, *63*(1), 1–49. doi:10.3102/00346543063001001

Chin-Parker, S., Hernandez, O., & Matens, M. (2006). Explanation in category learning. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 1098–1103*)*. Mahwah, NJ: Erlbaum.

Crowley, K., & Siegler, R. S. (1999). Explanation and generalization in young children's strategy learning. *Child Development, 70,* 304-316.

Fonseca, B. & Chi, M.T.H. (2011). The self-explanation effect: A constructive learning activity. In Mayer, R. & Alexander, P. (Eds.), *The Handbook of Research on Learning and Instruction*. Routledge Press.

Friedman, M. (1974). Explanation and Scientific Understanding. *Journal of Philosophy*, *71*, 5-19.

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

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371-395.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007) Unifying rational models of categorization via the hierarchical Dirichlet process. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society.*

Hayes, B. K., Foster, K., & Gadd, N. (2003). Prior knowledge and subtyping effects in children's category learning. *Cognition*, *88*(2), 171–199. doi:10.1016/S0010-0277(03)00021-0

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York, NY: John Wiley & Sons, Inc.

Heit, E. (2001). Background knowledge in models of categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and Categorization,* 155-178. Oxford University Press.

Heit, E. & Bott, L. (2000). Knowledge selection in category learning. *Psychology of Learning and Motivation, 39,* 163-199.

Heit, E., Briggs, J., & Bott, L. (2004). Modeling the effects of prior knowledge on learning incongruent features of category members. *Journal of Experimental Pyschology: Learning, Memory, and Cognition, 30,* 1065-1081.

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 829.

Khemlani, S. S., Sussman, A. B. & Oppenheimer, D. M. (2011). *Harry Potter* and the sorcerer's scope: latent scope biases in explanatory reasoning. *Memory & Cognition, 39*(3). 527-535.

Kim, N. S., & Keil, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory & cognition*, *31*(1), 155–165.

Kim, S., & Rehder, B. (2010). How prior knowledge affects selective attention during category learning: An eyetracking study. *Memory & cognition*, *39*(4), 649–665.

Kitcher, P. (1981). Explanatory Unification. *Philosophy of Science, 48*, 507-31.

Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In Philip Kitcher and Wesley Salmon (Eds.), *Minnesota Studies in the Philosophy of Science, Volume XIII: Scientific Explanation* (pp. 410-505). University of Minnesota Press.

Koslowski, B., Marasia, J., Chelenza, M., and Dublin, R., (2008). Information becomes evidence when an explanation can incorporate it into a causal framework. *Cognitive Development*, *23,* 472–487.

Kruschke, J. (2008). Models of categorization. In R. Son (Ed.), *The Cambridge handbook of computational psychology* (267-301). New York: Cambridge University Press.

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology*, *103*(3), 386–394.

Legare, C.H. (2010). Exploring explanation: Explaining inconsistent information guides hypothesis-testing behavior in young children. *Child Development*.

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

Legare, C.H. & Lombrozo, T. (2012). The unique and selective benefits of explanation for learning in early childhood.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*, 464-470.

Lombrozo, T. (2009). Explanation and categorization: how "why?" informs what?" *Cognition*, *110*, 248-253.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology, 61*, 303-332.

Lombrozo, T. & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition, 99*, 167-204.

Lombrozo, T & Gwynne, N. (under review). Explanation and inference: functional and mechanistic explanations guide property generalization.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*(4), 829–835.

Lupyan, G., & Rakison, D. (2007). Language is not Just for Talking Redundant Labels Facilitate Learning of Novel Categories. *Psychological Science*.

Maddox, W. T., & Ing, A. D. (2005). Delayed Feedback Disrupts the Procedural-Learning System but Not the Hypothesis-Testing System in Perceptual Category Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 100–107.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129*(4), 592-615.

Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1083.

McNamara, D.S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38*, 1-30.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge: The MIT Press.

Murphy, G.L. and Allopenna, P.D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory & Cognition,20*, 904–919.

Murphy, G.L. & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289-316.

Murphy, G. L., & Wisniewski, E. J. (1989). Feature correlations in conceptual representations. In G. Tiberghien (Ed.), *Advances in cognitive science, vol. 2: Theory and applications* (pp. 23-45). Chichester: Ellis Horwood.

Needham, D. R., & Begg, I. M. (1991). Problem-oriented training promotes spontaneous analogical transfer: Memory-oriented training promotes memory for training. *Memory & Cognition,* 19, 543–557.

Nokes, T. J., Hausmann, R. G. M., VanLehn, K., & Gershman, S. (2011). Testing the instructional fit hypothesis: The case of self-explanation prompts. *Instructional Science*.

Nosofsky, R., & Palmeri, T. (1994). Rule-plus-exception model of classification learning. *Psychological Review*.

Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and instruction*, *1*(2), 117–175.

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(3), 416.

Pennington, N. & Hastie, R. (1992). Explaining the evidence: tests of the story-model for juror decision making. *Journal of Personality and Social Psychology, 62,* 189-206.

Preston, J., & Epley, N. (2005). Explanations versus applications. *Psychological Science, 16*, 826-832.

Read, S. J., & Marcus-Newhall, A. R. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology, 65,* 429-447.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1141.

Rehder, B. (2006). When causality and similarity compete in category-based property induction. *Memory & Cognition, 34,* 3-16.

Rehder, B., & Ross, B. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1261-1275.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*(1), 1–29.

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary Educational Psychology*, 23, 90-108.

Rips, L. J. (1989). Similarity, typicality, and categorization. *Similarity and analogical reasoning*, 21–59.

Rittle-Johnson, B. (2006) Promoting transfer: the effects of direct instruction and self-explanation. *Child Development, 77*, 1–15.

Ross, B., Taylor, E., Middleton, E., & Nokes, T. (2008). Concept and category learning in humans. *Learning and Memory: A Comprehensive Reference*, *2*, 535–556.

Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. *Cognitive Psychology, 28*, 225-273.

Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31-58). New York: Cambridge University.

Sloman, S.A. (1994). When explanations compete: the role of explanatory coherence on judgments of likelihood. *Cognition*, *52*, 1-21.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411-1436.

Spalding, T. L., & Murphy, G. L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 525-538.

Spalding, T. L., & Murphy, G. L. (1999). What is learned in knowledge-related categories? Evidence from typicality and feature-frequency judgments. *Memory & Cognition, 27,* 856-867.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440-463.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435-467.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*(4), 732-749.

Walker, C. M., Williams, J. J., Lombrozo, T., & Gopnik, A. (2012). Explaining influences children's reliance on evidence and prior knowledge in causal induction. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Walker, C. M., Williams, J. J., Lombrozo, T., & Gopnik, A. (Under Review). The role of explanation in children's causal learning.

Wattenmaker, W., Dewey, G., Murphy, T., Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology*, *18*, 158-194.

Wellman, H. M., & Liu, D. (2007). Causal reasoning as informed by the early development of explanations. *Causal learning: psychology, philosophy, and computation*, 261–279.

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

Williams, J.J., Lombrozo, T., & Rehder, B. (2010). Why does explaining help learning? Insight from an explanation impairment effect. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2906-2911). Austin, TX: Cognitive Science Society.

Williams, J. J., Lombrozo, T., & Rehder, B. (2011). Explaining drives the discovery of real and illusory patterns. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Williams, J. J., Lombrozo, T., & Rehder, B. (Under Review). The hazards of explanation: overgeneralization in the face of exceptions.

Wisniewski, E. (1995). Prior knowledge and functionally relevant features in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21,* 449-468.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, *18*, 221–281.

Wong, R.M.F., Lawson, M.J., & Keeves, J. (2002). The effects of self-explanation training on students' problem solving in high-school mathematics. *Learning & Instruction, 12*, 233–262.

Table 1: Proportion of participants classified as using each basis for categorization in Experiment 1.

	Blank Labels		Informative Labels	
Pattern Use	**Free Study**	**Explain**	**Free Study**	**Explain**
Label-Relevant (100% Feet)	0.36	0.32	0.30	0.61
Label-Irrelevant (100% Antenna)	0.21	0.60	0.16	0.30
Item Similarity	0.42	0.07	0.50	0.07
Other	0.01	0.01	0.04	0.02

Table 2: Proportion of participants discovering each pattern in the free study conditions from Experiment 1 as a function of label type and self-reported explanation.

	Blank Labels		**Informative Labels**	
	Reported seeking explanations?			
Pattern Discovered	**No**	**Yes**	**No**	**Yes**
Both	0.04	0.06	0.03	0.14
Label-Relevant (100% Foot)	0.13	0.28	0.21	0.38
Label-Irrelevant (100% Antenna)	0.14	0.39	0.03	0.29
Neither	0.61	0.28	0.69	0.29

Table 3: Inferred pattern scope and relative pattern scope as a function of task and label type (blank vs. informative labels), in Experiment 3. Means are followed by standard deviations.

	Blank Labels (Glorp/Drent)		**Informative Labels (Outdoor/Indoor or Receiver/Transmitter)**	
Inferred Pattern Scope	**Write Thoughts**	**Explain**	**Write Thoughts**	**Explain**
Label-Relevant (75%)	69.2 (30.0)	68.3 (33.1)	63.8 (34.5)	71.3 (31.9)
Label-Irrelevant (100%)	80.3 (32.4)	85.4 (28.4)	82.0 (31.3)	77.8 (32.8)
Relative Pattern Scope	-11.1 (32.2)	-17.0 (38.5)	-18.3 (37.8)	-6.5 (33.7)

Table 4: Inferred pattern scope, relative pattern scope, and pooled pattern scope as a function of task and label assignment (random vs. representative labels), in Experiment 4. Means are followed by standard deviations.

	Random Labels		**Representative Labels**	
Inferred Pattern Scope	**Write Thoughts**	**Explain**	**Write Thoughts**	**Explain**
Label-Relevant Pattern	31.7 (60.2)	27.2 (66.8)	37.4 (58.1)	46.7 (62.7)
Label-Irrelevant Pattern	28.2 (60.4)	15.9 (53.6)	28.3 (54.6)	29.8 (58.6)
Pooled Pattern Scope	29.9 (60.3)	21.6 (60.2)	32.9 (56.3)	38.2 (60.6)
Relative Pattern Scope	3.5 (60.3)	11.3 (60.6)	9.1 (56.4)	16.9 (60.7)

Table 5: Inferred pattern scope, relative pattern scope, and pooled pattern scope as a function of task and label pair, in the representative labels conditions of Experiment 4. Means are followed by standard deviations.

	Receiver/Transmitter Labels		**Outdoor/Indoor Labels**	
Inferred Scope	**Write Thoughts**	**Explain**	**Write Thoughts**	**Explain**
Foot Pattern	37.6 (36.2)	40.8 (35.6)	52.4 (36.3)	63.7 (32.2)
Antenna Pattern	41.5 (36.9)	51.3 (36.4)	40.1 (39.9)	38.4 (37.5)
Pooled Scope	39.6 (36.6)	46.1 (36.0)	46.3 (38.1)	51.1 (35.0)
Relative Scope	-3.9 (31.4)	-10.4 (34.)	12.3 (37.3)	25.3 (33.6)

Figure 1. Study observations in Experiments 1 and 2. (a) Experiment 1 observations organized by category. (b) Experiment 2 observations organized by category.

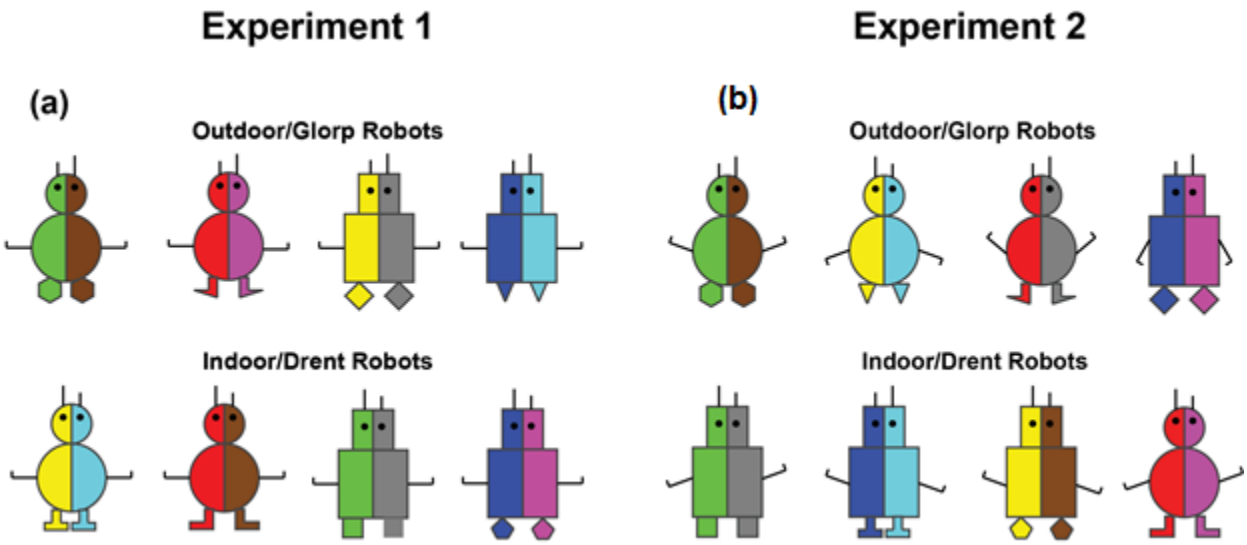

Figure 2. Results from Experiment 1 (a to c) and Experiment 2 (d to f). Error bars represent one standard error of the mean in each direction. **Pattern Discovery** (a & d): Proportion of participants who discovered the label-relevant and label-irrelevant patterns, and for Experiment 2, the additional partially reliable body shape and antenna patterns. **Number of Patterns Discovered** (b & e): Proportion of participants who discovered no patterns, exactly one pattern, or two or more patterns. **Conditional Discovery** (c & f): Of participants who discovered either the label-relevant or label-irrelevant pattern, the proportion that also discovered an additional pattern.

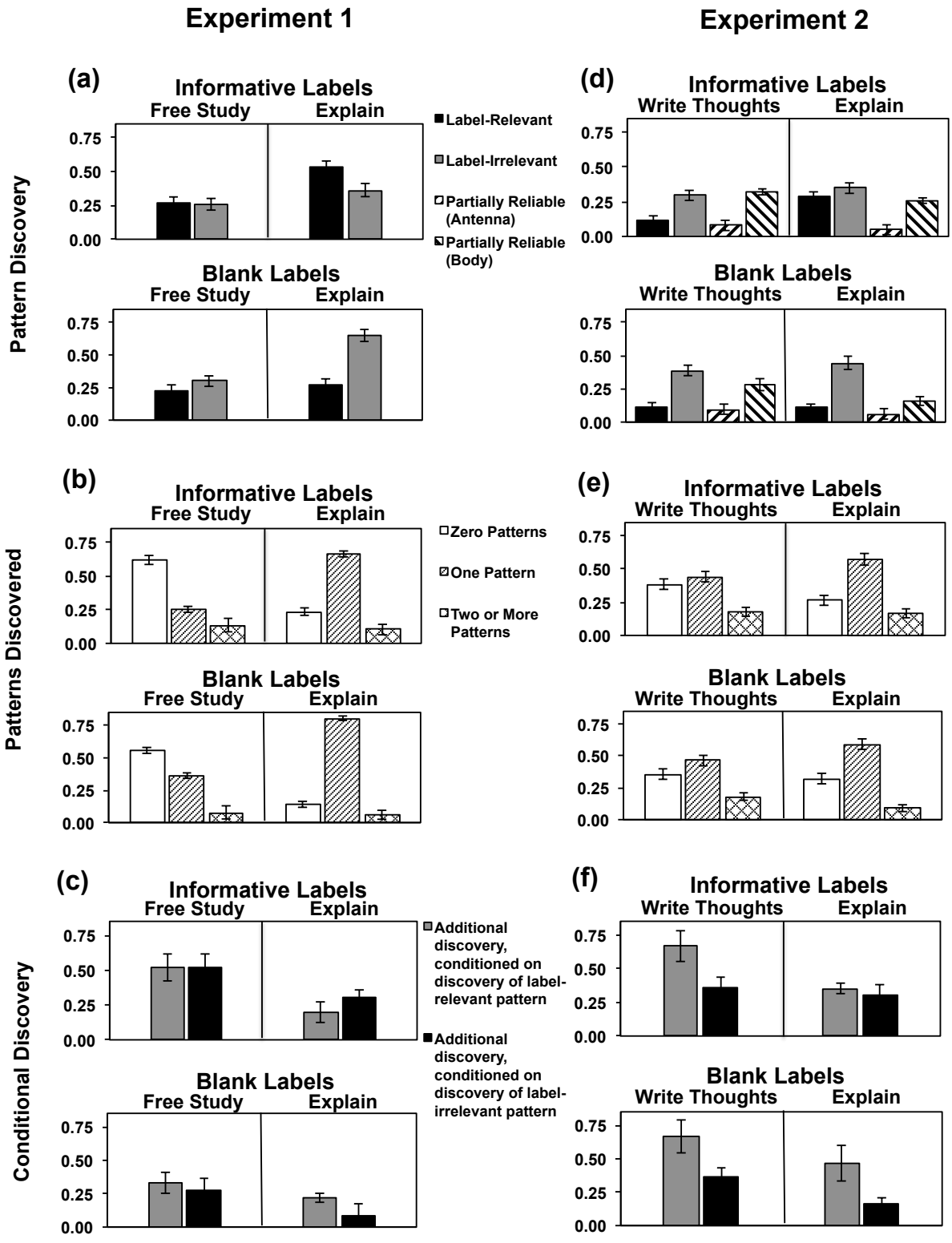
Experiment 1
Experiment 2
(a)
Informative Labels
Free Study
Explain
Blank Labels
Free Study
Explain
(d)
Informative Labels
Write Thoughts
Explain
Blank Labels
Write Thoughts
Explain
Pattern Discovery
Label-Relevant
Label-Irrelevant
Partially Reliable (Antenna)
Partially Reliable (Body)
(b)
Informative Labels
Free Study
Explain
Blank Labels
Free Study
Explain
(e)
Informative Labels
Write Thoughts
Explain
Blank Labels
Write Thoughts
Explain
Patterns Discovered
Zero Patterns
One Pattern
Two or More Patterns
(c)
Informative Labels
Free Study
Explain
Blank Labels
Free Study
Explain
(f)
Informative Labels
Write Thoughts
Explain
Blank Labels
Write Thoughts
Explain
Conditional Discovery
Additional discovery, conditioned on discovery of label-relevant pattern
Additional discovery, conditioned on discovery of label-irrelevant pattern
0.00
0.25
0.50
0.75

Figure 3: Study observations from Experiment 3: (a) when the foot pattern was the label-relevant pattern, (b) when the antenna pattern was the label-relevant pattern.

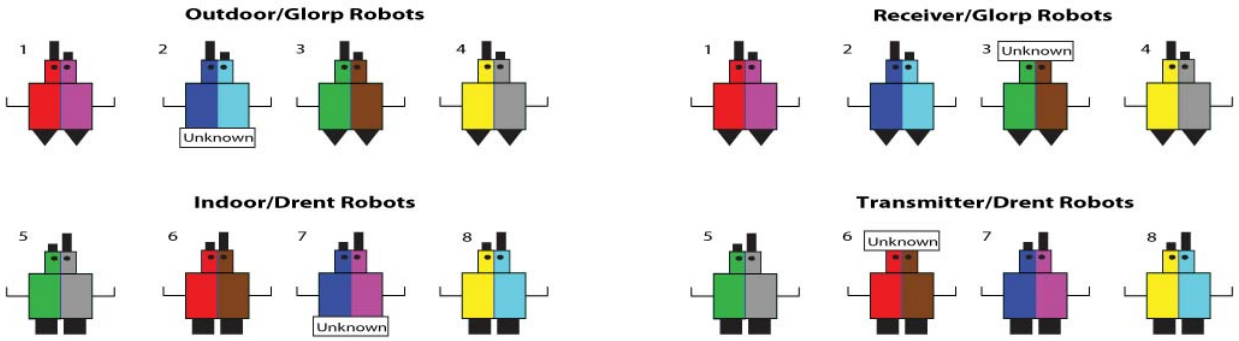

Figure 4: Extent to which the label-relevant pattern was used as a basis for generalizing category membership, as a function of task and label type, in Experiments 3 and 4. (a) Proportion of classifications consistent with label-relevant pattern in Experiment 3. (b) Average classification rating in Experiment 4, where higher numbers on 1-6 scale indicate greater consistency with the label-relevant pattern. Error bars correspond to one standard error of the mean in each direction.