COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 34 (2010) 776–806

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2010.01113.x

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

Joseph J. Williams, Tania Lombrozo

Department of Psychology, University of California, Berkeley

Received 15 June 2009; received in revised form 19 December 2009; accepted 28 December 2009

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.

Keywords: Explanation; Self-explanation; Learning; Transfer; Generalization; Category learning; Anomalies

1. 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?

Correspondence should be sent to Joseph J. Williams, Department of Psychology, University of California, Berkeley, 3210 Tolman Hall, Berkeley, CA 94720. E-mail: joseph_williams@berkeley.edu

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 (Carey, 1991; Lombrozo, 2009; Murphy & Medin, 1985), 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, Bassok, Lewis, Reimann, & Glaser, 1989), declarative learning from biology texts (Chi, de Leeuw, Chiu, & LaVancher, 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 with alternative study strategies like thinking aloud, reading materials twice, or receiving feedback in the absence of explanations (e.g., Amsterlaw & Wellman, 2006; Chi et al., 1994; Siegler, 2002; Wong, Lawson, & Keeves, 2002), 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* (DeJong & Mooney, 1986; Mitchell, Keller, & Kedar-Cabelli, 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 its 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, 2007), and by encouraging learners to posit unobserved causes (Legare, Gelman, & Wellman, in press; Legare, Wellman, & Gelman, 2009). Given the diversity of the processes that can underlie

learning (Nokes & Ohlsson, 2005), it is likely that explanation influences learning via multiple mechanisms.

1.1. Exploring the role of explanation in generalization

In this study, 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. Although 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, 2007). 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, 1989; 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.

Although 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, Marsh, Luhmann, & Lee, 2002). Moreover, compared with 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 in diverse ways (see Allen & Brooks, 1991; Ramscar et al, 2010; Yamauchi & Markman, 2000), 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 (Ashby & Maddox, 2004; Bruner, Goodnow, & Austin, 1956; Nosofsky, Clark, & Shin, 1989), rules with exceptions (Nosofsky, Palmeri, & McKinley, 1994), similarity to prototypical summary representations (Hampton, 2006; Posner & Keele, 1968; Rosch & Mervis, 1975), 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; Lee & Vanpaemel, 2008; Medin, Altom, & Murphy, 1984; Murphy, 2002).

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. 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 features of this series of experiments are worth emphasizing. First, the explanation condition is compared with *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

Table 1
Overview of experiments: key differences

	Introduction	Study Phase	Control Condition
Exp. 1	Informed about two categories	8 items × 50 s (accompanied by image of all 8)	Describe
Exp. 2	Informed about two categories, and memory and categorization tests; exposure to 3 repeated blocks of all 8 items	2 items × 90 s (accompanied by image of all 8); 2 items: consistent vs. anomalous	Think aloud
Exp. 3	Informed about two categories, and memory and categorization tests	Image of all 8 for120 s	Free study

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 is 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.

3. 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 it 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.

3.1. Methods

3.1.1. Participants

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

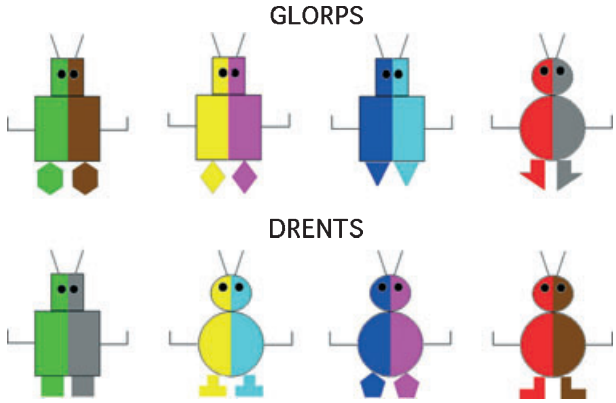

Fig. 1. Study items in Experiment 1.

3.1.2. Materials

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

3.1.2.1. 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, gray, 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.''

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.

3.1.2.2. 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).

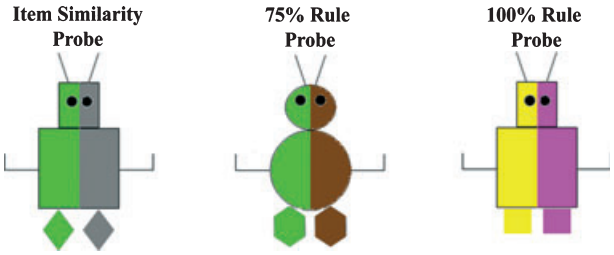

Fig. 2. Examples of three types of test items from Experiment 1.

3.1.2.3. 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, although the test items introduced new *combinations* of old features, the transfer items actually involved *new features* (new foot shapes).

3.1.2.4. 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).

3.1.3. Procedure

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

3.1.3.1. 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 s and kept it until the end of the study phase.

3.1.3.2. 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 s. Participants were not allowed to advance more quickly nor take extra time. After the study phase, the experimenter removed the sheet showing the eight robots.

3.1.3.3. 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 2 s. Participants were informed of this delay and the screen flickered after the 2-s period ended.

3.1.3.4. Memory phase: The 8 study items (35%) and 15 lures (65%) were presented in a random order, and participants judged whether each robot was one of the original robots from the introduction and study phases. As in categorization, items had to be onscreen for 2 s.

3.1.3.5. 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.

3.2. Results

3.2.1. 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 vs. flat feet), 75% rule (square vs. 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, whereas 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 [$\chi^2(1) = 15.89$, $p < .001$], while the 75% rule was more prevalent in the *describe* than the *explain* condition [$\chi^2(1) = 19.56$, $p < .001$]. ''Item similarity'' and ''other'' responses were not significantly associated with condition.

Although 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

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

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.

3.2.2. 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. Fig. 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) × 2 (*categorization measure*: test vs. transfer) mixed ANOVA was conducted on categorization accuracy. This revealed a main effect of *task* [$F(1,148) = 16.10$, $p < .001$], 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* [$F(1,148) = 13.46$, $p < .001$], 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 [$t(148) = 2.91$, $p < .01$] suggests that they not only recognized the importance of foot shape in determining category membership but also 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 [$t(148) = 4.41$, $p < .001$], whereas categorization of the *75% rule probes* was more consistent with the 75% rule in the *describe* than the *explain* condition [$t(148) = 3.77$, $p < .001$]. There was no difference for item similarity probes [$t(148) = 1.37$, $p = .17$]. 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

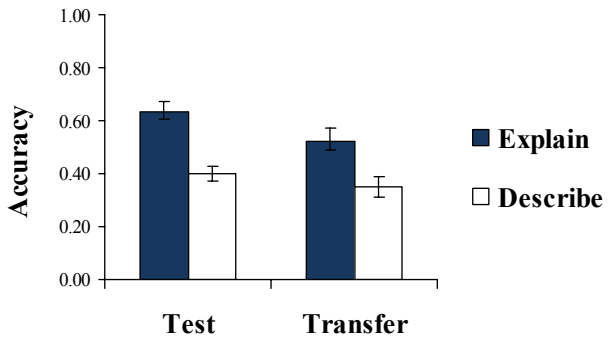

Fig. 3. Categorization accuracy on test and transfer items in Experiment 1.

additionally analyzed categorization accuracy using a nonparametric measure. Each participant was coded as relying on the 100% rule if 7 or 8 of the eight 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 [$\chi^2(1) = 10.37, p < .01$].

3.2.3. Memory for study items

Because engaging in explanation may have drawn special attention to the anomalous items (an issue addressed in Exp. 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. Fig. 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 [$t(148) = 2.24, p < .05$]. There was no difference in discrimination of *anomalous* study items [$t(148) = 0.82, p = .41$, 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 Fig. 5). There was no significant difference in memory performance for those who explicitly cited the foot rule [$t(30) = 0.88, p = .88$] or the body rule [$t(52) = 1.41, p = .16$), but there was a significant difference in memory for participants coded in the ''other'' category [$t(62) = 2.19$,

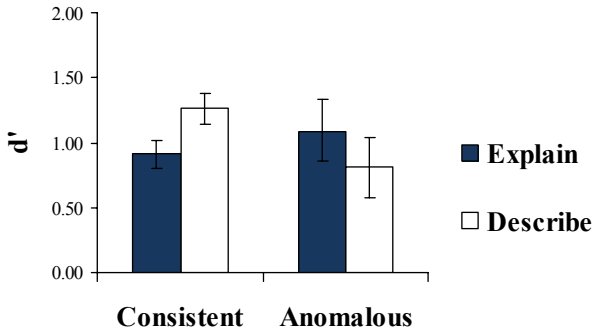

Fig. 4. Memory for consistent and anomalous items in Experiment 1.

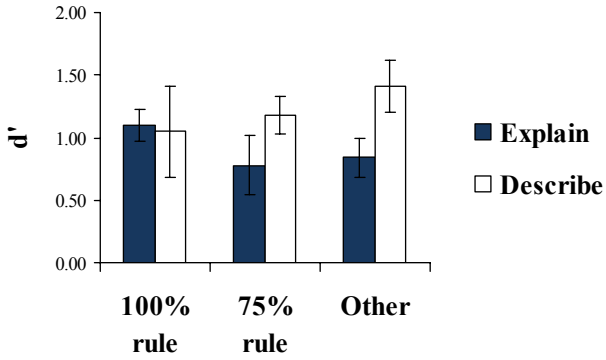

Fig. 5. Memory for consistent items as a function of basis for categorization in Experiment 1.

$p < .05$]. This suggests that the memory difference across conditions is not due to discovery of the 100% rule leading participants to ignore item details, but it may stem from a difference between the efficacy of explaining and describing for encoding specific details.

3.2.4. Coded content of explanations and descriptions

Each of the eight 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 Chin-Parker et al., 2006; Wisniewski & Medin, 1994). References were coded as *concrete* if they cited the actual feature, for example, triangle/square/L-shaped feet, square/round body, and yellow/green color. References were coded as *abstract* if they characterized a feature in more general terms, which could be applied to multiple features, for example, pointy/flat feet, big/strange body, and warm/complementary colors. Two experimenters coded explanations and descriptions independently, with agreement of 97% (analyses used the first coder's responses). Fig. 6 shows the number of features mentioned in each coding category as a function of *task*. Two separate 2 (*task*: explain vs. describe) × 3 (*feature*: feet vs. body vs. color) ANOVAs were conducted on the total number of *concrete* (*abstract*) features

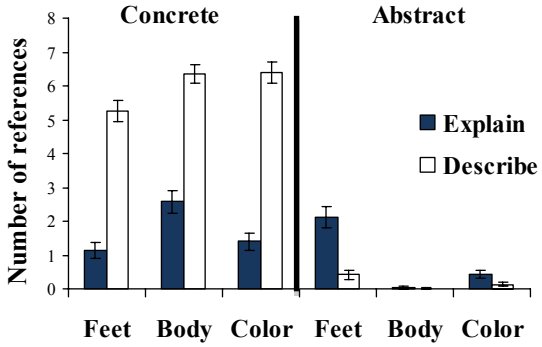

Fig. 6. Coding of feature references in explanations and descriptions in Experiment 1.

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*, $F(1,148) = 24.72$, $p < .001$], whereas those in the *describe* condition cited more concrete features than those who explained [a main effect of *task*, $F(1,148) = 164.65$, $p < .001$]. Individual t tests confirmed that these two findings were reliable for all features (all ps $< .05$) except abstract references to body shape [$t(148) = 0.82$, $p = .41$).

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 they are consistent with an attentional explanation for the enhanced memory found in the *describe* condition.

3.3. 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. Dye and Ramscar (2009) report a similar dissociation between feature discovery and memory in a prediction task, depending on whether category labels precede exemplars or exemplars precede labels.

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: $\chi^2(1) = 1.21$, $p = .27$]. 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.

4. 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, in Experiment 2 the participants who think aloud will have comparable benefits with 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 it 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, regardless of whether 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 consistent with 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 × 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 eight items, but the targets of explaining or thinking aloud were either *consistent* or *anomalous* observations.

4.1. Methods

4.1.1. Participants

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

4.1.2. 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.

4.1.3. Procedure

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

4.1.3.1. 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*.

4.1.3.2. Prestudy exposure: After participants received and viewed the sheet of robots, the introduction phase was augmented by presenting each of the eight robots onscreen. A block consisted of displaying each of the eight robots for 4 s 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 eight

study items, although only two items were displayed onscreen for the explain or think-aloud phase.

4.1.3.3. Study phase: Although participants provided explanations (descriptions) for all eight robots in Experiment 1, the Experiment 2 study phase only presented two robots (one glorp and one drent) for 90 s each, with a warning when 30 s were left. In the *consistent* condition, the two robots were randomly selected from the six consistent with the 75% rule, whereas 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, whereas the *think aloud* instructions were as follows: ''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.''

4.1.3.4. Test, transfer, and memory: The test, transfer, and memory phases were identical to Experiment 1, except that the restriction that responses could only be made after 2 s was removed.

4.1.3.5. Postexperiment 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?''

4.2. Results

4.2.1. 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

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

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: (a) reports reflecting discovery and use of the foot rule, and (b) all other responses. A hierarchical log-linear analysis with backwards elimination was carried out on the *task* × *item type* × *discovery* contingency table, revealing a highly significant interaction between *task* and *discovery* [$\chi^2(1) = 21.91$, $p < .001$]: Explaining was associated with discovery of the 100% rule. With post hoc tests comparing individual conditions, discovery was more frequent in both explain conditions than in either think-aloud condition [$\chi^2(1) = 8.71$, $p < .01$, $\chi^2(1) = 8.71$, $p < .01$; and $\chi^2(1) = 13.30$, $p < .001$, $\chi^2(1) = 13.30$, $p < .001$].

The benefit for explaining over thinking aloud was mirrored in categorization accuracy (see Fig. 7). A 2 (*task:* explain vs. think aloud) × 2 (*item type*: consistent vs. anomalous) × 2 (*categorization measure*: test vs. transfer) mixed ANOVA revealed a significant main effect of *task* [$F(1,236) = 21.90$, $p < .001$], with more accurate categorization in the *explain* condition. The effect of *item type* [$F(1,236) = 3.35$, $p = .07$] and the interaction between *task* and *item type* [$F(1,236) = 3.35$, $p = .07$] were marginal. There was additionally a significant effect of *categorization measure* [$F(1,236) = 14.38$, $p < .001$], with test accuracy higher than transfer, and significant interactions between *categorization measure* and *task* [$F(1,236) = 4.71$, $p < .05$] and *categorization measure* and *item type* [$F(1,236) = 4.71$, $p < .05$], 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 [$F(1,118) = 5.83$, $p < .05$] or in either think-aloud condition [$F(1,118) = 14.51$, $p < .001$, $F(1,118) = 12.68$, $p < .001$].

As in Experiment 1, categorization accuracy scores were not normally distributed, so a nonparametric 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* × *item type* × *inferred discovery* revealed an interaction between *task* and *inferred discovery* [$\chi^2(1) = 18.59$, $p < .001$] and also between *item type* and *inferred discovery* [$\chi^2(1) = 3.91$, $p < .05$]. This suggests that both explaining and the presence of anomalies contributed to discovery. The trend toward greater discovery in the explain-anomalous condition than the explain-consistent condition was marginal [$\chi^2(1) = 2.76$, $p = .10$].

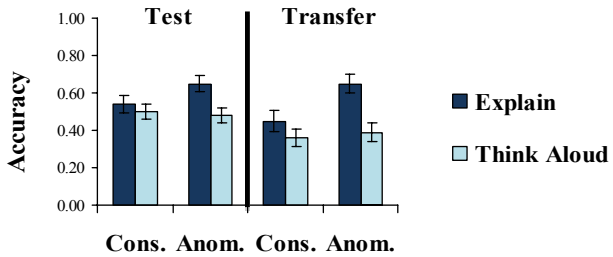

Fig. 7. Categorization accuracy in Experiment 2.

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

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: (a) explicit report of using the 75% rule, and (b) all other responses. A log-linear analysis revealed a significant three-way interaction between *task*, *item type*, and *body use* [$\chi^2(1) = 4.35$, $p < .05$]. Reliance on the 75% rule was more frequent in the explain-consistent than the explain-anomalous condition, approaching significance [$\chi^2(1) = 3.68$, $p = .055$], with no difference for the think-aloud conditions [$\chi^2(1) = 0.95$, $p = .33$].

4.2.2. Memory

Separate 2 × 2 ANOVAs were conducted on the discrimination measure d' for both consistent and anomalous items. There was an effect of *observation type* on the discrimination of anomalous items [$F(1,236) = 21.53$, $p < .001$], simply reflecting that the discrimination of anomalous items was better in the anomalous conditions. No other effects were significant (all ps > .30). Memory for the original items did not appear to be differentially influenced by explaining versus thinking aloud.

4.2.3. Postexperiment 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, the participant was 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); and 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 50%) of participants in the explain-consistent condition noticed and recalled the anomalies [$\chi^2(1) = 6.26$, $p < .05$], although they only provided explanations for the items consistent with the 75% rule.

4.3. 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 eight robots or recalling anomalies from the prestudy 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 arguably 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 are 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.

5. 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.

5.1. Methods

5.1.1. 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.

5.1.2. Materials

Participants saw all eight 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 one item similarity probe, one 75% rule probe, one item that received the same classification from all three bases, and four 100% rule probes. There were four transfer items. Memory items consisted of four old items and four lures.

5.1.3. 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 eight 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 min. 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 whether they thought there were, and if so 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 they were required to specify the order in which they would ask the questions if they had to decide whether the

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

obscured robot was a glorp or drent. The options were ordered randomly and were as follows: (1) What color is it? (2) What does its body look like? (3) What do its feet look like? and (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]

5.2. Results and discussion

5.2.1. Basis for categorization and categorization accuracy

Basis for categorization was coded from participants' explicit reports and the features they reported using in categorization, and this 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. Fig. 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 [$\chi^2(1) = 4.09$, $p < .05$] and in categorization accuracy [a main effect of *task* in a *task* × *categorization measure* ANOVA,

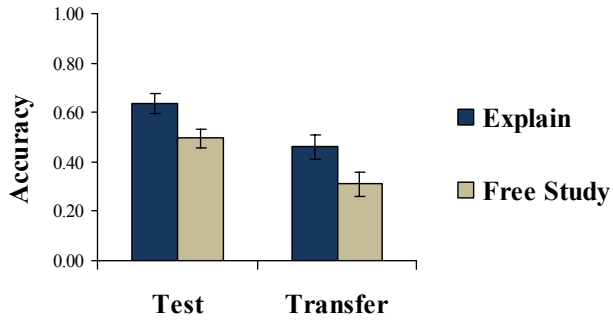

Fig. 8. Categorization accuracy in Experiment 3.

$F(1, 118) = 7.02$, $p < .01$]. Explaining was also significantly associated with reduced use of the 75% rule [$\chi^2(1) = 4.66$, $p < .05$].

5.2.2. Memory

There was no significant difference in memory (as measured by discrimination, d') for consistent items [$t(118) = 1.60$, $p = .11$: Explain: 0.89, Free study: 0.35] or anomalous items [$t(118) = 0.26$, $p = .80$: Explain: 0.74, Free study: 0.62].

5.2.3. 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 [$t(118) = 0.65$, $p = .52$; Explain: 37.9, Free Study: 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 ps $> .12$). This suggests that the effect of the prompt to explain was not simply to communicate to participants that there was a regularity present.

5.2.4. 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 [$\chi^2(1) = 2.41$, $p = .30$]. 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 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''). These data are displayed in Table 7. There were significant interactions between *task* and *discovery* [$\chi^2(1) = 4.17$, $p < .05$], and also between *explain-report* and *discovery* [$\chi^2(1) = 13.96$, $p < .001$]. 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.

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	Explain—yes	14	2	13
	Explain—other	3	7	21
Free study	Explain—yes	5	11	7
	Explain—other	3	8	26

6. 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 that drive 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 they 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 and 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.

Although 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 specific 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 simple pattern. 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. The subsumptive constraints account elucidates the mechanisms underlying explanation's effects, providing insight into the contexts in which explaining is or is not the most effective learning strategy.

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 insufficient data are available (for recent evidence that explanations are not 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.

6.1. 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 would not 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, although 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 concerns 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 Experiments 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. Although 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.

6.2. 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 toward 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; Lee & Vanpaemel, 2008; Sloman, 1996). 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 (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Nosofsky et al., 1994) 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. Reliance on rules, exemplars, or prototypes also does not exhaust the range of category-learning mechanisms that have been identified. Ramscar et al. (2010) report a ''Feature-Label-Ordering'' effect in which participants who learn a category structure through exposure to exemplars followed by category labels learn discriminating features better than those exposed to category labels followed by exemplars. It may be that explanation encourages the kind of processing observed in the former condition, although further research would be required to establish this connection.

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 they 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.

7. Future directions and conclusions

These experiments suggest the utility of subsumption and unification, but there is a great deal of future research to be carried out 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, e.g., 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 Kelemen, 1999; Lombrozo, 2009; Lombrozo & Carey, 2006) and explored the role simplicity plays in the evaluation of explanations (see E. Bonawitz & T. Lombrozo, unpublished data; Lombrozo, 2006). 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; DeJong & Mooney, 1986; Mitchell et al., 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.

Notes

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.
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 postexperiment debriefing unambiguously indicated a switch in category labels, that participant's categorization responses were reverse coded.
4. We interpret these 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 participants as having noticed the anomaly.

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

Acknowledgments

We thank Hava Edelstein and Ania Jaroszewicz for help with data collection, coding responses, analyzing data, discussion of this work, and feedback on previous drafts of this study; Jared Lorince and David Oh for help with data collection; and Randi Engle, Nick Gwynne, Cristine Legare, Luke Rinne, Steven Sloman, and the Concepts & Cognition Laboratory for discussion of this research and feedback on previous drafts of this study. This work was partially supported by the McDonnell Foundation Collaborative Initiative on Causal Learning. JJW was supported by a fellowship from the Natural Sciences and Engineering Research Council of Canada.

References

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

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

Ashby, F. G., & Maddox, W. T. (2004). Human category learning. *Annual Review of Psychology*, *56*, 149–178.

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

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 E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (fall 2008 edition)*. Available at: http://plato.stanford.edu/archives/fall2008/entries/laws-of-nature/. Accessed on May 27, 2010

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238), Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science education. *Review of Educational Research*, *63*, 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.

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

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

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

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, CA: Academic Press.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: John Wiley & Sons, Inc.

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.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, *48*, 507–531.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Minnesota studies in the philosophy of science, Volume XIII: Scientific explanation* (pp. 410–505). Minneapolis, MN: University of Minnesota Press.

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

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology*, *103*(3), 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., Gelman, S. A., & Wellman, H. M. (in press). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*.

Legare, C. H., Wellman, H. M., & Gelman, S. A. (2009). Evidence for an explanation advantage in naïve biological reasoning. *Cognitive Psychology*, *58*(2), 177–194.

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. (2009). Explanation and categorization: How ''why?'' informs ''what?'' *Cognition*, *110*, 248–253.

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

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

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.

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

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, MA: The MIT Press.

Murphy, G. L., & 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.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many quises. *Review of General Psychology*, *2*, 175–220.

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. 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.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

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.

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: Random House.

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

Salmon, W. C. (1989). *Four decades of scientific explanation*. Minneapolis, MN: University of Minnesota Press.

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

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

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

Strevens, M. (2008). *Depth: An account of scientific explanation*. Cambridge, MA: Harvard University Press.

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*. New York: Oxford University Press.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, *18*(2), 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, J. (2009). Scientific explanation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (spring 2009 edition)*. Available at: 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.