Title
The Subsumptive Constraints Account of why explaining "why?" helps learning

Permalink
https://escholarship.org/uc/item/9vk786tp

Author
Williams, Joseph Jay

Publication Date
2013

Peer reviewed|Thesis/dissertation

The Subsumptive Constraints Account of why explaining “why?” helps learning

by

Joseph Jay Williams

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Tania Lombrozo, Chair

Professor Thomas L. Griffiths

Professor Barbara Y. White

Spring 2013

The Subsumptive Constraints Account of why explaining “why?” helps learning

by

Joseph Jay Williams

Abstract

The Subsumptive Constraints Account of why explaining “why?” helps learning

by

Joseph Jay Williams

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor Tania Lombrozo, Chair

Constructing explanations in response to why-questions plays an important role in learning and generalization – in young children, students, adults and scientists – but much remains to be understood about the underlying mechanisms and their implications for practical learning and teaching. The first chapter of this dissertation presents theSubsumptive Constraints account, which proposes that explaining why a fact or observation is true drives learners to understand how it can be subsumed as an instance of a pattern or generalization. Chapters two through four present nine experiments which test whether explaining exerts this selective effect, rather than a general boost for learning, examining contexts ranging from learning artificial categories to learning to predict other people's behavior. In these nine experiments, three specific predictions of the Subsumptive Constraints account were supported. The second chapter presents the first three experiments, which provide evidence that explaining category membership promoted the discovery of broad underlying patterns that supported generalization. Four experiments reported in the third chapter further revealed that seeking explanations increased learners' consultation of existing knowledge that could be used to discover patterns in their observations. Explaining also interacted with existing knowledge to influence which patterns learners generalized to novel contexts. The final two experiments in chapter four confirmed the Subsumptive Constraints account’s counterintuitive prediction that seeking explanations can impair learning, by revealing that learning was driven by an interaction between explanation and the reliability of underlying patterns. Explaining drove people to ignore specific observations that were exceptions to patterns, so that explaining fostered overgeneralization at the expense of accurate learning. The fifth chapter concludes this work. By drawing together theory and methodology from cognitive psychology, education, and philosophy, this dissertation provides novel theoretical, empirical, and practical insight into how generating explanations influences and improves learning. Explaining does not simply promote an 'all-purpose' boost to engagement and metacognition – and is therefore not always practically beneficial or optimal. By characterizing the particular nature of the processing constraints promoted by explaining – as selectively constraining the search for and evaluation of patterns that underlie specific observations – this novel Subsumptive Constraints account explains why explaining especially impacts learners’ use of existing knowledge in transfer and generalization, and provides insight into when to expect and how to maximize explanation’s practical benefits.

Table of Contents

List of Figures

Figures (and Tables) numbering uses the first digit to represent chapter, and the digit after the decimal place to represent figure (table) number within that chapter. E.g., Figure 3.2 is the second figure in chapter three.

List of Tables

1. Introduction

Understanding how people learn is one of the central questions in cognitive science: How do people acquire knowledge in a way that supports transfer and generalization to novel contexts? One activity that has an intimate connection with learning and understanding is asking "Why?" – generating why-explanations. Most of us have had the experience that seeking and generating explanations can foster learning and understanding – whether spontaneous and covert, out loud to oneself, or for another person. In fact, there is extensive evidence for the importance of explanation in learning, from research in science and mathematics education (Fonseca & Chi, 2011), cognitive psychology (Lombrozo, 2012), development (Wellman & Liu, 2006), artificial intelligence (DeJong, 2006; Mitchell et al, 1986), and philosophy (Lombrozo, 2011; Salmon, 1989).

Investigating explanation's role in learning is particularly important because seeking explanations is so ubiquitous: whether a child asking why a person is sad, an adult wondering why traffic is slow, or a scientist explaining why the planets move the way they do. The present research focuses on why-explanations, although the common sense extension of *explanation* can be construed more broadly: such as thinking aloud or explaining what one is thinking about, explaining the meaning of a sentence, or explaining how something works. This focus on why-explanations is chosen based on the assumption that it is broad enough to encompass a significant swathe of reasoning and learning, but not too heterogeneous or vaguely defined to resist experimental study, or preclude empirical support for meaningful generalizations about the relationship between "explanation" and learning. Conceptual analyses of what makes for good scientific explanations have been pursued for decades in philosophy of science (Salmon, 1989), and provide reason to expect many why-explanations to share common features. Moreover, why-explanations have been important theoretical constructs and targets of inquiry in psychological theories of conceptual representation (Carey, 1985; Murphy & Medin, 1985), causal reasoning (Wellman & Liu, 2006), and conceptual development (Gopnik & Meltzoff, 1997).

Despite the importance of explaining, there is much work to be done in characterizing the precise mechanisms that account for *why* explaining helps learning. This is not merely a theoretical problem, but also of practical concern. Without an understanding of the underlying mechanisms, there is a practical challenge in anticipating the contexts in which explaining is *most* helpful, ineffective, or even harmful.

The remainder of this first chapter reviews previous research on explanation and learning, and proposes a novel theory– the *Subsumptive Constraints* account. Three lines of research to test the predictions of the Subsumptive Constraints account are outlined and presented in the second, third, and fourth chapters. The fifth chapter concludes the dissertation, synthesizing the key empirical findings, considering implications for understanding explanation's role in learning, and discussing practical implications and directions for future research.

1.1. The effects of generating explanations in mathematics and science education

In educational contexts, there is a range of evidence for the *self-explanation effect:* generating explanations – of the meaning of an expository text, a worked-out example solution, or a correct answer to a problem – can promote learning and generalization, exceeding the benefits of learners' typical study strategies, such as reading material twice, or thinking aloud. Benefits of self-explanation have been found across the domains of Newtonian physics, geometry, algebra, probability, biology of the circulatory system, learning chess, and reading science texts (Chi et al, 1989; Chi et al, 1994;

McNamara, 2004; Renkl, 1997; Renkl et al, 1998; Rittle-Johnson, 2006; Siegler, 2002; Wong, Lawson & Keeves, 2002).

Self-explanation can foster: comprehension (McNamara et al, 2006), the acquisition of declarative knowledge, understanding of biological systems (Chi 1994), procedural and conceptual knowledge underlying problem-solving ability in math, physics and statistics (Chi, 1989; Siegler, 2002; Renkl, 1997; Rittle-Johnson, 2006), and the encoding of problem solutions (Crowley & Siegler, 1999). These benefits have been found in age groups ranging from university students (Chi et al, 1989) to 9th graders (Wong et al, 2002) to 5-year olds (Siegler, 1995).

1.2. Explanation and learning in philosophy, cognitive psychology, and development

Seeking, generating and receiving explanations plays a key role in human understanding of and reasoning about the world. Work in philosophy of science suggests that why-explanations satisfy features critical for supporting learning and reasoning: providing cues to likelihood (Salmon, 1971), causal structure (Salmon, 1984; Strevens, 2008), and subsuming and unifying generalizations (Kitcher, 1981; 1989; Strevens, 2008). Not just knowing a fact, but being able to explain it is central to science, and explanations are key components of theories (Woodward, 2010).

Cognitive science accounts of the structure of conceptual knowledge identify explanations as a key representational component (Carey, 1985; Murphy & Medin, 1985). Moreover, possessing or generating explanations impacts reasoning about causation (Koslowski, 1996; Ahn et al, 1995), induction of properties (Lombrozo, 2007; Rehder, 2007; Sloman, 1994), and probabilistic judgment (Koehler, 1991; Pennington & Hastie, 1992).

Developmental work shows that children's seeking of explanations emerges soon after they learn to talk – children as young as two start seeking explanations for *Why*? and *How?* (Callanan & Oakes, 1992; Chouinard, 2008; Hickling & Wellman, 2001). When adults do not answer with relevant explanations, 2, 3, and 4-year-old children re-ask the question or attempt to generate explanations themselves (Chouinard, 2008; Frazier, 2009). Legare et al (2010) demonstrated that children attempt to explain observations that contradict rather than conform to their prior beliefs, providing further evidence for children's explanations serving the role of learning, rather than simple discourse about beliefs or opportunities to socially demonstrate current knowledge.

Children's attempts to generate explanations can also have profound effects on their learning, despite not having much training in using language to construct explanations, relatively impoverished metacognitive insight into their reasoning, and significantly less prior knowledge to utilize than adults. In formal schooling contexts, 3rd and 4th grade (8 or 9 year old) children's understanding of arithmetic and mathematical equivalence is fostered when they are prompted to explain why correct answers are correct and why incorrect answers are wrong (Siegler, 2002; Rittle-Johnson, 2006). The benefits of self-generation are observed even when children are also given direct instruction on the correct solution, and extend especially to transfer and generalization to novel problems (Rittle-Johnson, 2006). In fact, Brown and Kane (1988) found that preschoolers were more likely to transfer a solution to an analogous problem when prompted to explain how the first problem was solved than if its solution was explained to them.

Even in everyday informal settings that are less structured and not as obviously pedagogical as schooling environments, prompts to explain can advance understanding of physical systems like balance beams (Pine & Messer, 2000). Asking children to generate explanation can even drive five year old's conceptual change in the domain of number conservation (Siegler, 1995) and preschoolers' (3 to 4 years) false belief reasoning in theory of mind (Amsterlaw & Wellman, 2006).

1.3. Why does explaining "why?" help learning?

Despite the great deal of empirical work investigating the effects of explanation, much less is known about the exact mechanisms that underlie its effects on learning. A broad *learning engagement* account suggests that explaining increases engagement with the task of learning: establishing the goal of generating explanations out loud in a social context may increase motivation to learn, the act of having to actively explain can boost attention to the materials being studied or events observed, and attempting to explain extends the amount of time spent effectively processing the available information (see for e.g., Siegler, 2002). This account suggests that explaining should effectively ramp up whatever processing people naturally engage in, providing a general boost to learning. However, it does not make precise predictions about when explaining will help vs. have no effect on learning, and what kind of learning and understanding explaining will be most helpful for promoting.

1.4. The Subsumptive Constraints Account

The *Subsumptive Constraints* account of explanation's role in learning proposed in this dissertation suggests that explaining *selectively constrains* the processing that people engage in – rather than boosting overall processing. The hypothesis is that explaining "why?" drives people to seek underlying generalizations, understanding how the fact or observation being explained could be anticipated as an instance of a broader pattern. For example, explaining why 2 x 6 = 12 invokes the principle that multiplication is repeated addition, and an explanation like "John is a teacher because he's a caring person" appeals to a regularity – that caring people are more likely to become teachers. Explaining why a particular fact is true therefore drives people to discover underlying patterns, which then form a basis for generalizing to novel contexts.

Each of chapters two, three, and four is a published journal paper that represents a distinct line of research testing the Subsumptive Constraints account. Chapter two further articulates the Subsumptive Constraints account and reports three experiments that test whether people's sensitivity to subsumptive constraints drives them to discover subtle underlying regularities, privileging those patterns that have broader scope in that they account for more observations. Chapter three presents four experiments that consider learning when multiple patterns are present, probing the joint effects of engaging in explanation and possessing prior knowledge about the target of explanation. These studies extend the work in chapter two by testing whether explaining promotes the discovery and use of patterns that *existing knowledge* suggests are broad in scope and likely to apply to novel contexts. Chapter four investigates a unique and counterintuitive prediction of the Subsumptive Constraints account: that explaining does not produce a general-purpose benefit, but influences learning through an interaction with the reliability of underlying patterns. The Subsumptive Constraints account counterintuitively predicts that explaining can *impair* learning, as when there are exceptions to underlying patterns, explaining can promote *over*generalization at the expense of learning about specific observations. The presentation of these three lines of work is followed by the concluding chapter, which considers implications of and future directions for this work.

2. The Role of Explanation in Discovery and Generalization: Evidence From Category Learning

2.1. Abstract

Research in education and cognitive development suggests that explaining plays a key role in learning and generalization: when learners provide explanations – even to themselves – they learn more effectively and generalize more readily to novel situations. This paper proposes and tests a *subsumptive constraints* account of this effect. Motivated by philosophical theories of explanation, this account predicts that explaining guides learners to interpret what they are learning in terms of unifying patterns or regularities, which promotes the discovery of broad generalizations. Three experiments provide evidence for the subsumptive constraints account: prompting participants to explain while learning artificial categories promotes the induction of a broad generalization underlying category membership, relative to describing items (Exp. 1), thinking aloud (Exp. 2), or free study (Exp. 3). Although explaining facilitates discovery, Experiment 1 finds that description is more beneficial for learning item details. Experiment 2 additionally suggests that explaining anomalous observations may play a special role in belief revision. The findings provide insight into explanation's role in discovery and generalization.

2.2. Introduction

Seeking explanations is a ubiquitous part of everyday life. Why is this bus always late? Why was my friend so upset yesterday? Why are some people so successful? Young children are notorious for their curiosity and dogged pursuit of explanations, with one "why?" question followed by another. Equally curious scientific researchers might wonder: Why is explaining so important?

Psychologists and philosophers have independently proposed that in explaining observations about the past, we uncover underlying structure in the world, acquiring the knowledge to predict and control the future (e.g. Heider, 1958; Quine & Ullian, 1970; Lombrozo & Carey, 2006; Lombrozo, 2006; but see Keil, 2006). For example, in explaining a friend's behavior, you might come to appreciate the extent of his or her ambition, which informs expectations about future actions. Moreover, explanations have been posited as central, organizing elements within intuitive theories (Carey, 1985) and conceptual representations (Murphy & Medin, 1985; Carey, 1991; Lombrozo, 2009), suggesting that the process of explaining may be intimately related to learning concepts and theories.

Everyday experiences provide many illustrations of explanation's effects on learning. In the course of explaining a concept or a problem's solution to another person, the explainer may generate a new insight or acquire a deeper understanding of the material, despite not having received any additional input from the world. Attempting to explain what one is reading or learning about similarly seems to promote learning, beyond simply memorizing or passively encoding.

In fact, empirical research in education and cognitive development confirms that the process of explaining can foster learning. There are benefits in explaining to others (Roscoe & Chi, 2007; 2008), and even in explaining to oneself. This phenomenon is known as the *self-explanation effect*, and has been documented in a broad range of domains: acquiring procedural knowledge about physics problems (Chi et al., 1989), declarative learning from biology texts (Chi et al., 1994), and conceptual change in children's understanding of number conservation (Siegler, 1995; 2002) and theory of mind (Amsterlaw & Wellman, 2006), to name only a few. Compared to alternative study strategies like thinking aloud, reading materials twice, or receiving feedback in the absence of explanations (e.g. Wong et al, 2002;

Chi, 1994; Siegler, 2002; Amsterlaw & Wellman, 2006), self-explaining consistently leads to greater learning. Notably, the greatest benefit is in transfer and generalization to problems and inferences that require going beyond the material originally studied. Explanation's role in learning and generalization is further underscored by a tradition of research in machine learning and artificial intelligence known as *explanation-based learning* (Mitchell, Keller, Kedar-Cabelli, 1986; DeJong & Mooney, 1986).

Why is the process of explaining so helpful for learning, and especially for deep learning: acquiring knowledge and understanding in a way that leads to retention and use in future contexts? Researchers have generated a number of proposals about the mechanisms that underlie explanation's beneficial effects on learning. These include the metacognitive consequences of engaging in explanation (such as identifying comprehension failures), explanation's constructive nature, explanation's integration of new information with existing knowledge, and its role in dynamically repairing learners' mental models of particular domains (for discussion, see Chi et al, 1994; Chi, 2000; Siegler, 2002; Crowley & Siegler, 1999; Rittle-Johnson, 2006). Generating explanations may also scaffold causal learning by focusing attention on cases for which the outcome is known (Wellman & Liu, 2006), and by encouraging learners to posit unobserved causes (Legare, Wellman, & Gelman, in press; Legare, Gelman, & Wellman, under review). Given the diversity of the processes that can underlie learning (Nokes & Ohlsson, 2005), it is likely that explanation influences learning via multiple mechanisms.

2.2.1. *Exploring the role of explanation in generalization*

In this paper we explore a *subsumptive constraints* account of explanation's effects on learning, which provides an account of why explaining particularly facilitates transfer and generalization. The hypothesis is that engaging in explanation exerts constraints on learning which promote the discovery of broad generalizations that underlie what is being explained. This hypothesis is motivated by work on the *structure* of explanations. By the structure of explanations, we mean the relationship that must hold between an explanation and what it explains for it to be genuinely explanatory. Little research in psychology has addressed this question directly (see Lombrozo, 2006), but a rich tradition from philosophy provides candidate theories that offer useful starting points for psychological theorizing (see Woodward, 2009, for a review of philosophical accounts of scientific explanation).

Accounts of explanation from philosophy have typically emphasized logical, probabilistic, causal, or subsumptive relationships between the explanation and what it explains. While there is no consensus, we focus on *pattern subsumption* theories, which have been advocated in past research on explanation within psychology (Lombrozo & Carey, 2006; Wellman & Liu, 2006). Pattern subsumption theories propose that the defining property of an explanation is that it demonstrates how what is being explained is an instance of a general pattern (for discussion see Salmon, 1990; Strevens, 2008). For example, in explaining a friend's current cold by appeal to the contraction of a germ from another person, a specific event (Bob's cold) is subsumed as an instance of a general pattern (the transmission of germs produces illnesses in people). A subset of these accounts further emphasizes *unification*: the value of explaining disparate observations by appeal to a single explanatory pattern (e.g. Friedman, 1974; Kitcher, 1981, 1989). The general pattern that germ transmission produces illnesses not only accounts for Bob's cold, but also a diverse range of other data about the occurrence and spread of diseases.

Subsumption and unification accounts of explanation predict the privileged relationship between explanation and generalization that is demonstrated by the self-explanation effect. If the explanations people construct satisfy the structural demands of subsumption, then the process of explaining will exert particular constraints on learning: the beliefs and inferences generated will be those that play a role in demonstrating how what is being explained conforms to a general pattern. Explaining will therefore guide people to interpret observations in terms of unifying regularities, and the information constructed in successful explanations will result in the induction or explicit recognition of generalizations that underlie what is being explained. Discovering and explicitly representing such generalizations can in turn facilitate transfer from one learning context to novel but relevant contexts. For example, attempting to explain an instance of a person's behavior might lead to an explanation that posits an underlying personality trait, providing the basis to generalize about that person in a range of new situations.

While it may seem intuitive that explanations unify and subsume, this approach to understanding the effects of explanation on learning and generalization has not been fully developed, nor has it been tested empirically. Previous work has typically emphasized the ways in which explanation contributes to processes known to facilitate learning, such as metacognitive monitoring and strategy or belief revision. Our account complements this work by taking a different tack, emphasizing that the process of explaining may exert particular constraints on the knowledge constructed in learning by virtue of the properties of explanations. The specific constraints we explore are those motivated by pattern subsumption and unification theories of explanation. In sum, the key, novel idea in a subsumptive constraints account is that explaining facilitates generalization because satisfying the structural properties of explanations exerts constraints that drive learners to discover unifying regularities, allowing transfer to novel contexts.

To test our hypothesis that explaining promotes the discovery of unifying regularities, we employ a task from cognitive psychology: learning artificial categories from positive examples. Exploring the role of explanation in the context of category learning has two important benefits. First, there are already reasons, both theoretical and empirical, to suspect an important relationship between explanation and category structure. Previous work on category learning suggests that categories are judged more coherent to the extent they support explanations (Patalano, Chin-Parker, & Ross, 2006), that different explanations differentially influence conceptual representations (Lombrozo, 2009), and that background beliefs that explain feature combinations facilitate category learning (Murphy & Allopenna, 1994) and influence judgments of a category member's typicality (Ahn, 2002). Moreover, compared to learning a category through classification and feedback, explaining items' category membership can lead participants to rely more heavily on features that are meaningfully related to the type of category (e.g. a social club) and less heavily on features that are diagnostic but not meaningful (Chin-Parker, Hernandez, & Matens, 2006), suggesting that explanation and classification with feedback may differentially impact the category learning process.

A second benefit of studying the role of explanation in the context of category learning comes from the opportunity to employ well-controlled artificial materials in a relatively well-understood task. Category members can vary along many dimensions, and prior research has identified multiple ways in which category membership can be extended from known to novel items. For example, category membership could be generalized on the basis of rules or definitions (Bruner et al, 1956; Nosofsky, Clark & Shin, 1989; Ashby & Maddox, 2004), rules with exceptions (Nosofsky, Palmeri & McKinley, 1994), similarity to prototypical summary representations (Posner & Keele, 1968; Hampton, 2006), similarity to specific exemplars of a category (Medin & Schaffer, 1978; Nosofsky, 1986), or representations that combine prototypes and exemplars (Love, Medin, & Gureckis, 2004). These competing accounts are a source of contemporary debate (e.g. Allen & Brooks, 1991; Medin, Altom, & Murphy, 1984; Murphy, 2002; Lee & Vanpaemel, 2008).

Our aim here is not to evaluate competing theories of conceptual structure, but rather to capitalize on what is already known about category learning and categorization to inform the design of our experimental task and stimulus materials. Specifically, if explaining constrains learners to seek unifying and subsuming regularities, those who engage in explanation should be more likely than learners engaged in a comparison task to discover broad generalizations underlying category membership.

2.3. Overview of Experiments

In three experiments, we investigate the effects of explaining on the discovery of regularities underlying two artificial categories of alien robots. The principal hypothesis is that attempting to generate explanations of category membership will constrain learners to interpret their observations in terms of general unifying patterns, which will facilitate the discovery of a subtle regularity underlying category membership.

To test this, the categories we employ support two generalizations about category membership: a feature of body shape that accounts for the membership of 75% of study items (square vs. round bodies, termed "the 75% rule"), and a more subtle feature concerning foot shape that perfectly accounts for membership of all items (pointy vs. flat feet, termed "the 100% rule"). The prediction is that explaining will drive learners to discover the 100% rule. Although the 100% rule is harder to discover than the 75% rule, the 100% rule provides the most unified account of category membership.

In each of the three experiments, participants study category members, either *explaining* why a robot might belong to a given category or engaging in a control task: describing items (Exp. 1), thinking aloud during study (Exp. 2), or free study (Exp. 3). Participants then categorize new items, are tested on their memory for the original study items, and are explicitly asked to report what distinguishes the two categories. Table 1 provides a useful reference for key differences across experiments, which are discussed in detail in the methods section for each experiment.

Three feature of this series of experiments are worth emphasizing. First, the explanation condition is compared to *three* different control conditions, which have complementary strengths and weaknesses. In particular, the conditions allow us to examine alternative accounts of the effects of explanation. If the benefits of engaging in explanation stem from increased attention to item details, then tasks such as describing that likewise engage attention should yield a comparable benefit, and the explanation condition should only outperform control conditions in Experiments 2 and 3. If the benefits of engaging in explanation stem from the role of articulating thoughts in language, then the explanation condition should outperform free study (Exp 3), but not describing (Exp 1) or thinking aloud (Exp 2), which similarly involve language. Our hypothesis, in contrast, predicts a benefit for explanation across all three control conditions.

Second, the use of artificial categories allows us to investigate our proposal about the role of explanation in learning while minimizing a potential role for alternative mechanisms. In particular, because artificial categories evoke minimal prior knowledge, it's unclear how accounts of explanation that emphasize the integration of new information with prior knowledge would account for a tendency to discover or employ one rule over the other. There are also no existing mental models of the domain for explaining to repair or revise. In fact, some accounts of explanation's role in judgment provide reason to predict that explaining should promote generalization based on the more salient 75% rule: explaining why a hypothesis is true has been shown to increase belief in that hypothesis (for a review see Koehler, 1991), suggesting that requiring participants to provide explanations for membership could entrench belief in initial hypotheses rather than promote discovery of more unifying but subtle alternatives. More broadly, if people articulate hypotheses when they provide explanations and are biased in confirming these initial hypotheses (Nickerson, 1998), explaining could have adverse effects on discovery.

Finally, in Experiments 2 and 3, participants are explicitly informed of a later categorization test. Making the task very explicit to all participants minimizes the possibility that effects of explanation are due to implicit task demands, such as the prompt to explain simply directing participants to discover a basis for category membership.

2.4. Experiment 1

In Experiment 1, participants learned about artificial categories of alien robots. Half were prompted to *explain* while learning, the other half to *describe*. Description was chosen as a comparison because it requires participants to verbalize, attend to the materials, and be engaged in processing materials for an equivalent length of time, but does not impose the same structural constraints as explanation. If explaining drives participants to interpret observations in terms of general regularities, then participants prompted to explain should be more likely than those who describe to discover the subtle but perfectly predictive rule (the 100% rule), and to use it as a basis for categorization.

2.4.1. Methods

Participants.

150 undergraduates and members of the Berkeley community (75 per condition) participated for course credit or monetary reimbursement.

Materials.

The task involved study items, test items, transfer items, and memory items.

Study items. Participants learned about two categories of robots from an alien planet, glorps and drents (study items are shown in Fig. 1). Each item was composed of four features: left color (blue, green, red, yellow), right color (brown, cyan, grey, pink), body shape (square or circular), and foot shape (eight different geometric shapes). Color was uncorrelated with category membership: every right and left color occurred exactly once per category. Body shape was correlated with category membership: three of four glorps (75%) had square bodies, and three of four drents had round bodies. Finally, each robot had a unique geometric shape for feet, but there was a subtle regularity across categories: all glorps (100%) had pointy feet while all drents had flat feet.

This category structure supports at least three distinct bases for categorizing new robots. First, participants could fail to draw any generalizations about category membership, and instead categorize new items on the basis of their similarity to individual study items, where similarity is measured by tallying the number of shared features across items.[1] We call this ‘item similarity’.

[1] To confirm that our criterion for similarity (number of shared features) corresponded to that of naïve participants, 25 participants who were not in the main studies were presented with each item from the categorization tests, and asked to indicate which study item was most similar. Across all items, the study items our criterion identified were the most frequently chosen.

Alternatively, participants could detect the correlation between body shape and category membership, called the “75% rule”, as it partitions study items with 75% accuracy. Finally, participants could discover the subtle regularity about pointy versus flat feet, called the “100% rule”, as it perfectly partitions study items.

Test items. Three types of test item (shown in Fig. 2) were constructed by taking novel combinations of the features used for the study items. Each type yielded a unique categorization judgment (of glorp/drent) according to one basis for categorization (100% rule, 75% rule, item similarity), and so pitted one basis for categorization against the other two. We call these item similarity probes (2 items), 75% rule probes (2 items), and 100% rule probes (4 items).

Transfer Items. These items used completely novel foot shapes to distinguish participants who genuinely drew an abstract generalization concerning “pointy” versus “flat” feet from those who simply recognized the importance of particular foot shapes. For each item, the 100% rule was pitted against item similarity and the 75% rule. Critically, while the test items introduced new combinations of old features, the transfer items actually involved new features (new foot shapes).

Memory Items. Twenty-three robots were presented in a memory test at the end of the experiment: 8 were the old study items (35%), and 15 were lures (65%). The lures consisted of test items that were categorized in the test phase, study items with foot shapes switched to those of another robot, study items with left- and right-hand-side colors switched, study items with body and colors changed, and study items with entirely new features (new colors, body shapes, foot shapes).

Procedure.

The task involved several phases: introduction, study, testing, transfer, memory, and an explicit report.

Introduction phase. Participants were instructed that they would be looking at two types of robots, glorps and drents, from the planet Zarn. They were given a color sheet that displayed the eight study items, in a random order but with category membership clearly indicated for each robot. Participants studied the sheet for 15 seconds, and kept it until the end of the study phase.

Study phase. Each of the eight study items was presented onscreen with its category label. Participants in the *explain* condition received instructions to explain why the robot was of that type (e.g. “This robot is a GLORP. Explain why it might be of the GLORP type.”), and those in the *describe* condition received instructions to describe the robot of that type (e.g. “This robot is a GLORP. Describe this GLORP.”). All participants typed their responses into a displayed text box, with each robot onscreen for 50 seconds. Participants were not allowed to advance more quickly nor take extra time. After the study phase the experimenter removed the sheet showing the 8 robots.

Test and transfer phases. The eight *test* items were presented in random order, followed by the eight *transfer* items in random order, with participants categorizing each robot as a glorp or a drent. To discourage participants from skipping through items without paying attention, a response was only recorded after each robot had been displayed for two seconds. Participants were informed of this delay and the screen flickered after the two-second period ended.

Memory phase. The eight study items (35%) and 15 lures (65%) were presented in a random order, and participants judged whether or not each robot was one of the original robots from the introduction and study phases. As in categorization, items had to be onscreen for two seconds.

Explicit report. Participants were explicitly asked whether they thought there was a difference between glorps and drents, and if so, to state what they thought the difference was. Responses were typed onscreen.

2.4.2. Results

Basis for Categorization.

To understand how explaining influenced what participants learned about categories, we evaluated participants' bases for categorizing novel robots. Explicit reports were coded into four categories (displayed in Table 2): 100% rule (explicitly mentioning pointy versus flat feet), 75% rule (square versus circular body shape), "item similarity" (reliance on nearest match from study), and "other"[2]. Responses were coded independently by two coders with 91% agreement, and the first coder's responses were used for analyses.[3] Table 2 suggests that more participants learned and utilized the 100% rule in the *explain* than in the *describe* condition, while more participants drew on the 75% rule in the *describe* than the *explain* condition.

This pattern was evaluated statistically by tests for association between condition and a coding category: in each test the four rows were collapsed into two, the first being the target coding category and the second all other coding categories combined. Participants' basis for categorization was more likely to be the 100% rule in the *explain* than the *describe* condition ($\chi^2(1) = 15.89$, $p < 0.001$), while the 75% rule was more prevalent in the *describe* than the *explain* condition ($\chi^2(1) = 19.56$, $p < 0.001$). 'Item similarity' and 'other' responses were not significantly associated with condition.

While both groups of participants drew generalizations about the basis for category membership, these findings suggest that those in the *explain* condition were more likely to discover the subtle 100% rule, which drew on an abstraction about foot shape to account in a unified way for the category membership of all study items.

Categorization of test and transfer items.

For the purposes of analysis, participants' categorization responses were scored as accurate if they corresponded to the 100% rule. Figure 3 shows test and transfer accuracy as a function of condition. Note that accuracy near 50% does not reflect chance responding, because items pit bases for categorization against each other. For example, for transfer items the two most common accuracy scores were 0% (perfectly systematic use of the 75% rule) and 100% (perfectly systematic use of the 100% rule).

A 2 (*task*: explain vs. describe) x 2 (*categorization measure*: test vs. transfer) mixed ANOVA was conducted on categorization accuracy. This revealed a main effect of *task* ($F(1,148) = 16.10$, $p <$

[2] The "Other" category further consisted of blank, "no difference", any other basis, and unclear or uncodable responses.

[3] Coding revealed that some participants reversed the two category labels. An example would be stating that glorps had flat feet or that drents had square bodies, when in fact the opposite was true. For all three experiments, when a participant's verbal response or post-experiment debriefing unambiguously indicated a switch in category labels, that participant's categorization responses were reverse coded.

0.001), with participants in the *explain* condition categorizing test and transfer items significantly more accurately than those in the *describe* condition. There was also a significant effect of *categorization measure* (F(1,148) = 13.46, p < 0.001), as test accuracy was higher than transfer accuracy. It is worth noting that the more accurate categorization of transfer items by participants in the *explain* condition (t(148) = 2.91, p < 0.01) suggests that they not only recognized the importance of foot shape in determining category membership, but abstracted away from the specific shapes used on study items to recognize the subtle property of having 'pointy' or 'flat' feet.

Categorization performance was also analyzed separately for each of the three types of test item (displayed in Fig. 2). Participants' categorization of the *100% rule probes* was more consistent with the 100% rule in the *explain* than the *describe* condition (t(148) =4.41, p < 0.001), while categorization of the *75% rule probes* was more consistent with the 75% rule in the *describe* than the *explain* condition (t(148) = 3.77, p < 0.001). There was no difference for item similarity probes (t(148) = 1.37, p = 0.17). These patterns of significance mirror those for explicit reports.

The test and transfer accuracy scores did not follow a normal distribution, as the items used pit bases for categorization against each other, making the modal responses either very high or very low. To ensure the reliability of the categorization accuracy findings, we additionally analyzed categorization accuracy using a non-parametric measure. Each participant was coded as relying on the 100% rule if 7 or 8 of the 8 transfer items were accurately categorized (all others were coded as *not* using the 100% rule). We relied on transfer item categorization as the most sensitive measure for use of the 100% rule: test items could be perfectly categorized by remembering specific foot shapes. A chi-squared test for association substantiated the finding that explaining was associated with greater discovery of the 100% rule ($\chi^2(1) = 10.37$, p < 0.01).

Memory for study items.

Because engaging in explanation may have drawn special attention to the anomalous items (an issue addressed in Experiment 2), memory was analyzed separately for items *consistent* with the 75% rule and for those that were *anomalies* with respect to the 75% rule. Memory performance is reported using the d' measure of sensitivity (see Wickens, 2002). The d' measure reflects participants' ability to discriminate old study items from new lures, with larger values indicating better discrimination. Figure 4 shows d' for consistent and anomalous items as a function of condition. The ability to discriminate *consistent* study items from similar but new robots was significantly better in the *describe* condition than the *explain* condition (t(148) = 2.24, p < 0.05). There was no difference in discrimination of *anomalous* study items (t(148) = 0.82, p = 0.41, Explain: 1.09, Describe: 0.81), although the interpretation of this null effect is limited by the fact that there were many fewer anomalous items than consistent items, and therefore greater variability in performance.

One explanation for the memory difference is that participants who explained were more likely to discover the foot rule and then neglect the details of study items. An alternative is that the processing activities invoked by explaining are not as effective for encoding item details as are those invoked by describing. For example, it could be that describing allocates attention to details, that explaining exerts a cost on the resources available for encoding details, or both. To examine this issue, d' for consistent items was examined as a function of basis for categorization as determined by explicit reports (see Figure 5). There was no significant difference in memory performance for those who explicitly cited the foot rule (t(30) = 0.88, p = 0.88) or the body rule (t(52) = 1.41, p = 0.16), but there was a significant difference in memory for participants coded in the "other" category (t(62) = 2.19, p < 0.05). This suggests that the memory difference across conditions is not due to discovery of the 100% rule leading participants to ignore item details, but may stem from a difference between the efficacy of explaining and describing for encoding specific details.

Coded Content of Explanations and Descriptions.

Each of the 8 explanations (or descriptions) a participant provided was coded for whether a feature was mentioned (foot shape, body shape, and color), and if that feature was mentioned in an 'abstract' or a 'concrete' way (for similar coding categories, see Wisniewski & Medin, 1994, and Chin-Parker et al, 2006). References were coded as *concrete* if they cited the actual feature, e.g. triangle/square/L-shaped feet, square/round body, yellow/green color. References were coded as *abstract* if they characterized a feature in more general terms, which could be applied to multiple features, e.g. pointy/flat feet, big/strange body, warm/complementary colors. Two experimenters coded explanations and descriptions independently, with agreement of 97% (analyses used the first coder's responses). Figure 6 shows the number of features mentioned in each coding category as a function of *task*. Two separate 2 (*task*: explain vs. describe) *x* 3 (*feature*: feet vs. body vs. color) ANOVAs were conducted on the total number of *concrete* (*abstract*) features mentioned by each participant. Participants in the *explain* condition cited a greater number of abstract features than those in the *describe* condition (a main effect of *task*, $F(1,148) = 24.72$, $p < 0.001$), while those in the *describe* condition cited more concrete features than those who explained (a main effect of *task*, $F(1,148) = 164.65$, $p < 0.001$). Individual t-tests confirmed that these two findings were reliable for all features (all *p*s < 0.05) except abstract references to body shape ($t(148) = 0.82$, $p = 0.41$).

It is worth noting that participants who explained were more likely to discover the 100% rule, even though those who described made references to feet more frequently. The coding data provide evidence against an attentional account of the effects of explaining on discovery, but are consistent with an attentional explanation for the enhanced memory found in the *describe* condition.

2.4.3. Discussion

Relative to describing – a control condition matched for time, engagement, and verbalization – explaining promoted discovery of a subtle regularity underlying category membership. This was reflected in participants' explicit reports about category structure, as well as in categorization accuracy on test and transfer items. Coding of actual explanations and descriptions revealed that explanations involved a greater number of feature references coded as 'abstract' (operationalized as being applicable to multiple feature instances), not just for feet, but also for color. Descriptions involved a greater number of references to features that were coded as 'concrete' (operationalized as identifying specific feature values), suggesting that although the specific foot shapes were attended to and referenced by participants who described, merely attending was insufficient for participants to discover the regularity about foot shape. Despite the advantage of explanation for discovery and generalization, describing led to better encoding of details for most items and improved performance on a later memory test.

Critically, the category structure employed provides evidence for the role of subsumption and unification in explanation: participants in both conditions identified a generalization underlying category membership (the 75% rule or the 100% rule), but participants who explained were more likely to discover and employ the 100% rule, which accounts for category membership in a more unified way. The findings from Experiment 1 also contribute to existing work on the self-explanation effect by demonstrating that the effects of engaging in explanation can extend to learning artificial categories, and exceed the benefits of describing (a condition matched for time and attention) when it comes to learning category structure. Finally, the finding that participants generated more abstract references not only for feet – which matched the category structure – but also for color – which did not – suggests that explaining facilitates the discovery of unifying regularities by encouraging explainers to represent the material being explained in more diverse or abstract terms.

One potential concern is that the prompt to explain may exert implicit task demands such as cluing in participants that they should find a basis for category membership, or that they should seek features or rules that differentiate the categories. This predicts that overall rule discovery would be higher in the explain condition, but in fact there was no significant interaction between condition and discovery of *a* rule (collapsing the 75% and 100% rule to one cell: $\chi^2(1) = 1.21$, $p = 0.27$). Experiment 2 takes further measures to address this and other potential issues with our interpretations of the findings from Experiment 1, and additionally explores the role of anomalies to the 75% rule in prompting discovery of the 100% rule.

2.5. Experiment 2

The first goal of Experiment 2 is to provide a stronger test of the hypothesis that explaining promotes discovery, building on the results of Experiment 1. Accordingly, Experiment 2 uses a *think aloud* control condition instead of the *describe* condition. In Experiment 1, it is possible that the difference between performance in the *explain* and *describe* conditions resulted exclusively from a tendency of description to *inhibit* discovery. Thinking aloud places fewer restrictions than describing on how participants engage with the task, while matching explaining aloud for verbalization. A second concern with Experiment 1 is that the prompt to explain may have exerted implicit task demands, such as providing a cue to participants that they should find a basis for category membership, or that the category was likely to have a sufficiently simple structure to permit accurate generalization. To address this issue, all participants in Experiment 2 are explicitly instructed that they will later be tested on their ability to remember and categorize robots in order to generate equivalent expectations about the task and category structure. If the benefits of explanation derive solely from these expectations, the participants who think aloud will have comparable benefits to those who explain, eliminating a difference between conditions.

The second goal of Experiment 2 is to further investigate the process of discovery by examining the role of anomalous observations with and without explanation. Experiment 1 demonstrates that explaining category membership promotes discovery, but required explanations or descriptions for robots both consistent and inconsistent with the 75% rule. It is thus unclear whether noticing or explaining the category membership of items inconsistent with the 75% rule played a special role in discovery.

One possibility is that participants who explained were more likely to realize that the anomalous items were inconsistent with the 75% rule, and that the recognition of exceptions was sufficient to reject the 75% rule and prompt discovery of the 100% rule. If this is the case, then drawing participants' attention to the anomalies should match and eliminate the benefits of explaining. Another possibility is that engaging in explanation encourages participants to consider regularities they might not otherwise entertain, whether or not they are confronted with an item that is anomalous with respect to their current explanation. This hypothesis predicts that explaining will promote discovery whether participants are explaining consistent or anomalous items.

A third possibility is that providing explanations will increase participants' confidence in their explanations, reinforcing the use of features invoked in whichever explanation is first entertained. Because the 75% rule is more salient, participants who are prompted to explain items consistent with the

75% rule may perseverate in its use and ultimately discover the unifying 100% rule *less* frequently than those who think aloud about these items.

A final possibility, and the one we favor, is that the conjunction of explaining and anomalous observations will lead to the greatest discovery by constraining learning such that participants are driven to discover a basis for category membership that subsumes the anomalies. Explaining anomalies may thus play a special role in discovery, beyond merely drawing attention to anomalies or explaining consistent observations. This possibility is supported by previous work suggesting that noticing anomalies is insufficient for belief revision (see Chinn & Brewer, 1993). Without engaging in explanation, anomalies may be ignored, discounted, or simply fail to influence learning.

To investigate these issues the study phase in Experiment 2 was modified so that learners provided explanations (or thought aloud) for only *two* robots: a glorp and drent that were both either *consistent* or inconsistent (*anomalous*) with respect to the 75% rule. The result was a 2 x 2 between-subjects design with *task* (explain vs. think aloud) crossed with *observation type* (consistent vs. anomalous). Participants viewed all of the robots and retained the sheet of 8 items, but the targets of explaining or thinking aloud were either *consistent* or *anomalous* observations.

2.5.1. Methods

Participants.

240 undergraduates and members of the Berkeley community participated (60 per condition) for course credit or monetary reimbursement.

Materials.

The materials were the same as in Experiment 1, with minor changes to study items and a modified set of memory items: the number of *consistent* lures was reduced and the number of *anomalous* lures increased. There were 8 old items and 12 lures.

Procedure.

The procedure followed that of Experiment 1, with the following changes.

Task Instructions. The initial instructions explicitly informed participants: “You will later be tested on your ability to remember the robots you have seen and tested on your ability to decide whether robots are GLORPS or DRENTS.” Participants were also reminded of this before explaining (thinking aloud) in the *study phase*.

Pre-study exposure. After participants received and viewed the sheet of robots, the introduction phase was augmented by presenting each of the 8 robots onscreen. A block consisted of displaying each of the 8 robots for four seconds with its category label, in a random order. Three blocks were presented, with a clear transition between blocks. This portion of the experiment ensured that participants across conditions observed and attended to the 8 study items, although only 2 items were displayed onscreen for the explain or think aloud phase.

Study phase. While participants provided explanations (descriptions) for all 8 robots in Experiment 1, the Experiment 2 study phase only presented two robots (one glorp and one drent) for 90 seconds each, with a warning when 30 seconds were left. In the *consistent* condition the two robots were randomly selected from the 6 consistent with the 75% rule, while in the *anomalous* condition the two robots were those inconsistent with the 75% rule.

Instructions to *explain* and *think aloud* were provided before the robots were displayed, so the prompt accompanying each robot was omitted. Participants were instructed to explain out loud or think aloud, and their speech was recorded using a voice recorder. The *explain* instructions were identical to Experiment 1, while the *think aloud* instructions were: "You should say aloud any thoughts you have while you are looking at the robots on the screen or on the paper. Say aloud whatever you are thinking or saying in your head, whether you are having thoughts about the robots, memorizing what they look like, or anything at all -- even if it seems unimportant."

Test, Transfer & Memory. The test, transfer, and memory phases were identical to Experiment 1, except that the restriction that responses could only be made after 2 seconds was removed.

Post-experiment questions about body shape. After the explicit report, participants were asked to recall how many glorps (drents) from the study items were square (round). Four questions were posed to elicit responses for each type of robot with each type of body shape, of the form "How many of the original GLORPS [DRENTS] had square [round] bodies?".

2.5.2. Results

Basis for Categorization and Categorization Accuracy.

Data on participants' basis for categorization (as reflected by explicit reports) and categorization accuracy both provided evidence that explaining promoted discovery of the 100% rule more effectively than thinking aloud. Explicit reports were coded as in Experiment 1 and are shown in Table 3. Agreement between coders was 91%, with analyses based on the first coder. As in Experiment 1, the contingency table was analyzed by collapsing the coding of explicit reports to two categories, giving a *discovery* factor with two levels: (1) reports reflecting discovery and use of the foot rule, (2) all other responses. A hierarchical log-linear analysis with backwards elimination was carried out on the *task* x *item type* x *discovery* contingency table, revealing a highly significant interaction between *task* and *discovery* ($\chi^2(1) = 21.91$, $p < 0.001$): explaining was associated with discovery. With post-hoc tests comparing individual conditions, discovery was more frequent in both explain conditions than in either think aloud condition ($\chi^2(1) = 8.71$, $p < 0.01$, $\chi^2(1) = 8.71$, $p < 0.01$; and $\chi^2(1) = 13.30$, $p < 0.001$, $\chi^2(1) = 13.30$, $p < 0.001$).

The benefit for explaining over thinking aloud was mirrored in categorization accuracy (see Figure 7). A 2 (*task:* explain vs. think aloud) x 2 (*item type*: consistent vs. anomalous) x 2 (*categorization measure*: test vs. transfer) mixed ANOVA revealed a significant main effect of *task* ($F(1,236) = 21.90$, $p < 0.001$), with more accurate categorization in the *explain* condition. The effect of *item type* ($F(1,236) = 3.35$, $p = 0.07$), and the interaction between *task* and *item type* ($F(1,236) = 3.35$, $p = 0.07$) were marginal. There was additionally a significant effect of *categorization measure* ($F(1,236) = 14.38$, $p < 0.001$), with test accuracy higher than transfer, and significant interactions between *categorization measure* and *task* ($F(1,236) = 4.71$, $p < 0.05$) and *categorization measure* and *item type* ($F(1,236) = 4.71$, $p < 0.05$), with transfer accuracy being a more sensitive measure of the differences between explaining and thinking aloud. Contrasts revealed that categorization accuracy was significantly higher in the explain-anomalous condition than in the explain-consistent condition ($F(1,118) = 5.83$, $p < 0.05$) or in either think aloud condition ($F(1,118) = 14.51$, $p < 0.001$, $F(1,118) = 12.68$, $p < 0.001$).

As in Experiment 1, categorization accuracy scores were not normally distributed, so a non-parametric analysis based on transfer accuracy was also carried out. The basis for categorization inferred from transfer accuracy (criterion for 100% rule: 7 or 8 of 8 transfer items correct) is shown in Table 4, and is referred to as *inferred discovery*. A log-linear analysis of *task* x *item type* x *inferred discovery* revealed an interaction between *task* and *inferred discovery* ($\chi^2(1) = 18.59, p < 0.001$), and also between *item type* and *inferred discovery* ($\chi^2(1) = 3.91, p < 0.05$). This suggests that both explaining and the presence of anomalies contributed to discovery. The trend towards greater discovery in the explain-anomalous condition than the explain-consistent condition was marginal ($\chi^2(1) = 2.76, p = 0.10$).

Across all three measures (basis for categorization based on explicit reports, categorization accuracy, and basis for categorization based on transfer accuracy), explaining was significantly associated with facilitated discovery. However, only basis for categorization based on transfer accuracy revealed a reliable effect of anomalous observations, and only categorization accuracy revealed a reliable difference between the explain-anomalous and explain-consistent conditions, with the other measures providing consistent but marginal support.

Explaining anomalies may have facilitated discovery, in part, by fostering the rejection of the 75% rule. To analyze reliance on the 75% rule, the factor *body use* was created with two levels: (1) explicit report of using the 75% rule, (2) all other responses. A log-linear analysis revealed a significant three-way interaction between *task*, *item type*, and *body use* ($\chi^2(1) = 4.35, p < 0.05$). Reliance on the 75% rule was more frequent in the explain-consistent than the explain-anomalous condition, approaching significance ($\chi^2(1) = 3.68, p = 0.055$), with no difference for the think aloud conditions ($\chi^2(1) = 0.95, p = 0.33$),

Memory.

Separate 2 x 2 ANOVAs were conducted on the discrimination measure d' for both consistent and anomalous items. There was an effect of *observation type* on discrimination of anomalous items ($F(1,236) = 21.53, p < 0.001$), simply reflecting that discrimination of anomalous items was better in the anomalous conditions. No other effects were significant (all p*s* > 0.30). Memory for the original items did not appear to be differentially influenced by explaining versus thinking aloud.

Post-experiment questions about body shape.

Due to an experimental error, responses to the questions about how many robots in each category had a particular body shape only ranged over 1, 2, 3, and 4 (participants could not say '0'), and responses were only recorded for the final 103 participants. We therefore exclude a full analysis, and employ the data we do have only as an index of participants' awareness of anomalies to the 75% rule across conditions. As a measure of whether a participant realized there were exceptions to the trend in body shape, if a participant stated that there were 4 square glorps *or* 4 round drents they were coded as not noticing the anomaly.[4] According to this measure, the proportions of participants who noticed the anomalies were as follows: Think aloud-consistent, 65% (17 of 26); think aloud-anomalous, 64% (16/25); explain-consistent, 74% (20/27); explain-anomalous, 92% (23/25). This suggests that a sizeable number of participants noticed the anomalies in all conditions. In particular, the majority (more than

[4] We interpret this data as suggesting that a sizeable proportion of participants noticed the anomaly, and so we used this measure because it is conservative: using a '4' answer to *both* questions as the measure for not noticing the anomaly identifies even more people as having noticed the anomaly.

50%) of participants in the explain-consistent condition noticed and recalled the anomalies ($\chi^2(1) = 6.26$, $p < 0.05$).

2.5.3. Discussion

Building on the findings from Experiment 1, Experiment 2 found that engaging in explanation facilitated discovery relative to a think aloud control condition that exerted fewer restrictions on processing than describing. This effect of explanation occurred despite the fact that participants were informed that they would later have to categorize robots, and were given an opportunity to study each robot multiple times before the explain / think aloud manipulation.

The difference across explanation conditions additionally provides some suggestive evidence that explaining anomalous observations may be more effective for accurate learning and generalization than explaining observations consistent with current beliefs. Explaining anomalies seems to have prompted participants both to reject conflicting beliefs (the 75% rule) and to discover broader regularities (the 100% rule), although the former effect was more reliable than the latter. As suggested by the questions about body shape, it is possible that larger or more reliable effects were not observed because participants in the explain-consistent condition overwhelmingly noted the anomalies, and examining the sheet of all 8 robots or recalling anomalies from the pre-study phase may have led participants in this condition to seek a more unifying explanation for category membership even while explaining consistent items.

The two think-aloud conditions led to comparable rates of discovery, with hints of a benefit for thinking aloud while observing anomalies. However, even in the think-aloud–anomalous condition, discovery fell reliably short of that in the explanation conditions. Although attention was drawn to anomalies and the design provided implicit demands to incorporate these items into beliefs about category membership, only a small number of participants discovered and employed the 100% rule. This suggests that attending to, observing, and thinking aloud about anomalies is insufficient to promote discovery; a process like explaining is additionally required.

There were no significant differences in memory between the explain and think aloud conditions. This could suggest that the memory difference in Experiment 1 was driven by description's facilitation of memory, not a memory cost for explanation. However, a more conservative interpretation of the null effect may be warranted: participants received considerable exposure to study items outside of the explain vs. think aloud phase, potentially minimizing the effect of this manipulation on memory.

2.6. Experiment 3

The final experiment was a replication in which participants in the control condition were not instructed to perform a specific task, and all of the robots were simultaneously presented for study. This control condition aimed to address the possibility that the previous benefits of explanation were driven by describing and thinking aloud inhibiting discovery, not by explanation promoting discovery. If our interpretations of Experiments 1 and 2 are correct, explaining should promote discovery relative to a condition in which participants are not required to perform an alternative task.

2.6.1. Methods

Participants.

Participants were 120 undergraduate students enrolled in a psychology course who received course credit for completing the experiment as part of an hour of online surveys.

Materials.

Participants saw all 8 robots onscreen in an image that was identical to that in the previous experiments, except that each robot also had an associated number (the glorps were labeled 1 through 4, the drents 5 through 8). Due to time constraints, fewer test, transfer and memory items were presented. Test items consisted of 1 item similarity probe, 1 75% rule probe, 1 item that received the same classification from all 3 bases, and 4 100% rule probes. There were 4 transfer items. Memory items consisted of 4 old items and 4 lures.

Procedure.

Participants completed the experiment online. The instructions informed them that they would be learning about alien robots and that they would later be tested on their ability to remember and categorize robots. An image appeared onscreen that showed all 8 robots along with labels and numbers, and informed participants: "These are 8 robots on ZARN. This image will be onscreen for 2 minutes." In the *explain* condition participants were also told: "Explain why robots 1, 2, 3 & 4 might be GLORPS, and explain why robots 5, 6, 7 & 8 might be DRENTS." and typed their response into a text box. In the *free study* condition participants were told: "Robots 1, 2, 3 & 4 are GLORPS, and robots 5, 6, 7 & 8 are DRENTS."

The image was fixed to be onscreen for 2 minutes. After it was removed, participants categorized test and transfer items, completed the memory test, and answered several additional questions. Question 1 was "What do you think the chances are that there is one single feature that underlies whether a robot is a GLORP or a DRENT - a single feature that could be used to classify ALL robots?" and responses were 0, 25, 50, 75, or 100 %. Question 2 asked participants to report whether they thought there were noticeable differences between glorps and drents, and if they thought there were, what those differences were.

Question 3 showed a green screen ostensibly placed in front of a robot, obscuring all features except for the edges of its arms that extended beyond the sides of the screen. Participants were shown four questions they could ask about the robot, and required to specify the order in which they would ask the questions if they had to decide whether the obscured robot was a glorp or drent. The options were ordered randomly, and were: 1. What color is it? 2. What does its body look like? 3. What do its feet look like? 4. I would not ask any more questions - they will not be helpful. (The results from question 3 were redundant with other measures, and are hence not reported.)

Question 4 asked participants to state which features of glorps and drents they used in categorizing robots.

Question 5 asked "When the image of 8 numbered robots was onscreen, were you trying to explain why particular robots were glorps, and why particular robots were drents?" and the randomly ordered responses were "Yes", "Not sure", and "No".

Question 6 asked whether participants had previously been in an experiment that used these materials.[5]

2.6.2. Results and Discussion

[5] Three participants who indicated previous participation were dropped from the analysis.

Basis for categorization and categorization accuracy.

Basis for categorization was coded from participants' explicit reports and the features they reported using in categorization, and is shown in Table 5. As in previous experiments, the reports were independently coded by two experimenters: agreement was 87% and analyses are based on the first coder's responses. Figure 8 shows test and transfer accuracy as a function of condition.

Explaining was significantly associated with higher rates of discovery and use of the 100% rule, as revealed both in explicit reports ($\chi^2(1) = 4.09$, $p < 0.05$) and in categorization accuracy (a main effect of *task* in a *task* x *categorization measure* ANOVA, F(1,118) = 7.02, $p < 0.01$). Explaining was also significantly associated with reduced use of the 75% rule ($\chi^2(1) = 4.66$, $p < 0.05$).

Memory.

There was no significant difference in memory (as measured by discrimination, d') for consistent items (t(118) = 1.60, p = 0.11: Explain: 0.89, Free study: 0.35) or anomalous items (t(118) = 0.26, p = 0.80: Explain: 0.74, Free study: 0.62).

Likelihood of underlying feature.

There was no difference across conditions in how likely participants thought it was that there was a single feature underlying category membership (t(118) = 0.65, p = 0.52; Exp: 37.9, Con: 42.1). Moreover, there were no significant differences when the analysis was performed separately for each coded basis for categorization: feet, body shape, or "other" (all *p*'s > 0.12). This suggests that the effect of the prompt to explain was not simply to communicate to participants that there was a regularity present.

Self-report of explaining.

As expected, a greater number of participants reported explaining category membership in the explanation condition than in the control condition (see Table 6). However, there was not a significant association between condition and response cell ($\chi^2(1) = 2.41$, p = 0.30). It is interesting that the prompt to explain was effective even though a sizeable number of participants reported spontaneously trying to explain in the *free study* condition. It may be that explaining manifests its effects in a graded way: not simply as a function of whether or not participants attempt to explain, but in the frequency of generating explanations or in the degree to which participants persist in explaining.

To analyze the independent roles of the prompt to explain and reported efforts to explain, a log-linear analysis was performed on the following three factors: *discovery* (explicitly reported foot discovery, or not), *task* (explain vs. free study), and *explain-report* ('yes' response to question about explaining, vs. 'not sure' and 'no'), provided in Table 7. There were significant interactions between *task* and *discovery* ($\chi^2(1) = 4.17$, $p < 0.05$), and also between *explain-report* and *discovery* ($\chi^2(1) = 13.96$, $p < 0.001$). This suggests two additive effects, and provides further evidence for the importance of explaining in discovery. Prompts to explain tended to facilitate discovery, and to the extent that the prompt to explain was obeyed (in the *explain* condition) or that participants engaged in spontaneous explanation (in the *free study* condition), discovery was also promoted.

2.7. General Discussion

Experiments 1-3 find that participants prompted to explain why items belong to particular categories are more likely to induce an abstract generalization (100% rule) governing category membership than are participants instructed to describe category members (Exp. 1), think aloud during study (Exp. 2), or engage in free study (Exp. 3). These findings provide evidence for a subsumptive constraints account of explanation's effects: that explaining exerts constraints on learning which facilitate the discovery of regularities underlying what is being explained, and thereby support generalization.

Our findings support an account of explanation that emphasizes subsumption and unification. If good explanations are those that show how what is being explained is an instance of a general pattern or regularity, then trying to explain category membership should drive participants to discover patterns and regularities. And if explanations are better to the extent they unify a greater number of observations, explaining should drive participants to induce broad generalizations that surpass the 75% accuracy afforded by body shape, and support generalization to new contexts.

In addition to providing insight into the constraints exerted by explaining, Experiments 1 and 2 suggest that the mechanisms by which explaining promotes discovery involve *abstraction* and *anomalies*. In Experiment 1, participants who explained not only generated more abstract feature references about foot shape than did those who described, but also did so about color, even though the category structure did not support obvious generalizations about color. This suggests that explaining encourages learners to redescribe the material being explained in terms of new and potentially abstract features, because this redescription helps satisfy the demands of explanation: greater unification. Consistent with this possibility, Wisniewski & Medin (1994) reported that people's prior knowledge guided the construction of abstract features and hypotheses about category items. Explaining may invoke prior knowledge that guides such feature construction.

Experiment 2 provided some evidence for the value of explaining anomalies in driving discovery and revising beliefs. Even though the think aloud-anomalous condition drew attention to anomalies, attending to anomalies did not promote learning as effectively as explaining them. Providing explanations for anomalies may ensure that information inconsistent with current beliefs is not ignored or discounted, but used in a way that drives discovery and belief revision (for related discussion see Chinn & Brewer, 1993). In particular, explaining anomalies may lead to the rejection of beliefs inconsistent with the anomalies in addition to promoting the construction of more unifying alternatives.

While the reported experiments are the first to extend self-explanation effects to an artificial category learning task, we do not see this extension as the primary contribution of this work. After all, the powerful effects of explanation on learning and generalization have been well established in previous research using complex and educationally relevant materials. Rather, the current experiments help fill gaps in this previous research by testing a proposal about why explaining might play the role it does in generalization. Using a more controlled task and stimuli allowed a rigorous test of the hypothesis that explaining drives the discovery of regularities that support generalization, but necessarily reduced the richness of the explanations involved to concern a defining feature. Having established the current approach as a successful strategy for investigating explanation, an important direction for future research on explanation's role in discovery and generalization will be to reintroduce real-world complexity while maintaining experimental control.

Understanding why explaining promotes generalization has implications for both cognitive psychology and education. For example, the memory findings from Experiment 1 suggest that explanation and description may be complementary learning strategies, with explanation promoting the discovery of regularities, and description supporting memory for item details. In many learning contexts encoding facts and details is essential, and may even be a prerequisite to future learning. For example, in domains where learners have insufficient knowledge to induce underlying regularities, explaining is unlikely to facilitate generalization through discovery. Engaging in activities like description, memorization or receiving directed instruction may be more useful and promote the acquisition of background knowledge that supports future discovery. It follows that in certain contexts, explaining may not be the most effective strategy for learning.

In fact, one counterintuitive prediction of our account is that explaining should hinder learning under certain conditions. If explaining consistently exerts the constraint that observations are interpreted in terms of unifying patterns, it may be less helpful or even harmful in unsystematic domains, or when too little data is available (for recent evidence that explanations aren't always beneficial, see Kuhn & Katz, 2009; Wylie, Koedinger, & Mitamura, 2009). In the absence of true regularities, explaining random observations may lead people to induce incorrect generalizations. An anecdotal example of this might be elaborate 'conspiracy theories'. Explaining small samples of unrepresentative observations might also lead to the induction of incorrect patterns that do not generalize. Speculatively, this could be the case in inferring illusory correlations, such as in social stereotyping (e.g. Hamilton, 1981). One future direction is assessing whether documented biases or misconceptions can be understood from this perspective, and exploring the possibility that explaining can hinder accurate learning through "illusory discovery".

In the remainder of the discussion, we consider alternative interpretations for the effects of explanation and the relationship between explanation and other learning mechanisms. We conclude by highlighting a few promising future directions.

2.7.1. Alternative interpretations of the effects of explanation

An inherent difficulty in investigating the effects of prompts to explain is interpreting the differences between explaining and control conditions. In Experiment 1, it is possible that the difference between conditions was due to describing inhibiting discovery, with no benefit to explaining. However, explaining was also found to have an effect relative to thinking aloud (Exp. 2), which did not impose the restrictions that describing item features does, and relative to free study (Exp. 3). In Experiment 2, it is possible that thinking aloud distracted participants from crucial aspects of the task, but a difference was also found when participants were required to attend to items by describing (Exp. 1) or did not have to perform any potentially distracting task (Exp. 3). Finally, the findings from Experiment 3 might be explained in terms of explaining increasing attention to item features or requiring the use of language. But this kind of attentional account wouldn't predict the differences observed in Experiment 1, and appeals to language or articulation are less plausible in light of the benefits for explaining found in Experiments 1 and 2.

In sum, while each finding may allow for alternative explanations, the plausibility of these alternatives is decreased in the context of all three experiments. Moreover, there are reasons to expect describing, thinking aloud, and free study to *help* discovery rather than hurt it: by promoting attention, requiring articulation, and allowing participants to select *any* learning strategy. It is noteworthy that explaining had a beneficial effect above and beyond all three of these comparison conditions, which arguably intersect with activities typically engaged in by students and other learners.

Another set of alternative interpretations concern task demands. One possibility is that prompting participants to explain exerted its effects by indirectly communicating to participants that they should search for a basis for category membership. For example, the pragmatics of the explanation prompt might suggest the experimenter designed the categories to have differentiating features and expected participants to search for differences between categories. However, Experiments 2 and 3 explicitly informed participants that they would have a later categorization test in both the explain and control conditions. If explanation's only effect was to suggest to participants that they should find a feature that could be used to differentiate the categories, these instructions should have led to identical learning in the explanation and control conditions of Experiment 2 and 3. This alternative interpretation is also less plausible in light of the fact that participants in *both* the explain and control conditions identified body and foot shape features that figured in categorization rules: even without a prompt to explain participants sought differences between the categories. The critical difference was whether the differentiating rule they identified was the 75% rule or the 100% rule, which resulted in greater unification and subsumption.

Another task demand interpretation could be that being told to explain helps merely because it suggests to participants that they should find a defining feature underlying category membership. While this interpretation has some intuitive appeal, additional assumptions are needed in understanding *why* people would interpret a prompt to explain as concerning a defining feature, rather than some other structure. In fact, the subsumption account predicts that the prompt constrains learners to seek knowledge that shows how what they are explaining is an instance of a general pattern, which in this particular task could be knowledge about defining features or criteria that specify necessary and sufficient conditions for category membership. It is not clear that this particular 'task demand' interpretation competes with an account in terms of subsumptive constraints.

2.7.2. Relationship between explanation and other cognitive processes

In this section, we consider the relationship between explanation and other cognitive processes that could play a role in learning -- such as depth of processing, rule learning, hypothesis testing, and comparison.

Interpreting the effects of explaining raises the question of its relationship to depth of processing in memory research (Craik & Lockhart, 1972). For example, do effects of explaining reflect a standard depth of processing effect? On this point, it is worth noting that participants who explained processed items in a way that resulted in *worse* memory than did those in the describe control condition. One way to relate explanation and depth of processing is to interpret this work as a specific proposal about what the deeper processing prompted by explaining comprises. What seems most important about the prompt to explain is that it drives learners to allocate attention to the *right* features and patterns and to process items in an *appropriate* way for discovering regularities that can be constructed on the basis of current knowledge. We would argue that explaining exerts constraints that drive deeper processing of a specific kind: processing that is directed towards satisfying the subsumptive properties of explanation and so results in the discovery of regularities.

Some theories of category learning have emphasized the role of rules (e.g. Bruner et al, 1956), and aim to characterize the conditions under which categorization is more rule-like or more exemplar or prototype-based (Allen & Brooks, 1991; Sloman, 1996; Lee & Vanpaemel, 2008). It may be that the effect of explanation on category learning can be interpreted as increasing participants' use of rule-based strategies. However, explaining does not merely encourage the use of rules per se, as it promoted discovery of the 100% rule above the 75% rule. Models of category learning that favor rules with the fewest exceptions (Nosofsky, Palmeri, & McKinley, 1994; Goodman et al, 2008) predict this result and naturally correspond to explanation's subsumption and unification constraints. More broadly, if it is the case that "good" rules are those that make for good explanations, research on explanation and research on rule-based models may be mutually informing. However, to the extent that explaining exerts constraints other than subsumption and unification (such as relating observations to prior causal knowledge), people's learning about categories through explanation may be less amenable to rule-based accounts.

In these experiments, we interpret the findings of enhanced discovery as a consequence of explainers converging on knowledge that satisfies properties of explanation like subsumption and unification. But these results could also be understood in terms of hypothesis testing. Perhaps participants in the explain condition formulated and tested hypotheses about category membership, which facilitated rejection of the 75% rule and discovery of the 100% rule. Another possibility is that participants in the explain condition engaged in the comparison of items, so that processes like structural alignment of item features facilitated the induction of the subtle 100% rule (e.g. Yamauchi & Markman, 2000).

Instead of regarding these possibilities as mutually exclusive alternatives, they can be thought of as complementary proposals about which cognitive processes are recruited by explainers to satisfy the demands of explanation. Constructing explanations exerts a specific constraint on learning: that observations be interpreted in terms of unifying patterns. In satisfying this constraint, explainers may be driven to test different hypotheses when current beliefs are found to provide inadequate explanations, and may engage in comparison and structural alignment of category members in the service of identifying unifying patterns. More generally, explaining may recruit a range of cognitive processes in order to produce explanations that satisfy particular structural properties. The cognitive processes recruited will likely correspond to those identified by previous research as effective in facilitating learning and discovery: logical, inductive, and analogical reasoning, comparison, hypothesis testing, and so on. In fact, Chi et al's (1994) coding of self-explanations found that approximately one-third of explanations reflected the use of other learning mechanisms, such as logical and analogical reasoning.

2.7.3. Future Directions & Conclusions

These experiments suggest the utility of subsumption and unification, but there is a great deal of future research to be done in exploring how properties of explanation play a role in learning. A central question for future research concerns which kinds of patterns or regularities are judged explanatory, and hence likely to be discovered through explanation. Patterns that are consistent with prior knowledge and law-like are excellent candidates, but distinguishing law-like generalization from accidental generalizations is notoriously difficult (see, for example, Carroll, 2008 in philosophy, and Kalish, 2002 for a relevant discussion from psychology). Theories of explanation from philosophy of science provide proposals about other important properties of explanations, such as identifying the causes relevant to bringing about what is to be explained. Does explaining especially privilege the discovery of causal regularities?

Research in psychology has distinguished mechanistic and functional explanations (see Lombrozo & Carey, 2006; Lombrozo, 2009; Kelemen, 1999) and explored the role simplicity plays in the evaluation of explanations (see Lombrozo, 2007; Bonawitz & Lombrozo, under review). Do mechanistic and functional explanations play different roles in the acquisition of knowledge? Does people's preference for simple explanations have consequences for learning? If a function of explanation is to support generalization (see Lombrozo & Carey, 2006, for a proposal to this effect), then subsumption and unification may trade-off with other properties of explanations that support generalization.

The focus in this paper has been on human learning, but the proposal that the subsumptive properties of explanation exert constraints that can contribute to discovery and generalization may also inform machine learning, where algorithms involving explanation have been proposed (e.g. Lewis, 1988). Approaches in artificial intelligence referred to as "explanation-based learning" and "explanation-based generalization," for example, provide algorithms for learning generalizations by explaining one or a few examples (e.g. Ahn, Brewer, & Mooney, 1992; Mitchell, Keller, Kedar-Cabelli, 1986; DeJong & Mooney, 1986). These algorithms employ a circumscribed conception of explanation (as a process of deduction), but employing a broader notion of explanation that is informed by the kind of approach we adopt here may be useful in extending such algorithms.

Our experiment is the first (that we know of) to draw on theory from philosophy of science and methodology in cognitive psychology to examine the effects of explaining on learning, a phenomenon empirically established in educational and developmental psychology. We believe that the integration of these disciplines has a great deal of promise. Theories of explanation from philosophy can provide novel insights into the role of explanation in learning and generalization. And by using artificial categories, a research strategy from cognitive psychology, one can control participants' prior beliefs and provide a more precise characterization of the role of explanation in the discovery of generalizations. We hope that these experiments contribute to the utilization of philosophical work on explanation, and further explorations at the intersection of educational and cognitive psychology. Drawing on insights from each discipline offers the opportunity to gain a deeper understanding of the key role explaining plays in learning.

Figure 1. Study items in Experiment 1.

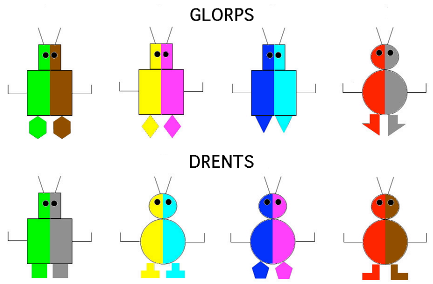


Figure 2. Examples of three types of test items from Experiment 1.

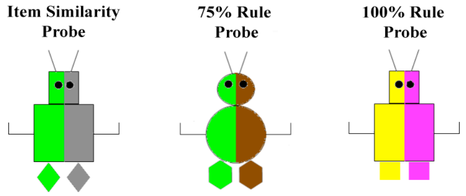


Figure 3. Categorization Accuracy on Test and Transfer Items in Experiment 1.

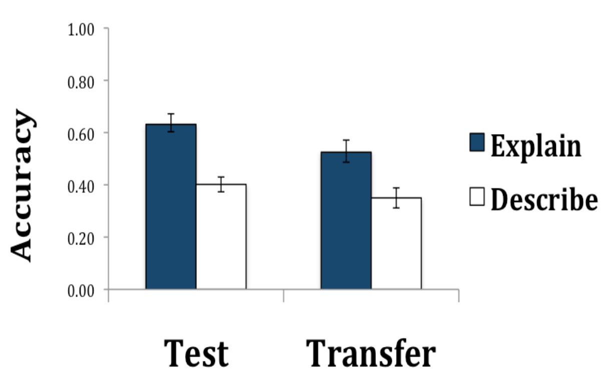


Figure 4. Memory for consistent and anomalous items in Experiment 1.

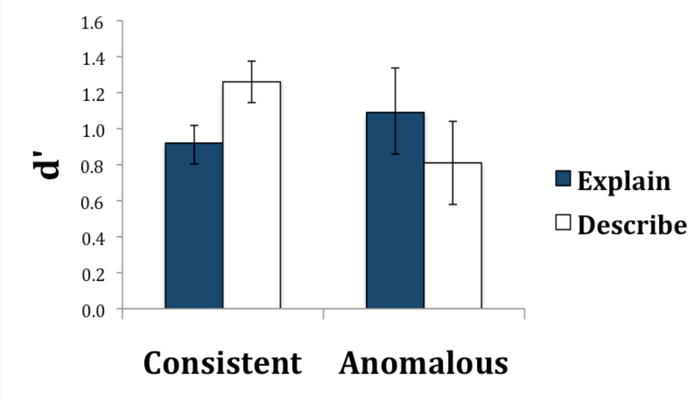


Figure 5. Memory for consistent items as a function of basis for categorization in Experiment 1.

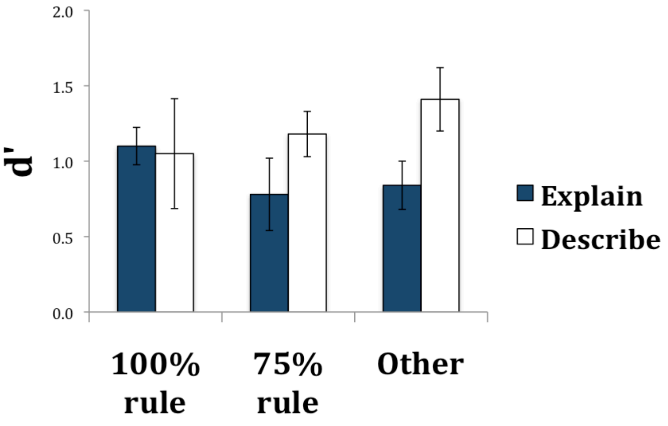


Figure 6. Coding of feature references in explanations and descriptions in Experiment 1.

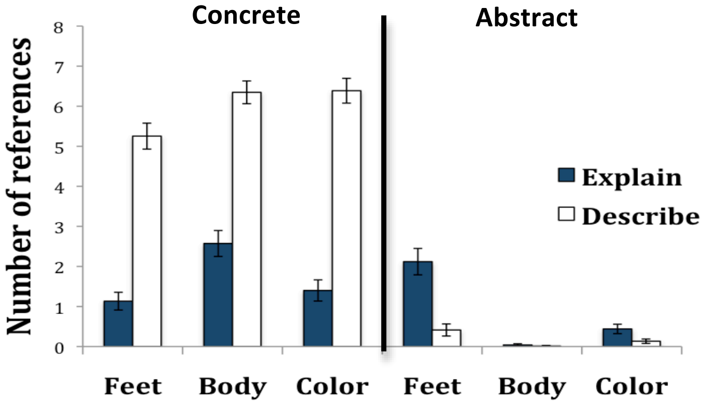


Figure 7. Categorization accuracy in Experiment 2.

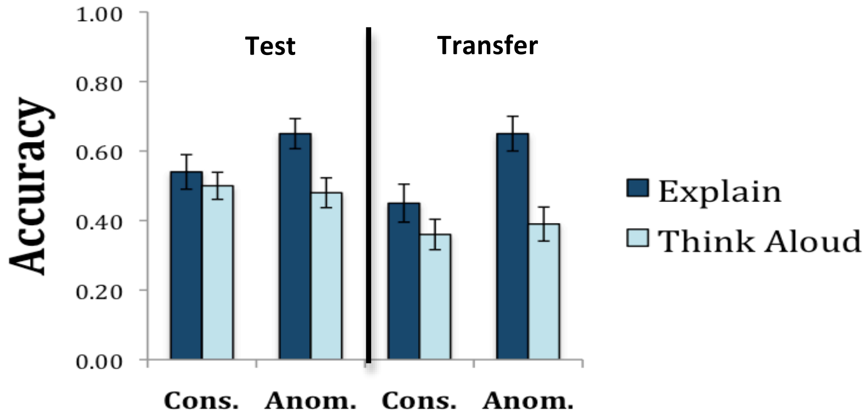


Figure 8. Categorization accuracy in Experiment 3.

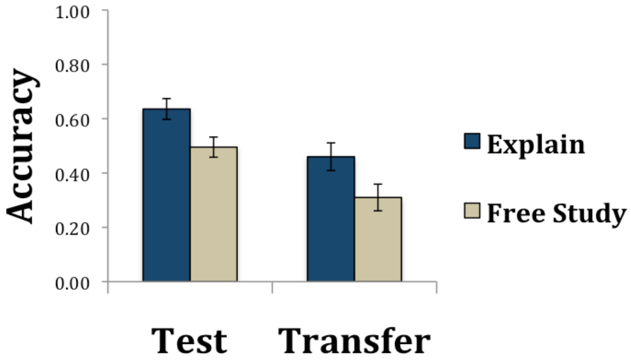


Table 1. Overview of experiments: key differences.

	Introduction	**Study Items**	**Control condition**
Exp. 1	Informed about two categories.	8 x 50 s	Describe
Exp. 2	Informed about two categories and Memory & Categorization tests; 3 blocks exposure	2 x 90 s (Consistent vs Anomalous)	Think Aloud
Exp. 3	Informed about two categories and Memory & Categorization tests	Sheet x 120 s	Free study

Table 2. Number of participants in Experiment 1 coded as providing each basis for categorization on the basis of explicit reports.

	100% rule-FOOT	75% rule-BODY	Item Similarity	Other
Explain	26	14	0	35
Describe	6	40	0	29

Table 3. Number of participants in Experiment 2 coded as providing each basis for categorization on the basis of explicit reports.

	100% Rule – foot	75% Rule – body	Other
Explain – Consistent	22	19	19
Explain - Anomaly	26	10	24
Think Aloud - Consistent	8	17	35
Think Aloud - Anomaly	8	22	30

Table 4. Number of participants in Experiment 2 corresponding to each basis for categorization as inferred from transfer accuracy.

	100% Rule – foot	NOT 100% rule
Explain – Consistent	21	39
Explain - Anomaly	30	30
Think Aloud - Consistent	8	52
Think Aloud - Anomaly	13	47

Table 5. Number of participants in Experiment 3 coded as providing each basis for categorization on the basis of explicit reports.

	100% rule-foot	75% rule-body	Other
Explain	17	9	34
Free study	8	19	33

Table 6. Number of participants reporting attempts to explain category membership in Experiment 3.

Engaged in explanation?	Explain	Free study
Yes	29	23
Not sure	20	19
No	11	18

Table 7. Number of participants in Experiment 3 coded as providing each basis for categorization on the basis of explicit reports, further subdivided by self-reported explaining.

	Engaged in explanation?	100% rule-foot	75% rule-body	Other
Explain	Exp-Yes	14	2	13
	Exp-Other	3	7	21
Free study	Exp-Yes	5	11	7
	Exp-Other	3	8	26

3. Explanation and Prior Knowledge Interact to Guide Learning

3.1. 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.

3.2. 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.[6]

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 the four experiments that follow.

3.2.1. 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"

[6] 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.

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.

3.2.2. 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.

3.2.3. 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?

3.3. 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.

3.4. 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.

3.4.1. Methods

Participants.

Four-hundred-and-seven UC Berkeley undergraduate students participated for course credit or monetary reimbursement.[7]

Materials.

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.[8] 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

[7] 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.

[8] 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).

colors, but these features were not diagnostic of category membership as they occurred equally often in each category.

Categorization probes. To assess which features participants used in generalizing category membership from the study observations to novel robots, participants classified fifteen 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.[9] 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.

Procedure.

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.

Test phase.

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.

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

[9] 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.

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."

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.

3.4.2. Results

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, $c^2(1, N = 407) = 11.65$, $p < 0.01$, with higher discovery rates for participants who explained, as well as an interaction between label type and discovery, $c^2(1, N = 407) = 11.61$, $p < 0.01$, 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, $c^2(1, N = 407) = 3.98$, $p < 0.05$: 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, $c^2(1, N = 207) = 1.43$, $p = 0.15$, nor by providing informative labels in free study conditions, $c^2(1, N = 200) = 0.55$, $p = 0.52$.

A parallel analysis on discovery of the *label-irrelevant* pattern also revealed a three-way interaction with task and label type, $c^2(1, N = 407) = 5.48$, $p < 0.05$, superseding interactions between task and discovery, $c^2(1, N = 407) = 17.39$, $p < 0.001$, and label type and discovery, $c^2(1, N = 407) = 11.47$, $p < 0.001$. 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, $c^2(1, N = 207) = 17.98, p < 0.01$.

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.

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, $c^2(2, N = 407) = 80.97, p < 0.001$, as well as between number and label type, $c^2(2, N = 407) = 8.53, p < 0.05$. 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, $c^2(1, N = 407) = 71.52, p < 0.001$, but more likely to discover exactly one, $c^2(1, N = 407) = 74.86, p < 0.001$, which was also more likely among participants receiving blank labels, $c^2(1, N = 407) = 7.64, p < 0.01$. There was no effect of explanation on discovering two patterns, although there was a marginal effect of label type, $c^2(1, N = 407) = 3.50, p = 0.062$, 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.

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, $c^2(1, N = 42) = 7.10, p < 0.01$. 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, $c^2(1, N = 69) = 6.73, p < 0.01$. 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, $c^2(1, N = 200) = 11.88, p < 0.01$), 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.

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, $c^2(3, N = 407) = 92.02, p < 0.0001$, as well as between label type and basis, $c^2(3, N = 407) = 17.34, p < 0.001$, with a marginal three-way interaction, $c^2(3, N = 407) = 6.89, p = 0.07$. 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, $c^2(1, N = 407) = 7.27, p < 0.01$, superseding the effects of explanation, $c^2(1, N = 407) = 4.98, p < 0.05$, and prior knowledge, $c^2(1, N = 407) = 4.21, p < 0.05$. Task and label type also interacted with use of the label-irrelevant pattern, $c^2(1, N = 407) = 7.18, p < 0.05$, with significant effects of task, $c^2(1, N = 407) = 17.39, p < 0.001$, and label type, $c^2(1, N = 407) = 11.47, p < 0.01$. One additional finding of note was that participants in the free study conditions were significantly more likely to generalize category membership by item similarity, $c^2(1, N = 407) = 3.90, p < 0.05$. 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.

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

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 (*p*s < .01), but the three-way interactions did not.[10] 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.

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 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.

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'

[10] 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 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.

3.5. 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).

3.5.1. Methods.

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.

Materials & Procedure.

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).

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.

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.

3.5.2. Results & Discussion.

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, $c^2(1, N = 554) = 5.31$, $p < 0.05$, which superseded the effects of task, $c^2(1, N = 554) = 7.00$, $p < 0.01$, and label type, $c^2(1, N = 554) = 8.64$, $p < 0.01$. 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, $c^2(1, N = 554) = 5.02$, $p < 0.05$. In addition, discovery of the body shape pattern was higher in the write thoughts than explain conditions, $c^2(1, N = 554) = 5.97$, $p < 0.05$. 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.

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, $c^2(1, N = 554) = 16.22$, $p < 0.01$, and label type, $c^2(1, N = 554) = 13.44$, $p < 0.01$, 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, $c^2(1, N = 554) = 3.88$, $p < 0.05$, while those in the explain conditions were more likely to discovery exactly one pattern, $c^2(1, N = 554) = 9.30$, $p < 0.01$. However, engaging in explanation and writing thoughts did not differ significantly in promoting discovery of multiple patterns, $c^2(1, N = 554) = 2.76$, $p = 0.10$. There were no additional significant effects.

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

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%), $c^2(1, N = 554) = 3.55$, $p = 0.06$.

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%; $c^2(1, N = 447) = 1.14$, $p = 0.\ 29$). However, for participants in both study conditions, informative labels were mentioned more frequently than blank labels (informative: 67%; blank 55%; $c^2(1, N = 447) = 6.78$, $p < 0.01$). 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.

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.

3.6. 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.

3.6.1. Methods

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.

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.

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.

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.

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.

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.

3.6.2. Results & Discussion

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).

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.[11] 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.

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.[12]

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 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

[11] 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.

[12] 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.

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.

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.

3.7. 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.[13] 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).

3.7.1. Methods

Participants.

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

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 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*.

[13] 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.

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.

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).

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.

3.7.2. 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.

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.

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.

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, $t(356) = 2.45$, $p < 0.05$, $d = 0.26$.

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 *p*s > 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, $c^2(1, N = 682) = 4.82$, $p < 0.05$. 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%, $c^2(1, N = 339) = 5.80$, $p < 0.05$), while no such difference existed for representative labels (explain: 41%; write thoughts: 37%; $c^2(1, N = 343) = 0.47$, $p = 0.51$).

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.

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.

3.8. 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).

3.8.1. 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.

3.8.2. 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).

3.8.3. 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.

3.8.4. 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.

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

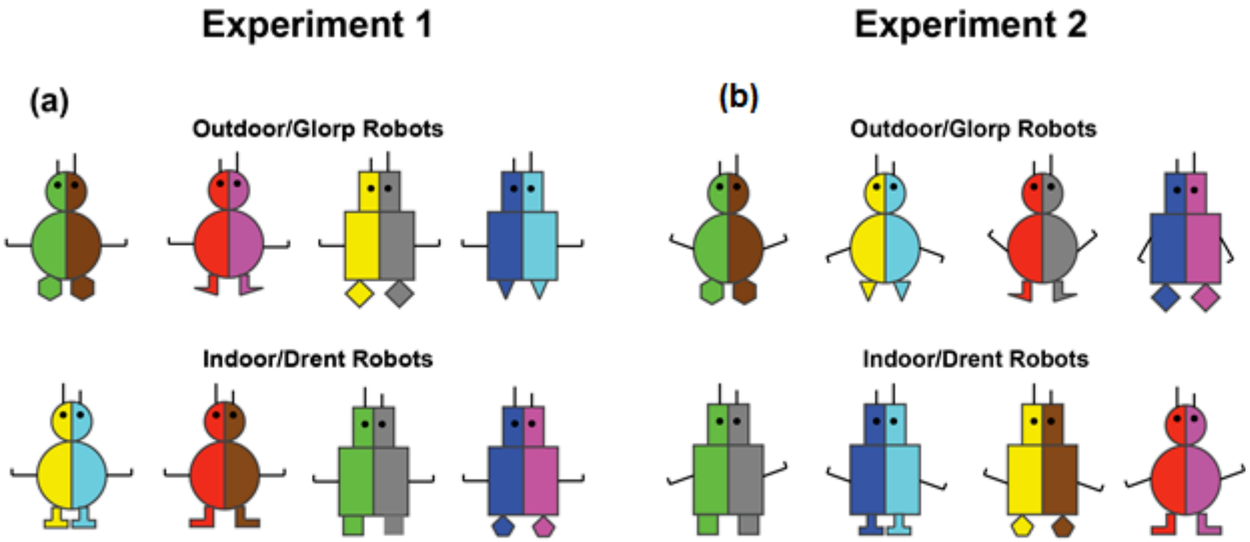


Figure 2. Results from Experiment 1 (a to c) and Experiment 2 (d to f). Error bars represent one standard error of the mean in each direction. **Pattern Discovery** (a & d): Proportion of participants who discovered the label-relevant and label-irrelevant patterns, and for Experiment 2, the additional 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.

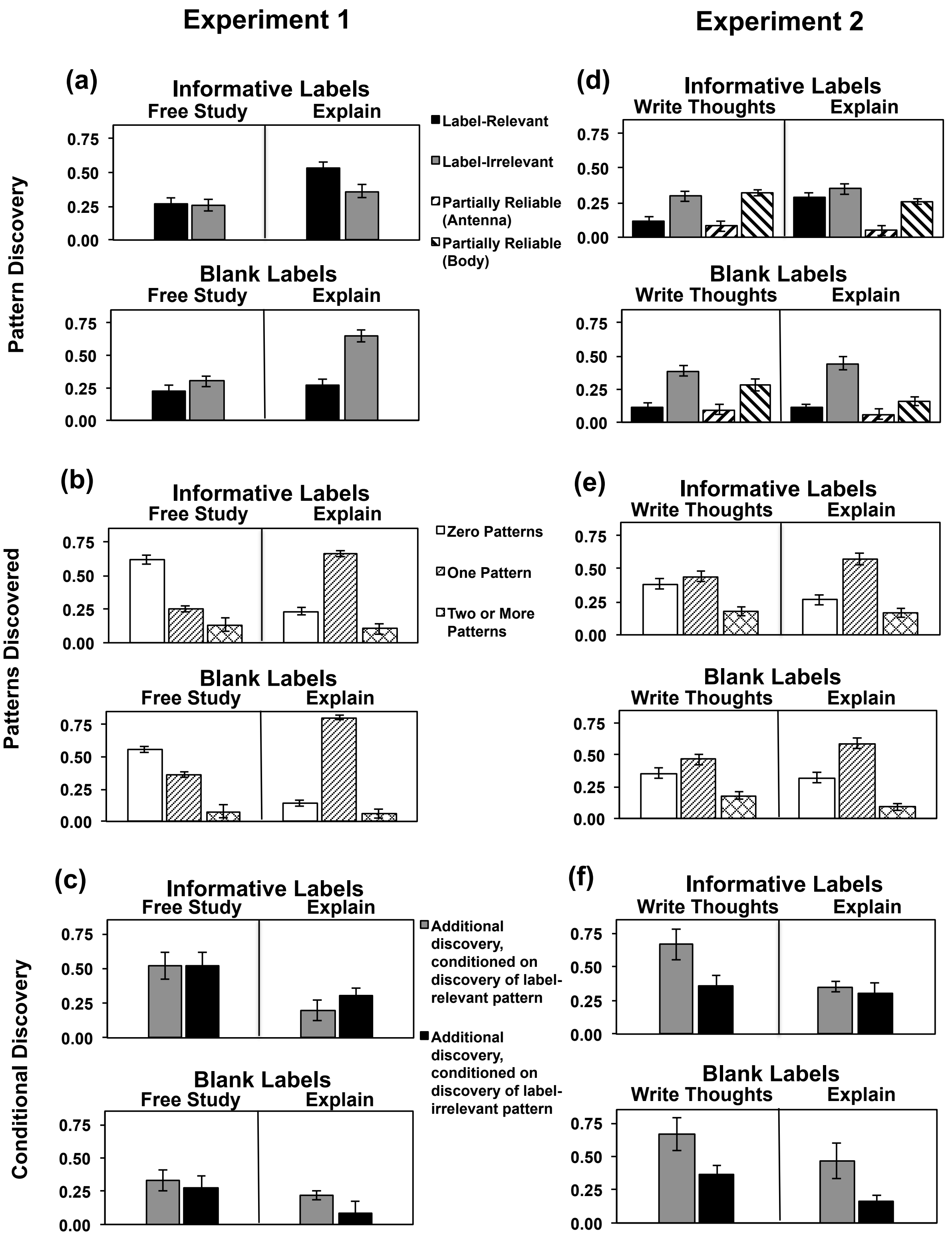

Experiment 1
Experiment 2
Pattern Discovery
Patterns Discovered
Conditional Discovery
(a)
Informative Labels
Free Study
Explain
Blank Labels
Free Study
Explain
0.00
0.25
0.50
0.75
■ Label-Relevant
■ Label-Irrelevant
▨ Partially Reliable (Antenna)
▧ Partially Reliable (Body)
(b)
Informative Labels
Free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

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.

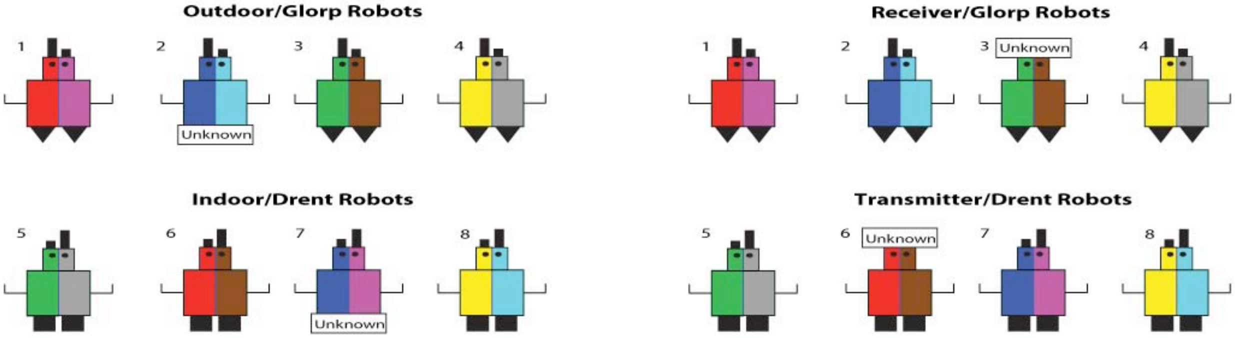


Figure 4. Extent to which the label-relevant pattern was used as a basis for generalizing category membership, as a function of task and label type, in Experiments 3 and 4. (a) Proportion of classifications consistent with label-relevant pattern in Experiment 3. (b) Average classification rating in Experiment 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.

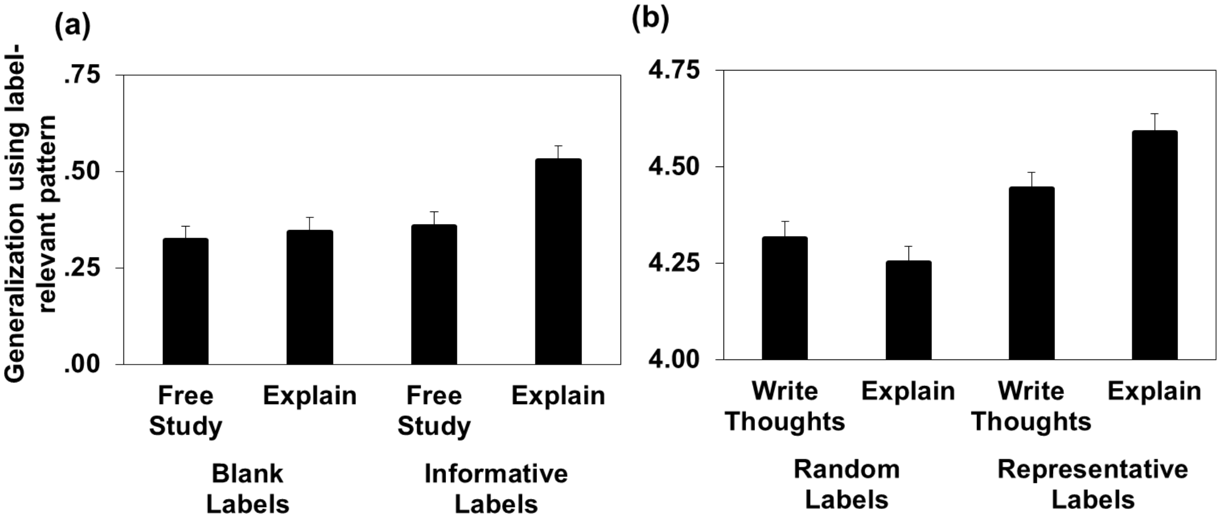


Table 1. Proportion of participants classified as using each basis for categorization in Experiment 1.

	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)

4. The hazards of explanation: overgeneralization in the face of exceptions

4.1. Abstract

Seeking explanations is central to science, education, and everyday thinking, and prompting learners to explain is often beneficial. Nonetheless, in two category learning experiments across artifact and social domains, we demonstrate that the very properties of explanation that support learning can *impair* learning by fostering overgeneralizations. We find that explaining encourages learners to seek broad patterns, hindering learning when patterns involve exceptions. By revealing how effects of explanation depend on the structure of what is being learned, these experiments simultaneously demonstrate the hazards of explaining and provide evidence for why explaining is so often beneficial. For better or for worse, explaining recruits the remarkable human capacity to seek underlying patterns that go beyond individual observations.

4.2. The hazards of explanation

People often have the impression of understanding something better after explaining it to someone else, whether it's why a person behaved the way she did or the solution to a math problem. In fact, prompting students to explain while they study can improve learning (e.g., Fonseca & Chi, 2010), and prompting children to explain can improve generalization to new problems (e.g., Amsterlaw & Wellman, 2006; Siegler, 2002). What is it about *explaining* that so effectively fosters learning? And if explaining is so beneficial, why don't people spontaneously explain more often?

Educational, developmental, and cognitive psychologists have proposed many answers to the first question. For example, explaining could increase attention, motivation, or engagement (e.g., Chi, 2009; Siegler, 2002), help learners identify and fill gaps in knowledge (e.g., Chi, 2000), or improve learning by facilitating the integration of novel information with prior beliefs (e.g., Chi et al, 1994; Lombrozo, 2006; Wellman & Liu, 2007). Given that these processes are demanding, people could fail to explain spontaneously – even when doing so would be beneficial – to avoid what they see as inessential costs in cognitive processing (e.g., see Fiske & Taylor, 1984; Gigerenzer, 2004).

Despite its appeal, a view of explanation as globally beneficial but inconsistently applied is at best incomplete. For one thing, such a view cannot account for previously-documented *hazards* of explanation. Needham and Begg (1991) found that participants prompted to explain solutions to riddle-like problems outperformed those prompted to memorize the solutions when it came to analogical transfer, but performed more poorly on memory for studied examples. Kuhn and Katz (2009) found that students prompted to explain causal claims were more likely to subsequently justify claims by appeal to potential mechanisms, ignoring relevant evidence from covariation. Finally, Berthold et al. (2011) found that a conceptually-oriented explanation prompt improved conceptual learning, but impaired procedural learning. These findings suggest that explanation does not have universally beneficial effects, and additionally highlight the need to specify more precisely *what* explanation directs attention or processing towards and precisely *why* it does so.

We propose that the very properties of explanation that make it a powerful mechanism for learning under some conditions lead to systematic errors under others. Specifically, we propose that explaining privileges broad generalizations over learning about individual instances, making learners susceptible to erroneous *over*generalizations. This idea is motivated from unification theories of explanation in philosophy, which propose that a good explanation is one that subsumes the fact or observation being explained as an instance of a broad pattern or generalization (e.g., Friedman, 1974; Kitcher, 1981, 1989). Explaining should therefore drive people to search for patterns that support satisfying explanations. In line with this prediction, we have found that explaining makes people more likely to discover patterns that account for a broad range of observations, even when alternative patterns are more salient (Williams & Lombrozo, 2010).

Here we test a novel and counterintuitive prediction of this account: that explaining can *impair* learning by leading to erroneous overgeneralizations when patterns involve exceptions. Such a finding would not only challenge the idea that explanation merely boosts processing or attention, but also shed light on what explanation directs effort and attention towards, and ultimately why explaining is so often beneficial for learning.

4.3. Experiment 1

Experiment 1 investigated effects of explanation in learning novel categories. Participants were prompted to explain or "think aloud" while studying ten labeled exemplars, where the underlying category structure involved a "reliable" pattern without exceptions or a "misleading" pattern with two exceptions. The exemplars additionally involved unique features that supported perfect classification. If explaining encourages learners to discover and privilege broad patterns in the face of exceptions, then participants who explain should fare more poorly than those who think aloud when the pattern is misleading.

The inclusion of a "think aloud" condition with an identical learning task was crucial to discriminate effects of explanation from previously-documented effects of verbalization or intentional learning, which can impair some kinds of category learning, memory, and implicit grammar acquisition (e.g., Ashby & Maddox, 2004; Love, 2002; Mathews et al., 1989; Toth, Reingold, & Jacoby, 1994). Experiment 1 can therefore isolate distinctive contributions of *explaining* to the specific impairment we predict: ignoring individual instances in favor of generalizations.

4.3.1. Methods

Participants.

Participants were 240 undergraduates and members of the UC Berkeley community who participated in exchange for pay or course credit.

Materials and procedure.

Learning phase. Participants learned to classify 10 novel objects (vehicles) into two categories through repeated classification, feedback, and study (see Figure 1). Participants received instructions and then completed a study trial consisting of (1) classifying an unlabeled object as "Dax" or "Kez" based on its description (e.g., blue, lightly insulated, etc.), (2) receiving feedback on category membership, and (3) studying the labeled object. During study, participants in the *explain* condition were prompted by a sentence on the screen to explain why the item might be a Dax [Kez], while those in the *think aloud* condition were prompted to say out loud whatever they were thinking.[14] This process

[14] Voice recorders were set up for both groups of participants in both Experiments 1 and 2, but unfortunately the data from all but a handful of these was lost due to a computer error.

was repeated for all ten items to form a single "block," and participants repeatedly classified blocks until achieving perfect classification or reaching the maximum of 15 blocks.

Each object description included one unique color feature, one feature relevant to a pattern (suitability for hot versus cold climates), and three features that were not diagnostic of category membership (see Table 1; materials were adapted from Kaplan & Murphy, 2000, by the addition of the unique color features). Participants could thus classify by remembering the ten *unique features* (e.g., "the red one is a Dax") or by finding a pattern in the ten *pattern-related features* (e.g., "a Dax is a vehicle for warm climates"). We manipulated whether this pattern was *reliable* (no exceptions, as in Kaplan & Murphy, 2000) or *misleading* (two exceptions in ten, created by randomly switching two pattern-related features in each block). We chose *two* exceptions as the most minimal manipulation of pattern reliability.

Post-learning measures. To investigate effects of explanation and pattern reliability on learning about category membership via unique versus pattern-related features, we included several post-learning measures (see Figure 1).

Individual feature classification. Each unique and pattern-related feature was presented individually in a random order. Participants were asked which category an item with that feature would belong to.

Conflict classification. Participants classified 10 "conflict items" which paired a *pattern-related* feature and a *unique* feature that corresponded to opposite classifications.

Reported differences. Participants reported differences across categories, typing their responses.

4.3.2. Results

Learning time.

A 2 (*study condition*: *explain, think aloud*) x 2 (*pattern reliability*: *reliable, misleading*) ANOVA on the mean number of blocks to reach the learning criterion revealed that participants learned more quickly when the pattern was reliable than misleading, $F(1,236) = 44.5$, $p < 0.001$, $\eta_p = 0.26$ (see Figure 2a). However, this effect was qualified by an interaction between *study condition* and *pattern reliability*, $F(1,236) = 6.3$, $p < 0.05$, $\eta_p = 0.03$.[15] We therefore evaluated effects of explanation separately for each pattern.

When the pattern was reliable, there was a trend for the *explain* group to learn faster than the *think aloud* group, $t(118) = 1.43$, $p = 0.16$, $d = 0.26$. When the pattern was misleading, however, participants in the *explain* condition were significantly *slower* to reach the learning criterion, $t(118) = 2.1$, $p < 0.05$, $d = 0.38$.[16] In fact, 52% of participants from the *misleading*/*explain* condition never

[15] To address concerns about non-normality, this analysis was repeated with a non-parametric test. We sorted the number of blocks to learning into five bins of three block increments and performed an ordinal regression with study condition and pattern reliability as factors. This analysis also revealed a significant interaction.

[16] To address concerns about non-normality, all *t*-tests reported in this experiment were checked with non-parametric Mann-Whitney U tests, which supported the same conclusions.

achieved perfect classification – significantly more than the 25% who failed to do so in the *misleading*/*think aloud* condition, $\chi^2(1) = 5.4$, $p < 0.05$.

Individual feature classification.

Performance was analyzed with a mixed ANOVA with *study condition* (2) and *pattern reliability* (2) as between-subjects factors, and *feature type* (2: *unique*, *pattern-related*) as a within-subjects factor (see Figure 2b). This analysis revealed an interaction between *study condition* and *feature type*, $F(1, 236) = 12.79$, $p < 0.001$, $\eta_p = 0.05$: explaining resulted in *fewer* errors on pattern-related features, $t(238) = 3.10$, $p < 0.01$, $d = 0.40$, but *more* errors on unique features, $t(238) = -2.37$, $p < 0.05$, $d = -0.31$. This interaction was independently significant when the pattern was reliable and supported learning as well as when the pattern was misleading and hindered learning, *p*s < .05.

There were also main effects of *pattern reliability*, $F(1, 236) = 7.10$, p < 0.01, $\eta_p = 0.03$, and *feature type*, $F(1, 236) = 24.00$, $p < 0.001$, $\eta_p = 0.09$, which were superseded by an interaction, $F(1, 236) = 15.04$, $p < 0.001$, $\eta_p = 0.06$. Participants in the *reliable pattern* conditions classified pattern-related features more accurately than those in the *misleading pattern* conditions, $t(238) = 4.44$, $p < 0.001$, $d = 0.58$, with a slight trend in the opposite direction for unique features, $t(238) = -1.43$, $p = 0.15$, $d = -0.19$.

Conflict classification.

A 2 (*study condition*) by 2 (*pattern type*) ANOVA revealed that a greater proportion of conflict items were classified in line with pattern-related features (as opposed to unique features) when the pattern was reliable rather than misleading, $F(1, 236) = 26.62$, $p < 0.001$, $\eta_p = 0.10$, and when participants explained rather than thought aloud, $F(1, 236) = 13.43$, $p < 0.001$, $\eta_p = 0.05$. The latter effect was independently significant for each pattern type, *p*s < 0.05 (see Figure 1c).

Reported differences.

Participants' typed reports about category differences were independently coded (with 84% agreement) for mention of the hot/cold pattern and/or the unique color features (see Figure 2d). Mention of the pattern was more frequent in the *explain* than *think aloud* conditions, whether the pattern was reliable, $\chi^2(1) = 4.04$, $p < 0.05$, or misleading, $\chi^2(1) = 9.79$, $p < 0.05$. However, mention of color differences was *less* frequent in the explain than think aloud conditions (pattern reliable: $\chi^2(1) = 4.82$, $p < 0.05$, misleading: $\chi^2(1) = 4.48$, $p < 0.05$).

4.3.3. Discussion

Experiment 1 confirmed our prediction that explaining can impair learning when patterns are misleading, with participants who were prompted to explain requiring more study time to reach the learning criterion than those who thought aloud. The findings additionally shed light on the basis for this impairment: Relative to thinking aloud, explaining improved learning of features that supported patterns but impaired learning of features unique to particular instances.

4.4. Experiment 2

Experiment 2 investigated effects of explaining on categorizing people's behavior, an important extension to Experiment 1 and past work on explanation in categorizing novel objects (Williams & Lombrozo, 2010). Because behavior is arguably explained in terms of unique features (Malle, 2011; Master, Markman, & Dweck, 2012) more readily than is the category membership of objects, finding an impairment here would bolster the evidence that explaining drives people towards broad patterns at the expense of idiosyncratic particulars.

Experiment 2 also went beyond Experiment 1 in comparing effects of generating explanations during study to a control condition in which participants *anticipated* having to later generate explanations. Expectations about the viability of explanations were thus matched while manipulating the degree to which participants explained *during study*. In addition, learning was evaluated by examining errors during a learning session of fixed time, avoiding the variability generated by learning to criterion.

4.4.1. Methods

Participants.

Participants were 164 undergraduates and members of the UC Berkeley community who participated in exchange for pay or course credit.

Materials and procedure.

Learning phase. Participants were instructed to learn whether each of ten hypothetical individuals rarely or frequently donated to charities (see Figure 3). *Explain* participants were told to explain each individual's behavior during study, while *control* participants were told that they would explain each individual's behavior at a later point. Participants then read a description of each individual and had 10 seconds to judge whether the individual rarely or frequently donated before receiving the correct answer and studying the individual for another 10 seconds. This process was repeated five times for each of the 10 individuals, with learning assessed by the proportion of errors over these trials.

Each description included the person's (ostensible) picture, name, and age, as well as a personality descriptor (e.g., friendly) and two additional features, college major and geographic location, which were not correlated with behavior (see Table 2). The unique picture and name supported perfect predictions concerning behavior (e.g., "Laura rarely donates"), while the age and personality features conformed to patterns that correlated with behavior (e.g., "younger people rarely [frequently] donate," "people with extraverted traits rarely [frequently] donate").[17] Using two patterns (age and personality) created a more complex learning context than Experiment 1, which involved a single pattern. In the *reliable patterns* conditions, the patterns involving age and personality correlated perfectly with behavior. In the *misleading patterns* conditions, each pattern involved two exceptions.

4.4.2. Results

Learning error was computed as the proportion of errors made during the learning phase, and was analyzed with an ANOVA that included *study condition* (2: explain, control) and *pattern reliability* (2: reliable, misleading) as between-subjects factors. This analysis revealed more errors in the *misleading patterns* conditions than the *reliable patterns* conditions, $F(1,160) = 32.41$, $p < 0.001$, $\eta_p = 0.25$, as well as the critical interaction between *study condition* and *pattern reliability*, $F(1,160) = 4.63$, $p < 0.05$, $\eta_p = 0.03$ (see Figure 4a). Explaining had no significant effect on errors when the patterns were reliable, $t(75) = 0.35$, $p = 0.73$, $d = 0.08$, but *increased* errors when the patterns were misleading, $t(85) = 2.50$, $p < 0.05$, $d = 0.54$.

[17] The direction of the association between age, personality, and behavior was counterbalanced across conditions, but this manipulation had no effect and is not discussed further.

To examine whether errors in the explain condition resulted from overgeneralizing patterns to exceptions, we performed an additional ANOVA on errors in the misleading patterns condition that considered two additional within-subjects factors: *item type,* where items were identified as *pattern-consistent* if the individual's behavior conformed to the age and personality patterns, and as *exceptions* if not (see Figure 4b), and *learning block* (see Figure 4c). The inclusion of item type and learning block did not alter the key finding that explaining increased errors relative to control, $F(1, 83) = 5.12, p < .05, \eta_p = 0.06$. However, there was additionally a main effect of learning block, $F(4,80) = 37.42, p < .001$, with errors decreasing over time, a main effect of item type, $F(1, 83) = 16.45, p < .001, \eta_p = 0.32$, with more errors for exceptions, and a (marginal) interaction between study condition and item type, $F(1, 83) = 3.82, p = 0.054, \eta_p = 0.04$: Participants in the *explain* condition made significantly more errors than those in the *control* condition for exception items, $F(1, 83) = 5.12, p < 0.05, \eta_p = 0.06$ (this difference was significant by block two, $p < 0.05$, and was still significant in block five, $p < 0.05$), with a similar but non-significant trend for pattern-consistent items, $F(1, 83) = 1.44, p = 0.23, \eta_p = 0.02$. We speculate that participants prompted to explain did not outperform control participants on pattern-consistent items because they may have switched between patterns (age versus personality) in the face of exceptions instead of abandoning patterns altogether.

4.4.3. Discussion

Experiment 2 replicated the key prediction that explaining can impair learning when patterns are misleading, extending the finding from Experiment 1 to a novel domain, a new measure of learning, and a different control condition. The analysis of error types suggests that explaining impaired learning through rapidly formed overgeneralizations that persisted in the face of repeated counterevidence.

4.5. General Discussion

Our findings reveal the double-edged nature of explanation. While explaining can be beneficial, it can also make learners vulnerable to overgeneralizations when categorizing artifacts (Experiment 1) or behaviors (Experiment 2), and for measures of both learning speed (Experiment 1) and learning accuracy (Experiment 2). In Experiment 1, explainers more accurately learned pattern-related features and less accurately learned unique features than those who thought aloud. In Experiment 2, explainers were especially inaccurate when it came to categorizing exceptions, and this effect emerged early in learning and persisted throughout training. These findings suggest that explainers focused on features that supported patterns at the expense of idiosyncratic information about individual items, and that they perseverated in seeking or applying broad patterns despite evidence against their generality.

Just as visual illusions shed light on the mechanisms by which visual perception is so often accurate, our findings shed light on why explanation is so often beneficial. In particular, our findings suggest that explaining "why?" recruits evaluative criteria for what constitutes a good explanation, directing learners to seek broad patterns that can accommodate what is being explained (see also Williams & Lombrozo, 2010; Lombrozo, 2012). Explaining can therefore support the remarkable human capacity to discover patterns and construct generalizations from sparse observations, but this capacity has associated risks: disregarded exceptions and overgeneralization.

Our findings also rule out the idea that explaining simply increases attention or processing to yield global improvements, and can help make sense of the seemingly disjointed set of negative effects of explanation reviewed in the introduction. For example, relative to instructions to remember examples, explanation promotes analogical transfer at the expense of memory (Needham & Begg, 1991), and this effect makes sense if explaining highlights broad patterns over individual examples. Similarly, explaining could encourage learners to invoke causal mechanisms over particular observations in justifying causal claims (Kuhn & Katz, 2009) because doing so relates what is being explained to broader regularities. Finally, explaining could encourage learners to draw generalizations over conceptual rather than procedural aspects of a domain when the content of what they are explaining is conceptual (Berthold et al, 2011).

Our finding that explanation can promote erroneous overgeneralizations goes beyond these previous results in isolating a prediction of our account, but additionally suggests that effects of explanation are not a simple consequence of directing limited attention or processing to the target of a particular study prompt (e.g., memory, causal mechanisms, or conceptual knowledge) at the expense of alternatives (e.g., relational structure, covariation, or procedural knowledge). Nothing about the content of our explanation prompts highlighted broad patterns over properties of individuals. If anything, a prompt to explain the category membership or behavior of a specific instance could have directed attention or processing to the particulars of that instance. Instead, the results support our proposal that explaining "why?" is inherently linked to patterns, with the content of the question potentially affecting the nature of the patterns considered (e.g., whether they involve conceptual or procedural regularities; see also Williams & Lombrozo, 2013).

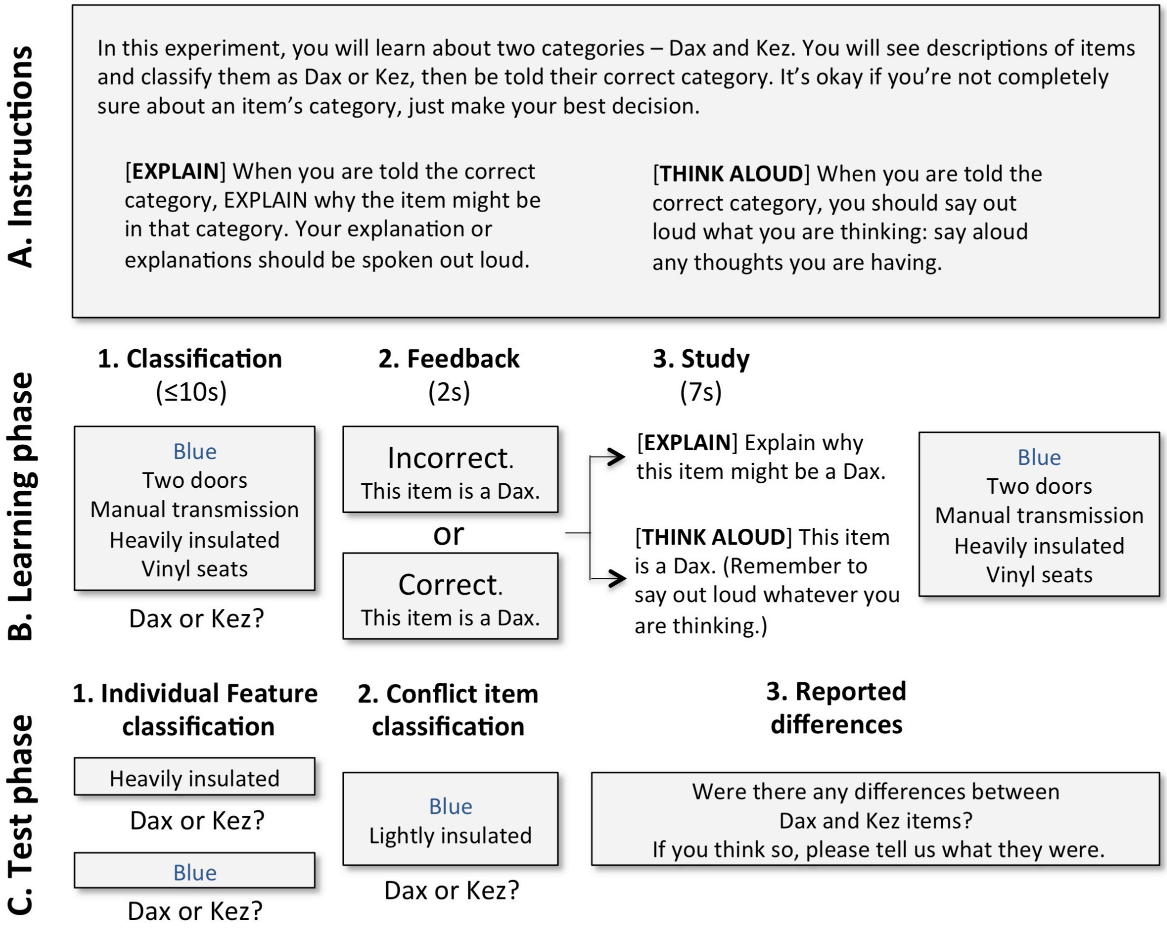
Our account also predicts conditions under which explanation should have minimal effects. For example, explanation prompts could have no effect when learners can successfully identify and apply broad generalizations without explaining (e.g., because they receive rich and effective instruction) or when learners lack the requisite knowledge to generate reasonable hypotheses about underlying patterns (for related discussion, see Matthews & Rittle-Johnson, 2009; Rittle-Johnson, 2006).

In their search for patterns, participants who explained could have recruited a host of well-documented processes, including verbal reasoning (e.g., Meissner & Memon, 2002; Schooler, 2002) and analogical comparison (Gentner, 2010), or triggered a more explicit and deliberative (Mathews et al, 1989), analytic and rule-based (Shanks & John, 1994; Ashby & Maddox, 2005), or intentional (Dienes et al, 1991; Reber, 1989) mode of learning. Given our closely matched study conditions, it's likely that these processes and strategies were also triggered in participants in control conditions, if to a lesser degree. Our account is not in conflict with these views concerning cognitive mechanisms or architectures, but instead suggests that if explanation did recruit these mechanisms or systems, it was in the service of finding broad patterns, and it is this feature of *explanation* that explains our results.

While we specifically designed conditions conducive to an explanation impairment using feature lists in a laboratory context, a range of real-world situations involve similarly sparse observations and unreliable patterns. For example, explaining a single (potentially unrepresentative) observation can generate the kinds of beliefs that underlie stereotypes (Risen, Gilovich, & Dunning, 2007), and trying to explain chance events could reinforce superstitious beliefs or conspiracy theories. Future work can investigate these hazards of explanation, and additionally aim to reconcile them with cases where explaining exceptions is useful in discovering novel regularities, as when anomalies presage scientific theory change (Kuhn, 1962; Chinn & Brewer, 1993) or guide children's causal learning (Legare, Gelman, & Wellman, 2010).

Finally, we should note that ignoring exceptions may sometimes *help* learning. For example, when there is substantial variability in observations, exceptions could erroneously lead learners away from noisy but reliable patterns. Moreover, there are many settings, such as those in mathematics and science education, where explaining has proven beneficial, for which the benefits of erring on the side of overgeneralization can outweigh minor costs. Providing a unified account of the positive and negative effects of explanation can not only help avoid the hazards of explanation, but also maximize and extend its benefits, whether in everyday, educational, or scientific contexts.

Figure 1. Schematic presentation of the learning task from Experiment 1. Steps 1-3 in the study phase were repeated for each item in each block, with participants repeating blocks until they achieved perfect classification or reached the maximum of 15.



In this experiment, you will learn about two categories – Dax and Kez. You will see descriptions of items and classify them as Dax or Kez, then be told their correct category. It's okay if you're not completely sure about an item's category, just make your best decision.

[**EXPLAIN**] When you are told the correct category, EXPLAIN why the item might be in that category. Your explanation or explanations should be spoken out loud.

[**THINK ALOUD**] When you are told the correct category, you should say out loud what you are thinking: say aloud any thoughts you are having.

[**EXPLAIN**] Explain why this item might be a Dax.

[**THINK ALOUD**] This item is a Dax. (Remember to say out loud whatever you are thinking.)

Were there any differences between Dax and Kez items?
If you think so, please tell us what they were.

Figure 2. (a) Average learning time from Experiment 1 as a function of study condition and pattern type; (b) Average accuracy in classifying individual features after training; (c) Proportion of pattern-consistent classifications for conflict items after training; (d) Explicitly reported differences across categories. Error bars correspond to one standard error of the mean in each direction.

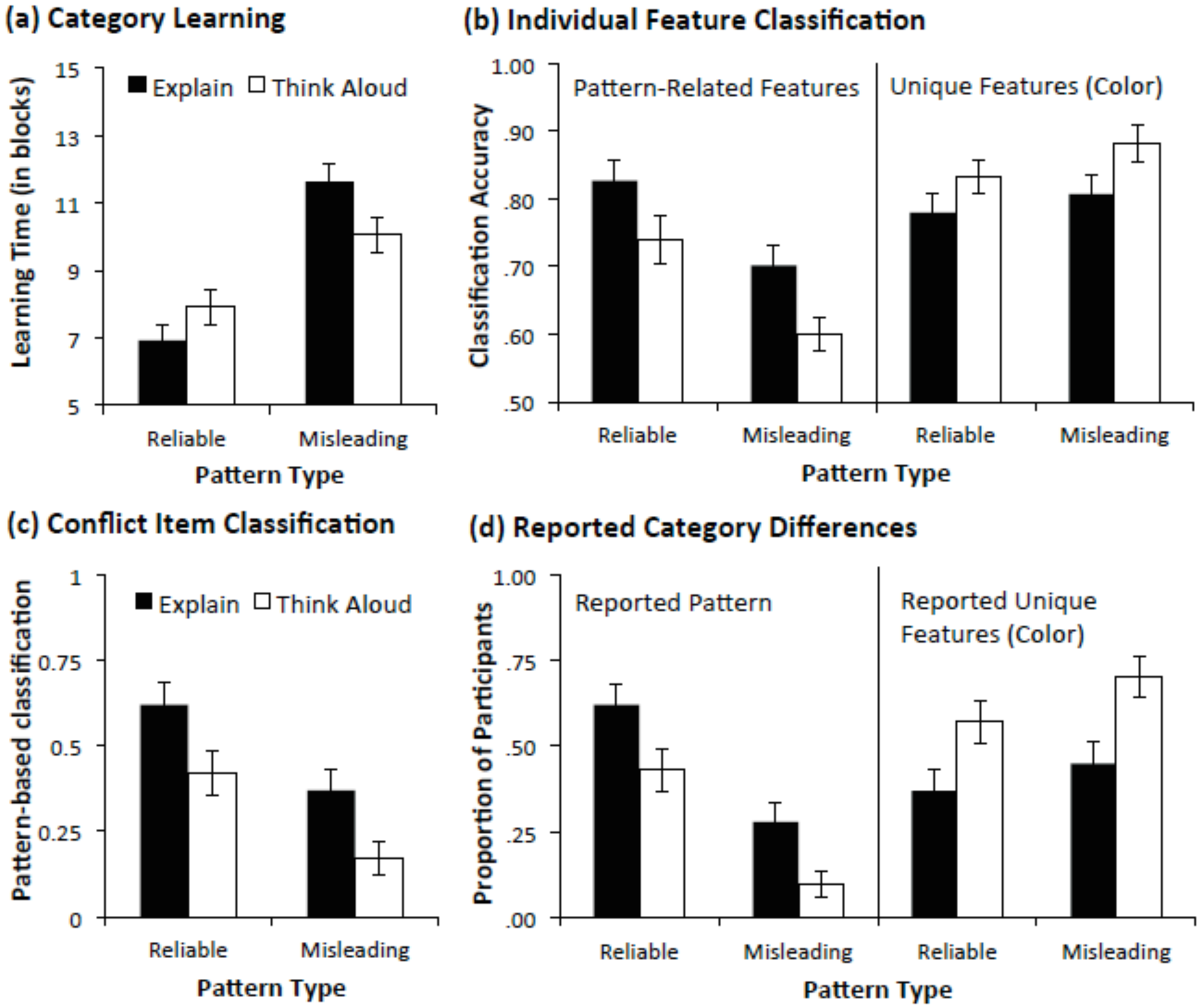
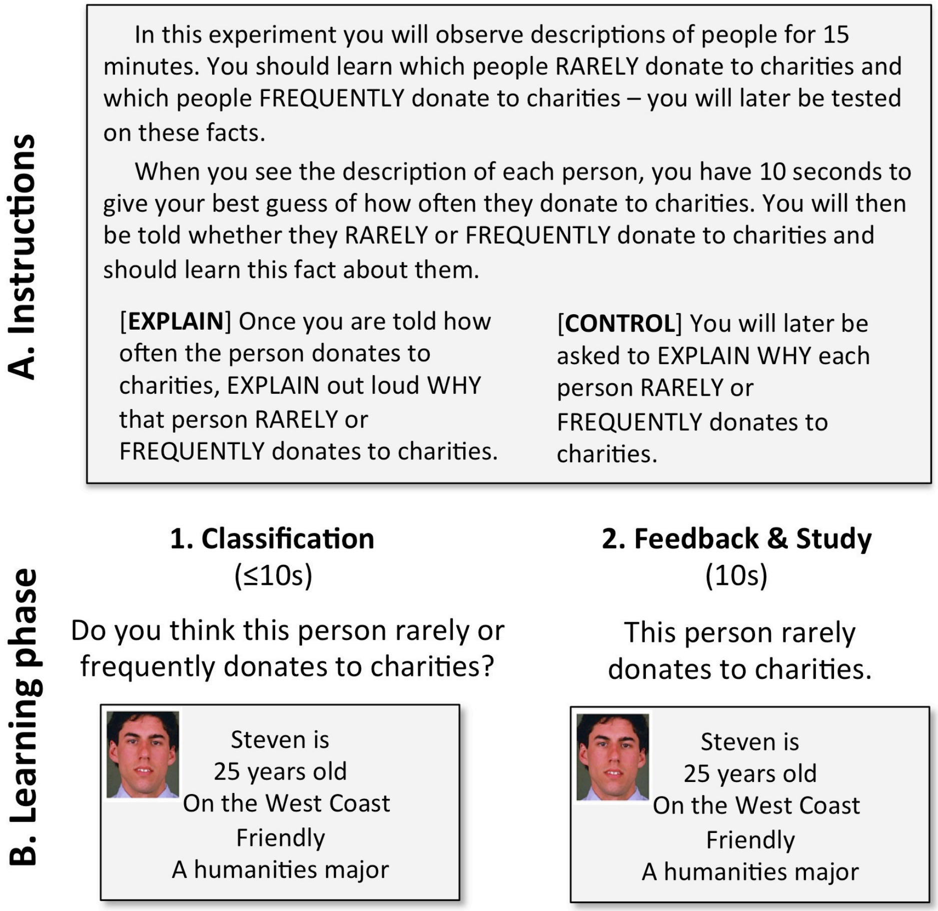


Figure 3. Schematic presentation of the learning task from Experiment 2. Steps 2-3 were repeated for each of the ten items in each block, with participants repeating each block five times.



In this experiment you will observe descriptions of people for 15 minutes. You should learn which people RARELY donate to charities and which people FREQUENTLY donate to charities – you will later be tested on these facts.

When you see the description of each person, you have 10 seconds to give your best guess of how often they donate to charities. You will then be told whether they RARELY or FREQUENTLY donate to charities and should learn this fact about them.

[**EXPLAIN**] Once you are told how often the person donates to charities, EXPLAIN out loud WHY that person RARELY or FREQUENTLY donates to charities.

[**CONTROL**] You will later be asked to EXPLAIN WHY each person RARELY or FREQUENTLY donates to charities.

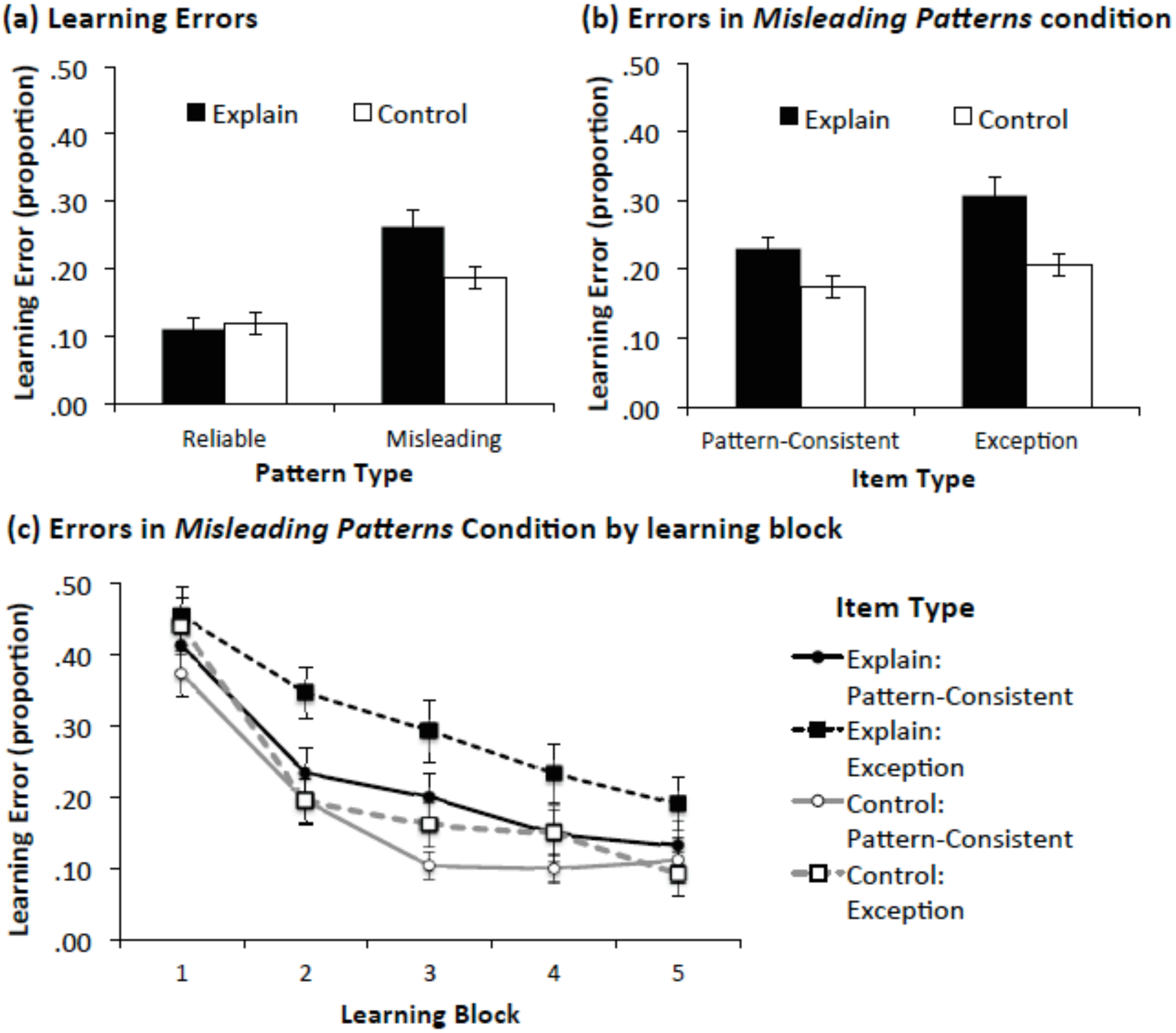
Do you think this person rarely or frequently donates to charities?

Steven is
25 years old
On the West Coast
Friendly
A humanities major

This person rarely donates to charities.

Steven is
25 years old
On the West Coast
Friendly
A humanities major

Figure 4. (a) Average proportion of errors during learning for Experiment 2 as a function of study condition and pattern type; (b) Errors during learning for the *misleading patterns* conditions as a function of study condition and item type; (c) Errors from the *misleading patterns* condition for pattern-consistent and exception items as a function of training block number. Error bars correspond to one standard error of the mean in each direction.



(a) Learning Errors

(b) Errors in *Misleading Patterns* condition

(c) Errors in *Misleading Patterns* Condition by learning block

Table 1. Stimuli from the *reliable pattern* condition in Experiment 1. In the *misleading pattern* condition a set of pattern-related features was randomly switched across the Dax/Kez categories.

Category	Unique features	Pattern–related features	Irrelevant features		
	Color	Cold/Warm climate	Transmission	Seat covers	Doors
Dax	Blue	Drives on Glaciers	Manual	Cloth	Two
	Silver	Made in Norway	Automatic	Vinyl	Two
	Purple	Used in Mountain Climbing	Automatic	Vinyl	Four
	Red	Heavily Insulated	Manual	Vinyl	Four
	Yellow	Has Treads	Manual	Cloth	Two
Kez	Cyan	Drives in Jungles	Manual	Vinyl	Four
	Magenta	Made in Africa	Manual	Cloth	Four
	Olive	Has Wheels	Automatic	Cloth	Two
	Maroon	Lightly Insulated	Manual	Vinyl	Two
	Lime	Used on Safaris	Automatic	Vinyl	Two

Table 2. Stimuli from Experiment 2. In the *misleading pattern* condition the age and picture of the fourth and eighth individuals were switched, as were the personality features of the fifth and tenth individuals.

Behavior	**Unique features**		**Pattern–related features**		**Irrelevant features**	
	Picture	**Name**	**Age**	**Personality**	**A graduate of a**	**Living on the**
Rarely donates to charities (Frequently donates to charities)		**Laura**	**30**	**dominating**	**science major**	**East coast**
		Steven	**25**	**friendly**	**humanities major**	**West coast**
		Jessica	**32**	**boastful**	**science major**	**West coast**
		Janet	**26**	**self-assured**	**science major**	**East coast**
		Kevin	**23**	**energetic**	**humanities major**	**West coast**
Frequently donates to charities (Rarely donates to charities)		**Joseph**	**37**	**cautious**	**science major**	**East coast**
		Josh	**47**	**discreet**	**humanities major**	**West coast**
		Karen	**39**	**studious**	**science major**	**West coast**
		Anna	**45**	**self-conscious**	**humanities major**	**West coast**
		Sarah	**42**	**quiet**	**science major**	**East coast**

5. Conclusion

The preceding three chapters presented the novel Subsumptive Constraints account of why generating explanations is key to learning and transfer, and provided empirical tests of three of its core predictions. The central idea is that explaining "why?" exerts a *subsumptive constraint* that drives people to seek underlying patterns and generalizations. Explaining why a fact is true *constrains* how learners reason, guiding them to understand the fact as an instance of a broader pattern – in other words, to *subsume* the fact as an instance of a generalization. Although past studies have identified patterns and principles in unstructured self-explanations (e.g., Chi et al, 1994; Renkl, 1997), the preceding chapters provide empirical evidence that explaining "why?" directly drives their discovery and construction.

The nine experiments reported in this dissertation provide evidence for the Subsumptive Constraints account's distinctive set of predictions about the effects of seeking why-explanations. Respectively: (1) Chapter two provided evidence that explaining "why?" selectively drives people to discover underlying patterns, particularly favoring broad, unifying generalizations. (2) Chapter three reported studies showing that explanation's constraint to prefer broader patterns also extends to those patterns that learners' prior beliefs suggest are likely to apply to a range of novel contexts. Explaining thus increases learners' consultation of their prior knowledge in guiding the discovery and generalization of patterns beyond what is being explained. (3) Chapter four suggests that although the constraint to seek patterns is helpful in many contexts, when regularities are spurious or misleading, explaining can *impair* learning through *overgeneralization*.

In this concluding chapter, I briefly review the evidence each chapter provided for its corresponding claim, considering the implications of each finding for existing research, practical applications, and future directions. I then close with a consideration of future research directions that build on the body of work presented in this dissertation.

5.1. Subsumptive Constraints on explaining are what drive the discovery of broad patterns that support generalization.

Chapter two explicated the Subsumptive Constraints account, testing the prediction that explaining "why?" drives people to search for underlying patterns. While the complex materials in past research attest to the consequential effects of explanation, the complexity poses a challenge in characterizing the underlying mechanisms. The work reported in Chapter two complements such work by using controlled stimuli in a laboratory context. We explored category learning because of its extensive study in past research and its potential generality – people learn novel categories across domains from mathematics to biology.

The learning task involved studying examples from two categories of novel alien robots. These could be partially categorized using an imperfect but salient pattern, or perfectly categorized using a broader but difficult to detect pattern. Prompts to explain why a robot belonged to the category it did promoted discovery of the subtle but broader pattern, as well as the generalization of this pattern to examples that had not been studied before. A comparison to closely matched control conditions revealed that explaining "why?" generated greater discovery than prompts to describe the robots (chosen to increase attention and verbal elaboration) and prompts to think out loud (chosen to promote metacognitive monitoring and awareness of thoughts through explicit verbalization). However, there was no advantage in memory for examples, and in some cases explaining even impaired memory.

This research indicates that explaining “Why?” does not serve simply as a generally elaborative or all-purpose boost to learning engagement, it can be useful in promoting induction or comprehension of underlying principles and causal patterns, rather than increasing memory or retention of facts– unless these are tied to underlying principles.

The effect of explaining may also be greater for domains in which underlying principles are key for learning and can be induced by the learner– previous demonstrations of explanation’s effects in learning physics or theory of mind may occur precisely because these are such domains.

Moreover, the current findings suggest that successful identification of principles and generalizations should stem from asking “why?” more than from other questions that increase general levels of elaboration and monitoring. One context in which this finding could be further tested would be learning from worked examples of problems in a cognitive tutor. One testable implication of the current dissertation work is that prompts to explain why an answer is correct might be most effective in constructing abstract generalizations, while explaining one’s reasoning should be less helpful for such generalizations, even if it were to make a greater contribution to metacognitive benefits.

5.2. Explanation & prior knowledge interact to guide learning.

The third chapter uses the Subsumptive Constraints account to provide novel insight into how explanation and prior knowledge interact in learning. Past research has found mixed results as to whether explaining helps more with low or high prior knowledge learners, possibly because this work has examined multiple kinds of explanations and numerous types of “prior knowledge”.

The prediction tested was that when learners evaluate explanations, they are driven to consult prior knowledge that is relevant to assessing how broad or unifying a pattern is. Two explanations might account equally well for the specific facts being explained, but prior beliefs can favor one as more likely to account for a diverse range of additional facts. The experiments manipulated whether or not people were prompted to explain the category membership of examples, and whether or not the category labels provided information favoring the more subtle of two patterns that *both* perfectly categorized all the examples. Neither explaining nor receiving this knowledge was sufficient to promote discovery of the subtle pattern, but together they interacted in increasing its discovery. Even when the patterns were made equally salient and discovery equated, a similar interaction was found in determining which pattern was generalized to novel examples, and participants’ judgments of how many novel examples the pattern would apply to.

Explaining may therefore promote learning by deploying prior knowledge that is relevant to discovering broad generalizations, preventing such knowledge from lying inert and failing to influence learning. This sheds light on what kind of knowledge is relevant to successful construction of “why?” explanations. Providing knowledge that helps students understand how a principle applies – to what is being explained and to related contexts – could support students’ construction of explanations.

5.3. The hazards of explanation: Overgeneralization in the face of exceptions.

Chapter four tested a unique prediction of the Subsumptive Constraints account: By driving learners to seek patterns, explaining "why?" can *impair* learning when patterns are misleading or violated by exceptions. We explored this in both the context of learning people's behavior and learning about object categories. Prompts that increased learners' search for explanations (relative to thinking aloud, or anticipating having to explain) *impaired* learning when patterns were misleading because exceptions to these patters existed. Poorer learning occurred because explaining drove people to ignore specific observations in favor of *overgeneralizations* (such as imperfect relationships between an individuals' age and donations to charity).

This finding provides strong evidence that explaining "why?" selectively constrains the search for generalizations, rather than producing an all-purpose boost to motivation or metacognition (which could still underlie the effects of other kinds of explanation prompts).

The finding that explaining carries both benefits and hazards provides a compelling case that explaining does not simply provide a general processing boost, and that closer investigation of its underlying processes and most effective pedagogical use is warranted. Not only can this avoid potential impairments, but it can elucidate *which* contexts and kinds of knowledge are *especially* well suited to learning by explaining "why?"

The importance of selecting cases that highlight relevant subsuming generalizations can be explored in future work. For example, in an educational setting, cases may need to be sufficiently diverse to target an abstract rather than shallow underlying generalization, but not so broad that explaining them requires induction of a subsuming generalization beyond the learner's current knowledge or reasoning capacity. In everyday learning contexts, explaining may be more likely to drive people towards misleading generalizations when they explain small data sets of observations, but useful for discovering true regularities when explaining is directed at sufficiently representative observations. This could be the case in phenomena such as detecting illusory correlations, forming stereotypes, or detecting systematicity in random data.

The findings also suggest a way to reconsider the role that students' unprompted and covert explanations might play in "buggy" reasoning in mathematics – where overgeneralization and undergeneralization from examples of mathematics problems and solutions results in systematic errors. Consideration of students' spontaneous explanations might also shed light on misconceptions or alternative conceptions in science education, and suggest which explanation prompts could avoid or revise those beliefs. Understanding might be best achieved by combining "why?" questions with other kinds of explanation prompts, and varying what knowledge is communicated through instruction. This will require both basic research – comparing the effects of different explanatory questions and providing different kinds of instructional explanations – as well applied projects that consider how best to combine them.

5.4. Future directions

The Subsumptive Constraints account provides a new framework within which to interpret past instances of explanation's effects, accounting for the importance of explanation in discovering principles that support generalization to novel contexts. It also adds precision to the operationalization of "explanation" by focusing on explanations that are responses to "Why?"

questions. By articulating a specific account of explanation and learning – that brings together research on explanation across education, cognitive psychology, development, philosophy, and artificial intelligence – this dissertation work also provides a fruitful framework for a number of future research directions. I now consider several of these.

5.4.1. Explanation's role in learning and revising beliefs in the face of anomalous observations.

In the more practical and educational context of learning about z-scores, current work (in collaboration with Caren Walker, Sam Maldonado, and Tania Lombrozo) is finding that prompting people to explain observations that conflict with their beliefs can revise existing knowledge, even though asking people to articulate their thoughts about these anomalous observations yields minimal learning benefits.

5.4.2. Children's causal learning through explaining.

In a collaboration led by Caren Walker (and including Tania Lombrozo, Alison Gopnik, and Cristine Legare), we have been testing whether children as young as five might be sensitive to Subsumptive Constraints when they explain "why?" Despite having less knowledge and language abilities than adults, we find that prompting children to explain novel causal outcomes makes them more likely to access their prior knowledge, and to use it in generalizing causal relationships.

5.4.3. Relationship between explanation and other cognitive processes.

The Subsumptive Constraints account identifies explaining as establishing the implicit goal of finding patterns, and so predicts that explaining will recruit other cognitive processes to do so. This suggests a fruitful exploration of the relationship between explanation and *comparison* of examples – a related and educationally-relevant instructional strategy. I have been collaborating with Brian Edwards, who is in examining the relative and joint effects of generating explanations and making comparisons on learning about patterns, using the materials first developed for the research reported in the second and third chapters.

5.4.4. Benefits of generating and receiving explanations.

One challenge I anticipate in learning from explaining is that the benefits of prompting students to explain may be only partially realized without additional support for constructing explanations. A straightforward way of providing the knowledge that helps people to construct an explanation may in fact be to also provide a correct explanation. Teachers and other educators might expect this to be redundant or find it strange for a student to be asked to both generate *and* be given an explanation. But if this is beneficial and can be done in a natural way, its ease of implementation could yield substantial practical benefits. For example, students could be asked to generate an explanation, and then asked to compare their self-generated explanation to a different explanation that was ostensibly generated by another student (although it could in fact be the correct or normative explanation, designed to instruct them).

This issue could be explored in the context of online education – adding explanation prompts and instructional explanations to improve students' learning from interactive online mathematics exercises. In addition to exploring the relative and joint impact of generating and receiving explanations, computational analyses and Natural Language Processing can be used to maximize learning by tailoring the explanations students are provided with to the explanations they generate.

5.5 Conclusion

The work presented in this dissertation has shed light on why explaining promotes learning – by driving people to seek and discover underlying patterns. This was shown to produce benefits – like discovering subtle regularities and increasing use of existing knowledge – as well as costs – ignoring individual contradictory observations and *over*generalizing. By fruitfully bridging cognitive psychology and education, this body of research provides a foundation for future work the blends theoretical research and practical applications for explanation. A particularly promising direction is to use online educational contexts to continue to bring together rigorous experimental studies, statistical modeling, and large practical impact in real-world settings.

6. References

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

Ahn, W., Brewer, W. F., & Mooney, R. J. (1992). Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory*, & Cognition. *18*(2), 391-412.

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*.

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., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation vs. mechanism information in causal attribution. Cognition, 54, 299-352.

Ahn, W., Marsh, J., Luhmann, C., & Lee, K (2002). Effect of theory-based feature correlations on typicality judgments. *Memory & Cognition, 30*, 107-118.

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.

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, *120*(1), 3–19.

Amsterlaw, J. A., & Wellman, H. M. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Child Development, 7(2),* 139-172.

Ashby, G., & Maddox, T. (2005). Human category learning. *Annual Review of Psychology*, *56*(1), 149-178.

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

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

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.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. Cognitive Psychology, 20(4), 493–523.

Callanan, M.A. & Oakes, L. (1992). Preschoolers' questions and parents' explanations: causal thinking in everyday activity. *Cognitive Development, 7,* 213-233.

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: Bradford Books, MIT Press.

Carey, S. (1991). Knowledge acquisition: enrichment or conceptual change? In S. Carey & R. Gelman (eds.), *The Epigenesis of Mind: Essays in Biology and Cognition* (pp. 257-291). Hillsdale, NJ: Erlbaum.

Carroll, J. W. (2008). Laws of nature, in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, URL = <http://plato.stanford.edu/archives/fall2008/entries/laws-of-nature/>.

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-constructive-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., DeLeeuw, 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.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in instructional psychology*, *5*, 161–238.Crowley, K., & Siegler, R. S. (1999). Explanation and generalization in young children's strategy learning. *Child Development, 70*, 304-316.

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.

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.

Chouinard, M. (2007). Children's questions: a mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, 72, 1-57.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, 11(6), 671-684.

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

DeJong, G. (2006). Toward robust real-world inference: A new perspective on explanation-based learning. In ECML06, the Seventeenth European Conference on Machine Learning, pp. 102-113.

DeJong, G., & Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine learning*, *1*(2), 145–176.

Dienes, Z., Broadbent, D., & Berry, D. C. (1991). Implicit and explicit knowledge bases in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 875.

Fiske, S. T., & Taylor, S. E. (1984). *Social cognition*. New York: Random House.

Fonseca, B. & Chi, M.T.H. (2010). The self-explanation effect: A constructive learning activity. In Mayer, R. & Alexander, P. (Eds.), *The Handbook of Research on Learning and Instruction* (pp. 270-321). New York, USA: Routledge Press.

Frazier, B.N., Gelman, S.A., Wellman, H.M. (2009). Preschoolers' search for explanatory information within adult-child conversation. *Child Development*, 80(6): 1592-1611.

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, *71*(1), 5-19.

Gentner, D. (2010). Bootstrapping the Mind: Analogical Processes and Symbol Systems. *Cognitive Science*, *34*(5), 752–775. doi:10.1111/j.1551-6709.2010.01114.x

Gigerenzer, G. (2004). *Fast and Frugal Heuristics: The Tools of Bounded Rationality*. (D. J. Koehler & N. Harvey, Eds.) *Blackwell handbook of judgment and decision …* (pp. 62–88). Malden, MA, USA: Blackwell Publishing Ltd. doi:10.1002/9780470752937.ch4

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.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

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.*

Hamilton, D. (1981). Illusory correlation as a basis for stereotyping. In D. Hamilton (Ed.), Cognitive processes in stereotyping and intergroup behavior (pp. 115-144). Hillsdale, NJ: Lawrence Erlbaum.

Hampton, J. A. (2006). Concepts as prototypes. The Psychology of learning and motivation: Advances in research and theory, vol. 46, pp. 79-113. San Diego: Academic Press.

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.

Hickling, A. K., & Wellman, H. M. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology*, *37,*668-683.

Kalish, C. W. (2002). Gold, Jade, and Emeruby: The value of naturalness for theories of concepts and categories. *Journal of Theoretical and Philosophical Psychology, 22*, 45-56.

Keil, F.C. (2006). Explanation and understanding. *Annual Review of Psychology, 57*, 227-254.

Kelemen, D. (1999). Functions, goals and intentions: Children's teleological reasoning about objects. *Trends in Cognitive Sciences*, 12, 461-468.

Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 829-846.

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*(4), 507-531.

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.

Kitcher and Wesley Salmon (Eds.), *Minnesota Studies in the Philosophy of Science, Volume XIII: Scientific Explanation* (pp. 410-505). University of Minnesota Press.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110*, 499-519.

Koslowski, B. (1996). *Theory and Evidence: The Development of Scientific Reasoning*. Cambridge, MA: MIT 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, T. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

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

Lee, M. D., & Vanpaemel, W. (2008). Exemplars, Prototypes, Similarities, and Rules in Category Representation: An Example of Hierarchical Bayesian Analysis. *Cognitive Science: A Multidisciplinary Journal*, *32*(8), 1403–1424.

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.M. (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.

Legare, C.H., Wellman, H.M., & Gelman, S.A. (in press). Evidence for an explanation advantage in naïve biological reasoning. *Cognitive Psychology.*

Lewis, C. (1988). Why and how to learn why: Analysis-based generalization of procedures. *Cognitive Science, 12,* 211-256.

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

Lombrozo, T. (2012). Explanation and abductive inference. K.J. Holyoak and R.G. Morrison (Eds.), *Oxford Handbook of Thinking and Reasoning (*pp. 260-276), Oxford, UK: Oxford University Press.

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass, 6*, 539-551.

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.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

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.

Malle, B. F. (2011). Attribution theories: How people make sense of behavior. In Chadee, D. (Ed.), *Theories in social psychology* (pp. 72-95). Wiley-Blackwell.

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

Master, A., Markman, E. M. and Dweck, C. S. (2012). Thinking in Categories or Along a Continuum: Consequences for Children's Social Judgments. *Child Development*, 83, 1145–1163. doi: 10.1111/j.1467-8624.2012.01774.x

Matthews, P. G., & Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of Experimental Child Psychology, 104,* 1-21.

Mathews, R., 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), 1083–1100.

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

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207–238.

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 333–352.

Meissner, C. A., & Memon, A. (2002). Verbal Overshadowing: A special issue exploring theoretical and applied issues. *Applied Cognitive Psychology*, *16*(8), 869-872.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization: A unifying view. *Machine learning*, *1*(1), 47–80.

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*(6), 543.

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*.

Nokes, T. J., & Ohlsson, S. (2005). Comparing multiple paths to mastery: What is learned? *Cognitive Science*, 29, 769-796.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

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

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–53.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 282–304.

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

Patalano, A.L., Chin-Parker, S., Ross, B.H. (2006). The importance of being coherent: Category coherence, cross-classification, and reasoning. *Journal of Memory & Language*, *54*, 407-424.

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.

Pine, K. J., & Messer, D. J. (2000). The effect of explaining another's actions on children's implicit theories of balance cognition and instruction. *Cognition and Instruction, 18(1),* 35-51.

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

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, *77*(3), 353.

Quine, W.V.O., & Ullian, J.S. (1970). *The Web of Belief.* New York, NY: Random House.

Ram, A., & Leake, D. B. (1995). *Goal-driven learning*. The MIT Press.

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.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219. doi:10.1006/cogp.2000.0743

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.

Risen, J. L., & Gilovich, T., Dunning, D. (2007). One-shot illusory correlations and stereotype formation. *Personality and Social Psychology Bulletin, 33, 1492-1502.*

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

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573–605.

Roscoe, R. & Chi, M. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research, 77*(4), 534-574.

Roscoe, R.D. & Chi, M.T.H. (2008). Tutor learning: the role of explaining and responding to questions. Instructional Science. 36(4), 321-350.

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

Salmon, W., 1989, *Four Decades of Scientific Explanation*, Minneapolis:University of Minnesota Press.

Salmon, W., 1984, *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.

Salmon, W., 1971a, ‘Statistical Explanation’, in *Statistical Explanation and Statistical Relevance*, W. Salmon, (ed.), 29–87, Pittsburgh: University of Pittsburgh Press.

Schooler, J. W. (2002). Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology*, *16*(8), 989–997. doi:10.1002/acp.930

Shanks, D. R., & John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and brain sciences*, *3*(17), 367–447.

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.

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

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3-22.

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.

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

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

Toth, J. P., Reingold, E. M., & Jacoby, L. L. (1994). Toward a redefinition of implicit memory: process dissociations following elaborative processing and self-generation. *Journal of experimental psychology: Learning, Memory, and Cognition*, *20*(2), 290.

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

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. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 261–279). Oxford, England: Oxford University Press.

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press, USA.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: evidence from category learning. *Cognitive Science*, *34*, 776-806. http://dx.doi.org/10.1111/j.1551-6709.2010.01113.x

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology*, 66(1), 55-84. http://dx.doi.org/10.1016/j.cogpsych.2012.09.002

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* (pp. 1352–1357). Austin, TX: Cognitive Science Society.

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.

Woodward, James, "Scientific Explanation", The Stanford Encyclopedia of Philosophy (Spring 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2009/entries/scientific-explanation/>.

Wylie, R., Koedinger, K. R., & Mitamura, T. (2009). Is Self-Explanation Always Better? The Effects of Adding Self-Explanation Prompts to an English Grammar Tutor. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1300-1305). Austin, TX: Cognitive Science Society.

Yamauchi, T., & Markman, A. B. (2000). Learning categories composed of varying instances: The effect of classification, inference, and structural alignment. *Memory & cognition*, 28(1), 64-78.