Explaining Promotes Discovery: Evidence from Category Learning

Joseph Jay Williams (joseph_williams@berkeley.edu)
Tania Lombrozo (lombrozo@berkeley.edu)
Department of Psychology, University of California, Berkeley

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 explores a potential mechanism underlying this effect, motivated by philosophical accounts of the structure of explanations: that explaining guides learners to interpret observations in terms of unifying patterns or regularities, which in turn promotes the discovery of broad generalizations. Experiment 1 finds that prompting participants to explain while learning artificial categories promotes the induction of a broad generalization underlying category membership. Experiment 2 suggests that explanation most readily prompts discovery in the presence of anomalies: observations inconsistent with current beliefs. Experiment 1 additionally suggests that explaining might result in reduced memory for details. These findings provide evidence for the proposed mechanism and insight into the potential role of explanation in discovery and generalization.

Keywords: explanation; learning; generalization; self-explanation; category learning

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

Psychologists and philosophers have independently proposed that in explaining observations about the past, we uncover underlying structure in the world, acquiring the knowledge to predict and control the future. For example, in explaining a friend's behavior, you might come to appreciate the extent of her ambition, which informs expectations about her future actions.

Research in education and cognitive development confirms that the process of explaining – even to oneself – can foster learning. This phenomenon is known as the *self-explanation effect*, and has been documented in a broad range of domains: acquiring procedural knowledge about physics problems (Chi et al., 1989), declarative learning from biology texts (Chi et al., 1994), and conceptual change in children's theory of mind (Amsterlaw & Wellman, 2006), to name only a few. Compared to alternative study strategies like thinking aloud, reading materials twice, or receiving feedback in the absence of explanations (e.g. Chi, 1994; Amsterlaw & Wellman, 2006), self-explanation consistently leads to greater learning, with the greatest benefit for transfer and generalization to problems and inferences that require going beyond the material originally studied.

Researchers have made 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, and its role in dynamically repairing learners' mental models of particular domains (e.g. Chi, 1989; 1994). Given the diversity of the processes which can underlie learning (Nokes & Ohlsson, 2005), it is likely that explanation influences learning via multiple mechanisms.

In this paper we explore why explaining plays such an important role in transfer and generalization. We investigate the hypothesis that engaging in explanation will promote the discovery of broad, abstract generalizations that underlie what is being learned. 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.

While there is no consensus, we focus on *pattern subsumption* theories, which identify good explanations as those that demonstrate how what is being explained is an instance of a general pattern (for discussion see Strevens, 2008; for suggestive empirical evidence see Lombrozo & Carey, 2006; Wellman & Liu, 2006). A subset of these accounts further emphasizes *unification*: the value of explaining disparate observations by appeal to a single explanatory pattern. 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), and this general pattern can account for both this observation and a range of other data.

Subsumption and unification accounts of explanation predict the privileged relationship between explanation and generalization demonstrated by the self-explanation effect. If the explanations people construct satisfy the structural demands of subsumption, then successfully engaging in explanation should result in the induction or explicit recognition of generalizations that underlie what is being explained. Generating or explicitly representing such generalizations should in turn facilitate the transfer of what is learned in one context to novel but relevant contexts. We therefore investigate the hypothesis that the *structure* of explanations contributes to the relationship between explanation and generalization, and that explaining will drive learners to discover broad, abstract generalizations that support effective transfer to novel contexts.

To test our hypothesis we employ a task from cognitive psychology: learning artificial categories from positive examples. Previous work on category learning suggests that categories are more coherent to the extent they support explanations (Patalano, Chin-Parker, & Ross, 2006), and that background beliefs that explain feature combinations facilitate category learning (Murphy & Allopenna, 1994) and influence judgments of a category member's typicality (Ahn, 2002). This research demonstrates that "explanatory" background knowledge impacts category representations, but has not considered the self-explanation effect or the properties of explaining that may support learning.

Examining generalization in the context of category learning has two benefits. First, learning about categories is a fundamental inductive generalization that children and adults regularly face, and one that has significant consequences for one's ability to reason about and successfully navigate a complex world. If explaining influences category learning, there is good reason to expect explanation to play a role in other forms of learning. Second, as category members vary along many dimensions, there are multiple generalizations one might draw about the basis for category membership, so that category learning provides a natural setting in which to precisely investigate the nature of the generalizations prompted by explaining.

Experiment 1

In the experiment that follows, participants were introduced to artificial categories that supported two generalizations about category membership: the salient and reasonably predictive feature of body shape, and the subtle, abstract, but perfectly predictive feature of foot shape. Half the participants were prompted to *explain* while learning, the other half to *describe*. We chose description as a comparison because it mirrors explanations in time, attention, and verbalization, but does not impose the same structural constraints as explanation. If explaining drives participants to interpret observations in terms of general regularities, then participants prompted to explain should be more likely than those who describe to discover the subtle but perfectly predictive rule as a basis for categorization.

Participants and Materials

150 undergraduate students participated for course credit or monetary reimbursement. The task involved *study items*, *test items*, *transfer items*, and *memory items*.

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

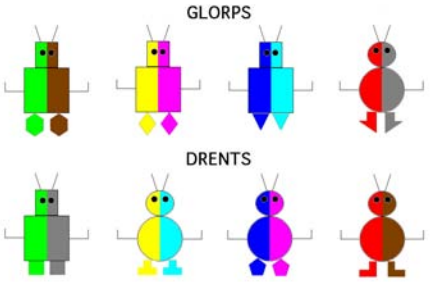

Figure 1: Study items.

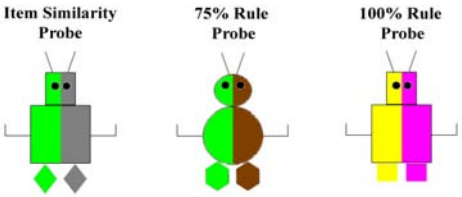

Figure 2: Three types of test items.

This category structure supported at least three distinct bases for categorization. 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, abstract regularity about pointy versus flat feet (called the '100% rule', as it perfectly partitions study items).

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

Transfer Items. These items used completely novel foot shapes to distinguish participants who genuinely drew an abstract generalization concerning "pointy" versus "flat" feet from those who simply recognized the importance of particular foot shapes. For each item, the 100% rule was pitted against item similarity and the 75% rule.

[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 study 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.

Memory Items. Twenty-three robots were presented in a memory test at the end of the experiment. Eight of these were the study items, four were selected from the previously presented test items, and 11 were totally new.

Procedure

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

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

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

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

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

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

Results

Bases for Categorization To understand how explaining influenced what participants learned about categories, we evaluated participants' basis for categorizing novel robots. Explicit reports were coded into four categories (displayed in Table 1A): '100% rule' (explicitly mentioning pointy versus flat feet), '75% rule' (square versus circular body shape), 'item similarity' (reliance on nearest match from study), and 'other'[2]. Responses were coded independently by two coders, with agreement of 86% and differences resolved by discussion.[3] Table 1B reports the analogous *categorization pattern* coding categories: each basis for categorization predicts a particular pattern of responses across the test item probes, so participants were classified as using a basis for categorization if their responses were most consistent with that basis, with ties coded as 'other'.

For both measures, Table 1 suggests that more participants learned and utilized the 100% rule in the *explain* than in the *describe* condition, while more participants drew on the 75% rule in the *describe* than the *explain* condition. For each measure, these suggestive patterns were 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. For both the *explicit response* and *categorization pattern* measures, participants' basis for categorization was more likely to be the 100% rule in the *explain* than the *describe* condition ($\chi^2(1) = 15.89$, $p < 0.001$; $\chi^2(1) = 17.65$, $p < 0.001$), while the 75% rule was more prevalent in the *describe* than the *explain* condition ($\chi^2(1) = 19.56$, $p < 0.001$; $\chi^2(1) = 9.54$, $p < 0.01$). For both measures, 'item similarity' and 'other' responses were not significantly associated with condition.

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

	A ***Explicit response***		**B** ***Categorization pattern***	
Basis	**Exp.**	**Desc.**	**Exp.**	**Desc.**
100% rule	26	6	36	12
75% rule	14	40	17	35
Item Sim.	0	0	4	7
Other	35	29	18	21

Table 1: Bases for categorization, by condition.

Categorization of test and transfer items Participants' categorization responses were scored as accurate if they corresponded to the '100% rule'. Figure 3 shows test and transfer accuracy as a function of condition. A 2 (*task*: explain vs. describe) x 2 (*categorization measure*: test vs. transfer) mixed ANOVA was conducted on categorization accuracy. This revealed a main effect of *task* ($F(1,148) = 16.10$, $p < 0.001$) with participants in the *explain* condition

[2] The "Other" category further consisted of blank, "no difference", 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. When a participant's verbal response or post-experiment debriefing unambiguously indicated a switch in category labels, that participant's categorization responses were reverse coded.

categorizing test and transfer items significantly more accurately than those in the *describe* condition.[4] There was also a significant effect of *categorization measure* ($F(1,148) = 13.46$, $p < 0.001$) as test accuracy was higher than transfer. It is worth noting that the more accurate categorization of transfer items by participants in the *explain* condition ($t(148) = 2.91$, $p < 0.01$) suggests that they not only recognized the importance of foot shape in determining category membership, but abstracted away from the specific shapes used on study items to recognize the subtle property of having 'pointy' or 'flat' feet.

Categorization performance was also analyzed separately for each of the three types of test item (displayed in Fig. 2). Participants categorization of the *100% rule probes* was more consistent with the '100% rule' in the *explain* than the *describe* condition ($t(148) = 4.41$, $p < 0.001$), while categorization of the *75% rule probes* was more consistent with the '75% rule' in the *describe* than the *explain* condition ($t(148) = 3.77$, $p < 0.001$).

Memory for study items Correct identification of original study items (*hits*) was similar across conditions ($t(148) = 0.06$, $p = 0.95$). However, participants in the *explain* condition were less likely to correctly classify novel items as novel (*correct rejections*) than those in the *describe* ($t(148) = 2.12$, $p < 0.05$), suggesting that they were less attentive to the details of items during study (see Fig. 3).

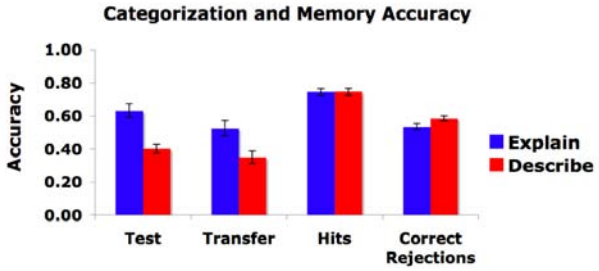

Figure 3: Categorization and Memory accuracy.

Typed Explanations and Descriptions Each of the 8 explanations (or descriptions) a participant provided was coded for whether a feature was mentioned (foot shape, body shape, and color), and if that feature was cited in an abstract or concrete way. References were coded as *concrete* if they cited the actual feature: e.g. triangle/square/L-shaped feet, square/round body, yellow/green color. References were coded as *abstract* if they characterized a feature in more general terms, which could be applied to multiple features: e.g. pointy/flat feet, big/strange body, warm/complementary colors. Figure 4 shows the number of features mentioned in each coding category, as a function of *task*. Two separate 2 (*task*: explain vs. describe) *x* 3 (*feature*: feet vs. body vs. color) ANOVAs were conducted on the total number of *concrete* (*abstract*) features mentioned by each participant. Participants in the *explain* condition cited a greater number of abstract features than those in the *describe* condition (a main effect of *task*, $F(1,148) = 24.72$, $p < 0.001$), while those in the *describe* condition cited more concrete features than those who explained (a main effect of *task*, $F(1,148) = 164.65$, $p < 0.001$). Individual t-tests confirmed that these two findings were reliable for all features (all *p*s < 0.025) except abstract references to body shape ($t(148) = 0.82$, $p = 0.41$).

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

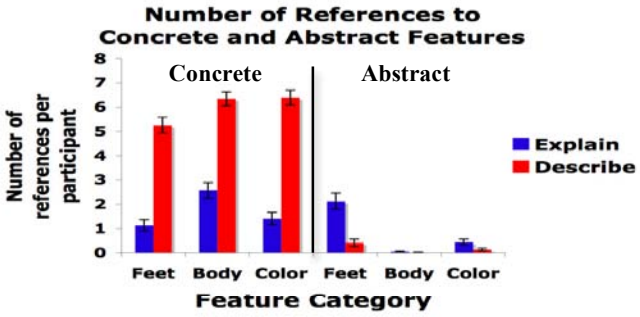

Figure 4: Coding of feature references in participants' explanations and descriptions.

Experiment 2

The first goal of Experiment 2 was to provide a stronger test of the hypothesis that explaining promotes discovery, building on the results of Experiment 1. Firstly, Experiment 2 used 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 from a tendency of description to *inhibit* discovery. Thinking aloud places fewer restrictions than describing on how participants engage with the task, while controlling for having to verbalize. Secondly, participants were explicitly instructed in the introduction phase that they would be later tested on their ability to remember and categorize robots, and reminded of this before the study phase. This manipulation aimed to increase motivation and direct participants' attention to processing items in a way that would be useful for the later tests, reducing the chances that engaging in explanation would have an additional effect on discovery as a result of task demands.

Experiment 2 also aimed to test two hypotheses about the role of explaining in discovery. By virtue of the subsumptive properties of explanations or processing characteristics of explaining, explaining observations may always drive learners to discover underlying regularities that are present. Alternatively, in order to promote discovery, explanation may need to be directed at anomalous

[4] Accuracy near 50% does not reflect chance responding as 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% (100% rule).

observations that demonstrate the inadequacy of current beliefs (or current explanations). For example, these might serve to guide the search for better explanations which capture deeper underlying structure. To investigate this issue the study phase was modified so that learners provided explanations of category membership for only **two** robots: a glorp and drent that were both either *consistent* or inconsistent (*anomalous*) with the '75% rule'. The result was a 2 x 2 between-subjects design with *task* (explain vs. think aloud) crossed with *item type* (consistent vs. anomalous). The materials were the same as in Experiment 1, with minor changes to study items and a new set of memory items.

Procedure Except for the following changes, the procedure was the same as in Experiment 1. (1) 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 reminded of this before explaining (thinking aloud) in the *study phase*. (2) After participants received and viewed the sheet of robots, the introduction phase was augmented by presenting individual study items for study. A block consisted of displaying each of the 8 study items for four seconds with its category label, in a random order. Three blocks were presented, with a clear transition between blocks.

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

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

The test, transfer, and memory phases were identical to Experiment 1.

Results The results reported are for 160 participants (40 per condition), although data collection is ongoing. As the data on explicit responses and categorization patterns generally mirrored that for categorization accuracy, we only present results for categorization. In the interests of space we do not discuss the memory data.

A 2 (*task:* explain vs. describe) x 2 (*item type*: consistent vs. anomalous) x 2 (*categorization measure*: test vs. *transfer*) mixed ANOVA was conducted on categorization accuracy. Fig. 5 shows test and transfer categorization accuracy. There was a significant main effect of *task* ($F(1,156) = 8.33$, $p < 0.005$) with higher accuracy in the explain than think aloud condition, while the effect of *item type* was marginal ($F(1,156) = 2.45$, $p = 0.12$) and the interaction between *task* and *item type* was also marginal ($F(1,156) = 2.54$, $p = 0.11$). There was a significant effect of *categorization measure* ($F(1,156) = 14.38$, $p < 0.001$) with test accuracy higher than transfer, and a significant interaction between *categorization measure* and *item type* ($F(1,156) = 6.33$, $p < 0.05$), with transfer accuracy being particularly high in the anomalous condition. A contrast of the *explain-anomalous* against the *explain-consistent* condition revealed a significant difference ($F(1,156) = 4.99$, $p < 0.05$): explaining anomalous category members promoted accurate generalization significantly more than explaining consistent members.

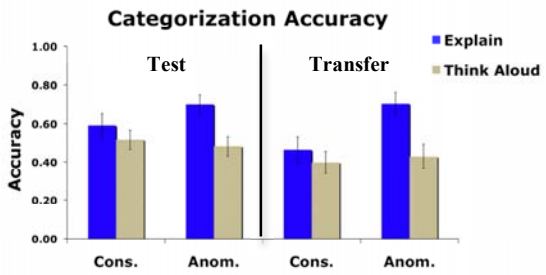

Figure 5: Categorization accuracy (Exp. 2).

Discussion

These findings support our hypothesis that engaging in explanation can facilitate the discovery of regularities underlying category membership. Participants prompted to explain why items belong to particular categories were more likely to induce the abstract generalization ('100% rule') governing category membership than were participants instructed to describe category members (Exp. 1) or to think aloud during study (Exp. 2). Experiment 2 further suggests that anomalous data mediates the relationship between explanation and discovery: neither explaining typical category membership nor studying anomalous members was as effective as the conjunction of these conditions - explaining the membership of anomalous items.

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 explaining category membership should drive participants to discover regularities or patterns. And if explanations are better to the extent they unify a greater number of observations, explaining should drive participants to induce abstract generalizations that are broader and surpass the 75% accuracy afforded by using body shape.

While explaining promoted the discovery of category structure, there was suggestive evidence that participants who explained fared more poorly than those who described on a memory test for studied items. Two possible reasons

for this could be that less attention is available for encoding item details when people are actively trying to generate explanations and/or that highlighting features relevant for category membership (such as foot or body shape) indicates which other features can be safely ignored.

An alternative interpretation of Experiment 1 is that describing inhibits discovery, not that explaining facilitates it. However, the think aloud control condition of Experiment 2 did not require participants to focus on or report item features, yet discovery of the generalization was still higher when participants explained. Another possibility is that prompting participants to explain merely altered the implicit demands of the task. However, in Experiment 2, participants in both conditions were explicitly instructed that they would have to categorize and remember items, and reminded of this goal before the study phase. Taken together, the findings from Experiment 1 and 2 suggest that explaining played a significant role in facilitating learning by promoting discovery of an underlying regularity.

This research has potential implications for both cognitive psychology and education. Our findings suggest that explaining will be most beneficial when learning material or interpreting observations that contain systematic regularities or reflect broad underlying principles. Explaining may be less helpful – or even harmful – in less systematic domains, or when learners' prior knowledge is insufficient to support the induction of relevant principles. An interesting question for future research is whether learners spontaneously explain precisely when it's most likely to be beneficial. Our finding that explaining anomalous information benefits learning more than explaining consistent information (Exp 2) is consistent with this speculation.

The importance of the confluence of explanatory efforts and anomalous observations is also interesting. An important component of effective instruction may be incorporating or making salient those facts and observations that expose the insufficiency of a learner's current knowledge and beliefs, or demonstrate the unsatisfactory nature of current explanations, so that engaging in explanation can drive the induction of generalizations that subsume anomalies. A similar set of considerations may apply in facilitating discovery, construed more broadly.

The memory findings from Experiment 1 suggest that it is not necessarily always productive to engage in explanation: this learning strategy has particular advantages and disadvantages. In many learning contexts encoding facts and details is essential, and may even be necessary to support future learning. Explanation and activities like description may therefore be complementary learning strategies.

Beyond human learning, research on the relationship between explanation, learning and generalization may also inform machine learning, where algorithms involving explanation have been proposed (e.g. Lewis, 1988).

Our experiment is the first (that we know of) to use tools from the cognitive psychology of categorization to examine the effects of explaining on learning, a topic most commonly pursued in educational and developmental psychology. We believe that the integration of these traditions has a great deal of promise. By using artificial categories, we were able to exert more precise control over participants' prior beliefs. And by generating category structures that supported multiple generalizations, we were able to provide a more precise characterization of the role of explanation in the discovery of generalizations. We hope that this experiment contributes to further explorations at the intersection of educational and cognitive psychology.

Acknowledgments

We thank Ania Jaroszewicz, Hava Edelstein, Randi Engle, Nick Gwynne, Cristine Legare, Jared Lorince, David Oh, Luke Rinne, and three anonymous reviewers for helpful feedback. This work was partially supported by the McDonnell Foundation Collaborative Initiative on Causal Learning.

References

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

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.

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.

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

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

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

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.

Strevens, M. (in press). *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: Oxford University Press.