Explanation Recruits Comparison in a Category-Learning Task

Brian J. Edwards[a], Joseph J. Williams[b], Dedre Gentner[a], and Tania Lombrozo[c,d]

[a]Department of Psychology, Northwestern University, Swift Hall 102, 2029 Sheridan Rd., Evanston, IL 60208, USA

[b]Department of Computer Science, University of Toronto, Bahen Centre, 40 St. George St., Room 7224, Toronto, ON M5S 2E4, Canada

[c]Department of Psychology, University of California, Berkeley, 2121 Berkeley Way, Berkeley. CA 94720, USA

[d]Department of Psychology, Peretsman Scully Hall, Princeton University, Princeton, NJ 08540, USA

Corresponding Author: Brian J. Edwards, Museum of the Mind, 111 Lawrence St., Ste. 13F, Brooklyn, NY 11201, USA, Brian.Edwards@mind-museum.org, (646) 623-1366

Running Head: Explanation, Comparison, and Category Learning

Abstract

Generating explanations can be highly effective in promoting category learning; however, the underlying mechanisms are not fully understood. We propose that engaging in explanation can recruit comparison processes, and that this in turn contributes to the effectiveness of explanation in supporting category learning. Three experiments evaluated the interplay between explanation and various comparison strategies in learning artificial categories. In Experiment 1, as expected, prompting participants to explain items' category membership led to (a) higher ratings of self-reported comparison processing and (b) increased likelihood of discovering a rule underlying category membership. Indeed, prompts to explain led to more self-reported comparison than did direct prompts to compare pairs of items. Experiment 2 showed that prompts to compare all members of a particular category ("group comparison") were more effective in supporting rule learning than were pairwise comparison prompts. Experiment 3 found that group comparison (as assessed by self-report) partially mediated the relationship between explanation and category learning. These results suggest that one way in which explanation benefits category learning is by inviting comparisons in the service of identifying broad patterns.

Keywords: Explanation, comparison, categorisation, learning

Explanation Recruits Comparison in a Category-Learning Task

1. Introduction

Explanation and comparison are pervasive in our everyday lives, and they often appear to operate in tandem. For example, asking someone to explain why children prefer chocolate to broccoli will prompt a comparison between chocolate and broccoli. Considering why-questions such as these motivates a search for understanding, and can invite a comparison between two alternatives, even if the alternatives are not explicitly stated (Chin-Parker & Bradner, 2017). For example, if asked "Why do flames burn upward?" one might mentally convert the question into "Why do flames burn upward (rather than downward)?" and consider relevant factors (such as that hot air rises). Indeed, comparison may be a valuable cognitive strategy for producing richer, more accurate, and more satisfying explanations.

Despite these intuitive links between explanation and comparison, researchers have typically studied explanation and comparison separately (for reviews, see Gentner, 2010, on analogy and comparison; Lombrozo, 2012, 2016, on explanation). In this paper, we take steps towards a more integrated approach (see also Chin-Parker & Bradner, 2017; Hummel, Landy, & Devnich, 2008). Broadly, we strive to improve our understanding of why and how engaging in explanation and comparison can enhance learning, and especially how these processes might work together to do so. More specifically, our aim is to explore whether generating explanations in the context of a category-learning task (i.e., explaining why a labelled exemplar belongs to a particular category) encourages learners to engage in comparison. If so, what comparison strategies result, and how do these strategies affect subsequent learning?

We focus on category learning because of its importance in mental life, as reflected by the vast amount of psychological research in this area (e.g., Ashby & Maddox, 2005; Murphy, 2002; Smith & Medin, 1981). Prior research and theory also suggest potentially powerful relationships between explanation and comparison in this domain. For example, comparison is essential for discovering the similarities and differences within and between categories that can form the basis for an explanation of category membership. Additionally, as noted above, the search for an explanation often creates a specific implicit contrast (e.g., why is a tomato a fruit *and not a vegetable*?) (e.g., Chin-Parker & Bradner, 2017), leading people to engage in spontaneous comparison. We are especially interested in understanding how engaging in explanation might influence the specific comparison strategies people use when engaged in a category-learning task.

We begin by reviewing the respective literatures on how explanation and comparison support learning. We then present three experiments that address these central questions in the context of category learning. Finally, we consider whether and how our findings shed light on broader questions about the relationship between explanation and comparison.

1.1. How does engaging in explanation support learning?

For present purposes, we define explaining as the process of answering a why-question to achieve understanding of what is being explained.[1] Prior work shows that explaining, either to others (Roscoe & Chi, 2007, 2008) or to oneself (Chi et al., 1994; for reviews, see Bisra et al., 2018; Fonseca & Chi, 2011), can substantially boost learning. This "self-explanation" effect has

[1] Accounts of explanation typically define *explanation* (the product) rather than *explaining* (the process); for a discussion see Wilkenfeld & Lombrozo (2015). Given that the present paper focuses on the mechanisms by which the *process* of explaining affects learning, and on how this process invokes various forms of comparison processing, a process-focused definition allows us to pursue our central research questions while remaining somewhat agnostic as to the structure of explanations themselves.

been demonstrated across a range of cognitive domains and experimental protocols. For example, generating self-explanations has been shown to improve students' understanding of physics (Chi et al., 1989), biology (Chi et al., 1994), and math (Aleven & Koedinger, 2002; McEldoon, Durkin, & Rittle-Johnson, 2013; Wong, Lawson, & Keeves, 2002; for a review, see Rittle-Johnson, Loehr, & Durkin, 2017). These findings suggest a number of ways in which generating explanations might promote learning (for reviews, see Fonseca & Chi, 2011; Lombrozo, 2006, 2012, 2016).

One way in which explaining can support learning is by increasing metacognitive awareness. Metacognitive processes are essential for enabling people to identify deficiencies in their own knowledge. Explanation can help people detect (Rozenblit & Keil, 2002) and fill gaps in their knowledge (Chi, 2000), as well as resolve inconsistencies (Johnson-Laird, Girotto, & Legrenzi, 2004). Relatedly, the process of generating explanations can increase attention and cognitive engagement (e.g., Siegler, 2002), promoting deeper cognitive processing and improving learning outcomes.

Recent work additionally suggests that engaging in explanation can influence learning by leading people to selectively seek and extend some types of patterns, particularly those with broader scope (i.e., that apply to more cases; Walker et al, 2017; Williams & Lombrozo, 2010, 2013) or that exhibit other "explanatory virtues" (Lombrozo, 2012, 2016; Wilkenfeld & Lombrozo, 2015), such as simplicity (Walker, Bonawitz, & Lombrozo, 2017). As a result, explaining can promote abstraction (Walker & Lombrozo, 2017) and point to inductively rich, causal properties (Legare & Lombrozo, 2014; Walker et al., 2014). Williams and Lombrozo (2010) proposed and tested the "Subsumptive Constraints" account, which predicted that asking people to generate explanations would lead them to focus preferentially on broad patterns that

can account for a greater proportion of the observed evidence. The experimental paradigm used in our studies is based on that developed in Williams and Lombrozo (2010), so we review their studies in some detail.

Across three experiments (Williams & Lombrozo, 2010), adults learned how to categorize robots labeled as Glorp robots or Drent robots. Of the eight study examples, six robots (75%) could be categorized by the 75% body-shape rule that Glorp robots have rectangular bodies and Drent robots have round bodies, but the other two robots were anomalous with respect to this rule (i.e., one Glorp robot had a round body and one Drent robot had a rectangular body). There was also a more subtle 100% foot rule that could perfectly categorize all eight robots. While each of the eight robots had a unique foot shape, all four Glorp robots had feet with pointy bottoms and all four Drent robots had feet with flat bottoms.

Compared to participants in a variety of control conditions, participants who were asked to explain why each robot belonged to its particular category were more likely to discover the 100% foot rule, but less likely than control participants to report discovering the 75% body-shape rule. These results suggest that engaging in explanation has fairly selective effects: it leads people to discover rules that account for all study examples; it does not simply increase the discovery rate of all possible categorization rules (Rehder, 2007; Williams & Lombrozo, 2013). However, it should be noted that favoring broad, exceptionless rules is not always beneficial, and explaining therefore has the potential to hinder learning. When the only way to achieve perfect classification is to memorize the idiosyncratic properties of individuals, seeking explanations can impair learning by encouraging learners to overgeneralize (i.e., disregard exceptions) or perseverate in seeking patterns (Williams, Lombrozo, & Rehder, 2013).

There is also evidence that engaging in explanation can affect learning by recruiting prior knowledge. People often attempt to integrate the phenomenon being explained with their prior beliefs (Chi et al., 1994; Kuhn & Katz, 2009; Lombrozo, 2006), and in so doing to accommodate it within a larger framework (Wellman & Liu, 2007). In particular, there is evidence that people often recruit prior knowledge when trying to explain and understand category structure (Murphy & Medin, 1985; Rips, 1989; Spalding & Murphy, 1996). In a study of the effects of explanation and prior knowledge on category learning, Williams and Lombrozo (2013) presented adults with a similar set of robots that could be classified by two 100% rules, one based on foot shape and one based on the relative length of the antennae. One group of participants was asked to explain why each robot belonged to its particular category and the other group of participants engaged in free study.

Within each group, participants were further divided based on whether the robots were given category labels that were uninformative (Glorp robots vs. Drent robots) versus informative (Indoor robots vs. Outdoor robots). The informative labels were intended to cue prior knowledge relevant to whether the antenna or foot rule was relevant to category membership. For example, robots suited for indoor versus outdoor environments might need different types of feet, while relative antenna length would be less relevant. For participants who were prompted to explain, those receiving the informative category labels were more likely to discover the more subtle foot rule than the antenna rule, while the opposite was true for those presented with the uninformative category labels. In contrast, the type of category label did not affect which rules participants in the free-study condition discovered.

Williams and Lombrozo (2013) argued that explanation invokes prior knowledge in the search for broad patterns, with patterns being judged broader (i.e., more likely to generalize) to

the extent that they conform not only to current evidence, but also to prior beliefs. These findings are also consistent with the proposal that categorization can be construed as inference to the best explanation (Murphy & Medin, 1985; Rips, 1989), where explanations that are consistent with prior beliefs provide "better" explanations—because they have a broader scope, supply a causal mechanism, support a more coherent set of beliefs, or exhibit other virtues (Lombrozo, 2012, 2016).

While these findings help characterize the precise learning consequences of engaging in explanation, many questions remain about the mechanisms by which explanation generates these effects. Lombrozo (2012) hypothesized that engaging in explanation may recruit (or be recruited by) a variety of cognitive processes, including inductive reasoning, deductive reasoning, categorization, causal reasoning, and analogical reasoning. In the present work, we focus on one candidate mechanism: analogical comparison. The very act of explaining could stimulate comparison in the service of fostering the discovery and generalization of broad patterns. This fits with research that has identified abstraction and generalization as two of the principal benefits of comparison (Gentner, 2010). As previously suggested, seeking and evaluating explanations often involves an explicit or implicit contrast between two alternatives: Why is a tomato a fruit (as opposed to a vegetable)? Why is this robot a Glorp (as opposed to a Drent)? (see also Chin-Parker & Bradner, 2017; McGill, 2002; van Fraassen, 1980). The search for explanations—particularly when learning categories from examples—may therefore initiate a process of comparison, which could in part underlie explanation's effects on learning. In the following section, we provide a more detailed review of the evidence that comparison can promote the kind of learning observed in experiments involving explanation.

1.2. How does engaging in comparison support learning?

Comparison is the process of identifying similarities and differences between two cases. Like explanation, engaging in comparison can provide significant learning benefits (e.g., Christie & Gentner, 2010; Gentner, 2003, 2010; Gentner & Namy, 1999; Kotovsky & Gentner, 1996; Loewenstein, Thompson, & Gentner, 2003; Richland, Zur, & Holyoak, 2007; Rittle-Johnson & Star, 2009, 2011; Thompson & Opfer, 2010; for a review, see Alfieri, Nokes-Malach, & Schunn, 2013). One way in which making comparisons can enhance learning is by promoting structural alignment of the cases being compared. In our discussion we use Gentner's (1983, 2010) structure-mapping theory, which describes how analogical comparison can be used to uncover a common relational structure. According to this theory, analogical comparison is geared towards finding a structurally consistent set of one-to-one correspondences that maximizes the common relational structure. The idea is that people implicitly prefer structurally consistent alignments in which lower-order matches are connected by higher-order relational matches (the systematicity principle) (Gentner, 1983; Gentner & Markman, 1997; Gentner, Rattermann, & Forbus, 1993). A computational model of the structure-mapping process, SME, uses a three-stage local-global matching process to arrive at a maximal or near-maximal alignment (Falkenhainer, Forbus, & Gentner, 1989; Forbus et al., 2017). Although there are a number of other computational models of analogy (e.g., LISA: Hummel & Holyoak, 1997, DORA: Doumas, Hummel, & Sandhofer, 2008, DRAMA (Eliasmith & Thagard, 2001), most current models of analogy share structure-mapping theory's core assumption that inferences are based on finding a structurally consistent alignment.

On this view, comparison is an especially powerful learning process because it helps people do more than merely notice feature-level similarities and differences between two items. By highlighting common systems of features, the mapping generated by structural alignment

supports the formation of an abstract relational schema. Thus, comparison plays an important role in the acquisition of abstract knowledge (Gentner & Medina, 1998). This schema can in turn facilitate successful analogical transfer, including far transfer to problems with vastly different surface features or in different cognitive domains (e.g., Catrambone & Holyoak, 1989; Gentner et al., 2009; Gick & Holyoak, 1983; Loewenstein, Thompson & Gentner, 2003). Further, in addition to highlighting the common system, structural alignment also highlights differences that play corresponding roles in the two systems (alignable differences) (Markman & Gentner, 1993; Sagi, Gentner, & Lovett, 2012). Both of these are relevant to category learning.

Notions of similarity have played a prominent role in theories of categorization (Posner & Keele, 1968; Reed, 1972; Rosch & Mervis, 1975; for reviews, see Goldstone, 1994; Sloman & Rips, 1998; Smith & Medin, 1981), and there is considerable evidence that comparison processes are important in category learning (Gentner & Medina, 1998, Markman & Wisniewski, 1997; Spalding & Ross, 1994). Indeed, performing comparisons can help adults learn to categorize birds (Higgins & Ross, 2011) or learn new relational categories (Goldwater & Gentner, 2015; Kurtz, Boukrina, & Gentner, 2013). There is also evidence that comparison processes help young children select taxonomic choices over perceptually similar distractors in a categorization task (Gentner & Namy, 1999; Namy & Gentner, 2002) and learn challenging relational categories (Christie & Gentner, 2010; Gentner, Anggoro, & Klibanoff, 2011; Kotovsky & Gentner, 1996). Interestingly, in some of the developmental studies, children were asked "Do you see why these are both jiggies?"—which could be seen as an invitation to *explain* why these exemplars belong to the "jiggy" category. That this prompt seems to promote both comparison and explanation suggests a close relationship between these processes. Indeed, we hypothesize that when learning

a category, people often invoke comparison processes in the service of generating or evaluating explanations.

1.3. Research examining both explanation and comparison

One way to examine a possible relationship between explanation and comparison is to explore their effects on the same experimental task. Few studies have done so, and when they have, the aim has often been to isolate the effect of each process (that is, to have participants explain only or compare only), rather than consider their potentially interactive effects (Gadgil, Nokes-Malach, & Chi, 2012; Nokes-Malach et al., 2013; Richey, Zepeda, & Nokes-Malach, 2015). For example, Nokes-Malach et al. (2013) evaluated the relative effectiveness of three cognitive strategies for helping college students learn to solve physics problems: reading solutions to worked examples and solving practice problems, explaining the solutions to the worked examples, or comparing and contrasting the worked examples. Participants in the reading and explanation conditions achieved greater near transfer than participants in the comparison condition, while participants in the explanation and comparison conditions achieved greater far transfer than participants in the reading condition.

Similarly, Gadgil, Nokes-Malach, and Chi (2012) investigated the roles of explanation and comparison in acquiring more accurate theories of the circulatory system. They found that comparing an incorrect model of the circulatory system (that was consistent with the participant's prior beliefs) with an expert model of the circulatory system was more effective than explaining the expert model of the circulatory system. These findings suggest that explanation and comparison are both beneficial for learning, but also make it clear that they do not generate equivalent outcomes.

Three additional studies hint at whether and how explanation and comparison might interact. In one study, Kurtz, Miao, and Gentner (2001) presented college students with two superficially dissimilar examples of heat flow. Participants who compared the two scenarios by analyzing the scenarios jointly and listing correspondences between the scenarios later rated the two scenarios as more similar than did both participants who analyzed the two scenarios separately (and who did not list correspondences) and control participants who did not previously analyze the scenarios. Furthermore, in a difference-listing task, participants who had previously engaged in the comparison task (including stating correspondences) listed differences that were more causally relevant to the principle of heat flow than did control participants. This provides evidence that intensive comparison supported the discovery of a common causal system, and that engaging in comparison can help participants identify principles that can serve as a basis for causal explanations.

In another study, Sidney, Hattikudur, and Alibali (2015) gave college students math problems and analyzed the roles of explanation and comparison in students' learning. Most relevant for our purposes, participants who received explanation prompts noticed more similarities and differences between the problems than those who did not—suggesting that the explanation task facilitated comparison processing.

Finally, Hoyos and Gentner (2017) asked whether children's explanations would be influenced by the kinds of comparisons that were readily available. Six-year-old children were asked to explain why a building with a diagonal brace is strong (more precisely, why it is stable, such that its shape cannot be changed without breaking a piece or a joint). They compared the effectiveness of three study-example conditions: (a) a single model building with diagonal braces (single-model condition), (b) a perceptually similar pair of model buildings, one with diagonal

braces and one without such braces (high-alignability condition), or (c) a perceptually dissimilar pair (low-alignability condition). Children were asked "Which building is stronger?" (or in the single model case, "Is this building strong?"). Then they were asked to explain why that building was strong(er). Children in the high-alignability condition were most likely to produce brace-based explanations for the strength of the model with diagonal braces, and were also more likely to succeed on a far-transfer task than children in the low-alignability and single-model conditions. These results show that children can use information acquired through comparison to inform their causal explanations and support transfer to a novel problem.

While these findings reveal that explanation and comparison can work in tandem to support learning, they do not target the central question we explore here: namely whether engaging in explanation recruits comparison, and, if so, which comparison strategies are deployed and how they affect learning. In the next section, we develop a proposal for how explanation and comparison might work together to promote learning in a categorization task.

1.4. Explanation and comparison in a category-learning task

In a basic category-learning task like that in Williams and Lombrozo (2010), participants must learn a classification rule that allows them to successfully classify items into one of two novel categories, and they must do so on the basis of a small number of labelled examples. In this situation, explaining involves answering the question: why does this example belong to this category (as opposed to another)? An answer could identify a classification procedure (e.g., "I know it is an even number, rather than an odd number, because it ends in a '2'"), or it could invoke a deeper basis for category membership (e.g., "it is an even number because it is divisible by 2"). Often, people treat superficial bases for classification as indicative of deeper, category-defining properties (Gelman, 2003), so what look like fairly superficial explanations (e.g., "it is a

Glorp because of its feet") could reflect much deeper explanatory commitments. Bearing this in mind, how might explanation recruit comparison under these conditions?

We propose that when trying to explain why an item belongs to a particular category, people initially invoke comparison through an implicit or explicit contrast: people are likely to think about the problem as a question about why the item belongs to one category and not to another (Chin-Parker & Bradner, 2017; Williams & Lombrozo, 2013). Given that comparison can be essential for explaining an item's category membership, a deeper question than whether explanation recruits comparison is which of several comparison strategies is engaged. One strategy is to compare single pairs of items, either within the same category or across two categories. For example, the person might first carry out pairwise comparisons of individual items within a category to identify similarities, and then between the categories to identify differences. Another strategy is to carry out comparisons of all members of a single category, what we refer to as a *group-level* comparison. For example, a participant might carry out a series of comparisons within each category to identify features that are common (if not universal), and that form the basis for a prototype or some other, more abstract representation.

We hypothesize that such group-level comparisons are likely to be especially helpful in promoting the re-representation of features and exemplars to identify a subtle rule that underlies category membership. For example, a pairwise comparison between Robot A and Robot B in Figure 1 is likely to support the conclusion that Glorp robots differ in their foot shape, and that this is not a basis for categorization. However, a group-level comparison of all four Glorp robots invites a re-representation of foot shape at a more abstract level—that despite surface-level differences, all these robots have feet with pointy bottoms. A comparison with the four Drent robots, all of which have feet with flat bottoms, suggests that foot shape is indeed diagnostic of

category membership. (See Forbus et al., 2017; Kuehne et al., 2000; for a computational model of iterative alignment and abstraction.)

If this proposal is correct, then we should expect that prompting learners to explain will lead them to engage in more comparison processing, and that the strategy they pursue (i.e., pairwise or group-level comparisons) might affect what they ultimately learn. Our experiments are correspondingly designed to address the following questions. (1) Does explanation increase the extent to which participants engage in comparison? (2) If so, what comparison strategies does explanation recruit? (3) Do these comparison strategies contribute to the effectiveness of explanation in this task?

1.5. Overview of experiments

To answer these questions, we adapted the category-learning task from Williams and Lombrozo (2010; 2013). In the present experiments, some participants received prompts to explain and some received prompts to compare. We varied the nature of these prompts to consider both within- and between-category comparisons, as well as pairwise versus group-level comparisons. In order to assess whether participants used comparison when generating explanations, we asked participants to report the extent to which they engaged in comparison, as well as the extent to which they engaged in explanation. Across experiments, these self-reports[2] enabled us to gain insight into both the extent and nature of the specific comparison strategies (i.e., within-category, between-category, and group comparisons) that participants in each

[2] A further motivation for collecting self-reports, particularly for comparison processing, was that, as can be seen in Figure 1, the robots have a high degree of alignable surface similarity. Because of the high similarity of the stimuli and because all eight robots were displayed on-screen simultaneously, we were concerned that participants would engage in comparison spontaneously, even in the control condition. To preview the results, this concern turned out to be correct.

condition were performing, including how these different comparison strategies relate to learning.

In each experiment, we examined the effects of our experimental manipulations on category learning, as well as (1) whether instructions to generate explanations would lead participants to engage in comparison and (2) whether comparison (either directly prompted or as assessed by self-report) would promote category learning. Across experiments we varied the nature of the comparisons that were prompted and assessed, with Experiment 1 focusing on comparisons of pairs of robots within the same category, and Experiments 2 and 3 additionally focusing on group-level comparisons. We predicted that prompts to explain or to compare would be beneficial (replicating prior research), that they would lead to more self-reported comparison, and that comparison would in turn be associated with positive learning outcomes. To foreshadow our results, we found that our predictions were confirmed, but only for group-level comparisons. Pairwise comparisons of individual pairs of robots (as we explored in Experiment 1) were not associated with positive learning outcomes.

2. Experiment 1

Experiment 1 evaluated the effects of prompts to explain and of prompts to compare on how people learn novel categories from examples, with a focus on the comparison of pairs of individual robots within the same category. Additionally, we included measures of the extent to which participants actually engaged in explanation and comparison in response to each prompt, enabling us to see whether engaging in explanation would recruit comparison processing, and whether these forms of processing were correlated with learning in our task.

Participants either received explanation prompts or did not, and received comparison prompts or did not, resulting in four possible study conditions. The explanation prompts involved explaining the category membership of individual exemplars. The comparison prompts involved comparisons between pairs of robots within the same category. This configuration of conditions allowed us to investigate the effects of prompts to elicit each strategy—explanation and comparison—relative to each other, in conjunction, and relative to the absence of either prompt, which served as a control condition.

2.1. Method

2.1.1. Participants

Participants were 157 adults recruited from the website Amazon Mechanical Turk, which has been used by a wide range of psychological and other behavioral science research (e.g., Ahn & Yo, 2011). For all experiments, we restricted participation to IP addresses within the U.S., and to people with a task approval rating of at least 95%. Participants received a small amount of monetary compensation. An additional 60 participants were tested, but excluded from the analyses because they failed a "catch trial," had previously completed a similar experiment, or because of a duplicate or missing IP address, which suggested possible repeat participation. This exclusion rate is typical of psychology studies conducted on Mechanical Turk (Ahn & Yo, 2011). In all experiments, the proportion of participants excluded did not vary across conditions.

2.1.2. Materials

The stimuli were eight robots adapted from stimuli used by Williams and Lombrozo (2010, 2013). Four robots (A-D) were labeled "Glorp robots" and four robots (E-H) were labeled "Drent robots." The stimuli are shown in Figure 1.

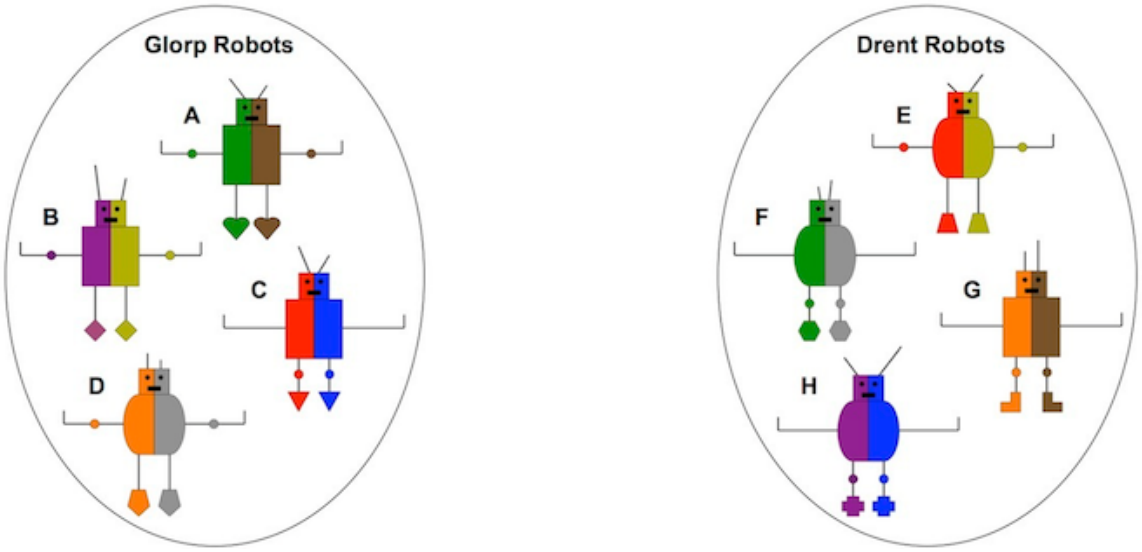

Figure 1: Robot stimuli used in Experiments 1-3.

There were four rules that could be used to categorize robots as either Glorp robots or Drent robots. Two rules were labeled "100% rules" because they successfully differentiated all eight study robots. The other two rules were labeled "75% rules" because they successfully differentiated six out of the eight (or 75%) of the study robots, with two robots (one Glorp and one Drent) anomalous with respect to each 75% rule. The four rules were as follows:

(1) ***Foot rule (100%).*** All Glorp robots have feet with pointy bottoms and all Drent robots have feet with flat bottoms.

(2) ***Antenna rule (100%).*** All Glorp robots have a right antenna that is taller than the left antenna and all Drent robots have a left antenna that is taller than the right antenna.

(3) ***Body-shape rule (75%).*** Glorp robots have rectangular bodies and Drent robots have round bodies. (Glorp robot D and Drent robot G are exceptions.)

(4) ***Elbows/knees rule (75%):*** Glorp robots have elbows (but no knees), and Drent robots have knees but no elbows. (Glorp robot C and Drent robot E are exceptions.)

While the eight robots had different color patterns, these did not vary systematically across categories.

2.1.3. Procedure

The procedure consisted of a study phase followed by a rule-reporting phase, self-report questions, and end-of-study questions. At the beginning of the study phase, participants were told that they would study a set of robots and then answer questions about how to decide whether robots are Glorp robots or Drent robots. Each participant was randomly assigned to study the robots in one of four ways, based on a 2 × 2 design: prompts to explain (yes/no) × prompts to compare (yes/no). The total study time in each condition was 640 seconds (80 seconds for each of the eight robots).

In all conditions, the picture of all eight robots including the category labels (Figure 1) was visible during the entire study phase. Below the picture, there was a study prompt followed by a response text box. As described below, participants were presented with a series of study prompts, displayed one at a time, which varied across conditions. Each prompt asked participants to study one (control and explanation prompts) or two (comparison prompts) of the eight robots displayed in the picture. Aside from the study prompt, participants did not receive specific instructions about how they should structure their responses. The study prompts and procedures for each condition were as follows:

Comparison prompts only condition. Participants were given prompts of the form "What are the *similarities and differences* between Glorp [Drent] robot X and Glorp [Drent] robot Y?" Participants were given 160 seconds to respond to each prompt and could not advance until the time had elapsed. After 160 seconds, participants automatically advanced to the next comparison pair. The order of the comparison pairs was as follows: A and B, F and H, C and D, and E and G.

This order was selected so that participants would compare the robots that were consistent with respect to both 75% rules before comparing the robots that were anomalous with respect to one of these rules. In this experiment, all comparison prompts were within-category; this was varied in subsequent studies.

Explanation prompts only condition. Participants were given prompts of the form "Try to explain *why* robot X is a Glorp [Drent] robot." Participants were given 80 seconds to respond to each prompt and could not advance until the time had elapsed. After 80 seconds, participants automatically advanced to study the next robot. The order in which participants studied the robots matched that of the comparison condition (A, B, F, H, C, D, E, G), except that participants studied the robots one at a time.

Both comparison and explanation prompts condition. Participants studied the robots by responding to both the comparison and explanation prompts. For each pair of robots, participants were given both types of prompts (e.g., explain A, explain B, compare A and B) before moving on to the next pair. The order of the comparison and explanation prompts was counterbalanced across participants; however, each participant always performed the explanation task before the comparison task, or vice versa. To match the conditions for total study time, the duration of each comparison prompt was 80 seconds and the duration of each explanation prompt was 40 seconds. The study order was the same as in the other conditions.

Control condition. Participants were given prompts of the form "*Write out your thoughts* below as you learn to categorize Glorp [Drent] robot X." This prompt was intended to engage participants in actively studying the robots and providing a written response, but without explicitly asking them to make comparisons or generate explanations. Participants studied each

robot for 80 seconds, and as in the other conditions, could not move on to the next robot until this time had elapsed. The study order was the same as in the explanation condition.

To ensure that participants' attention was not diverted to other tasks while studying and answering prompts, participants in all conditions received simple math questions (e.g., 9 + 7) to solve in between studying category items; one question appeared after each 160 seconds of study (four questions total). Participants who took longer than 60 seconds to answer or who answered one or more questions incorrectly were excluded from the analyses, as they were not likely to be paying attention. We refer to these as the "catch trials."

After studying the robots in one of these four ways, participants advanced to the rule-reporting phase. At the beginning of the rule-reporting phase, participants were told, "We're interested in any patterns that you noticed that might help differentiate Glorps and Drents. Report any patterns that you noticed, even if they weren't perfect and even if you don't think they're important." This language was chosen to maximize reporting of both the 75% and 100% rules. Participants then reported each rule they discovered one at a time by entering a description of the rule in a text box.[3]

Participants' responses were evaluated by a coder who was masked to the experimental condition. The coder determined which of the four rules each participant discovered, choosing "yes" or "no" for each rule. Additionally, 25% of the data were coded independently by a second masked coder. In all three experiments, inter-coder reliability was greater than 95%.

[3] Participants also answered two questions about each reported rule. First, participants reported how many of the eight study robots could be categorized using that rule. Second, participants answered a rule-generalization question: "Out of 100 new Glorp and Drent robots from planet ZARN, how many *new* robots do you think you could accurately categorize as either Glorps or Drents *using only the pattern you just described*?" Because this information was only collected when participants reported discovering a categorization rule, the resulting sample sizes were small and these data were not analyzed.

After completing the rule-reporting phase, participants answered self-report questions about the extent to which they engaged in explanation and in comparison. Specifically, we asked: (1) "Regardless of the task instructions, did you notice yourself explaining what makes particular robots Glorp robots or Drent robots when the image of the eight robots was on-screen?," and (2) "Regardless of the task instructions, did you notice yourself making comparisons between pairs of Glorp robots and pairs of Drent robots when the image of the eight robots was on-screen?" Participants answered each of these questions on a 1-7 scale with one-point intervals, where 1 = "not at all," 4 = "some of the time," and 7 = "all of the time."

The self-report questions had three purposes. First, they allowed us to test the prediction that explanation prompts would invoke comparison processing. Second, they allowed us to evaluate the effects of explanation *processing* and comparison *processing* on category learning (over and above whether participants had received explanation or comparison *prompts*). Third, they served as manipulation checks, allowing us to ask whether prompts to explain or to compare successfully evoked the corresponding process.

After completing the self-report questions, participants advanced to the end-of-study questions. They reported their age and gender and whether they had previously completed a similar experiment, and they answered an additional "catch trial" question (adapted from Oppenheimer, Meyvisb, & Davidenkoc, 2009) to see whether they were reading the instructions carefully. Participants who reported having previously completed a similar experiment or who failed a catch trial were excluded from the analyses.

2.2. Results

To evaluate the effects of the explanation and comparison prompts on categorization rule discovery, we analyzed the following across conditions: (1) self-reported explanation and

comparison processing, (2) discovery of at least one 100% rule, and (3) discovery of at least one 75% rule. The results are shown in Table 1 and Figure 2. We focus on discovery of at least one rule of each type (rather than the number of such rules discovered) because once participants discovered a categorization rule that they judged adequate, they tended not to search for additional rules (see also Williams & Lombrozo, 2013). In the Supplementary Materials, we present tables that show the percentage of participants who discovered each combination of rules (e.g., one 100% rule and two 75% rules) for each experiment.

In the following analyses, we treated the four study conditions as comprising a 2 × 2 design: participants received or did not receive the explanation prompts, and independently, received or did not receive the comparison prompts. For all analyses, we report all significant effects, as well as non-significant effects of particular interest.

2.2.1. Self-reports

Table 1 reports the average levels of self-reported explanation and comparison across study conditions. To analyze effects of study condition on self-reported processing, we conducted a pair of ANOVAs. First, we performed a 2 × 2 ANOVA with explanation prompts (yes vs. no) and comparison prompts (yes vs. no) as independent variables and the amount of self-reported explanation as the dependent variable. Participants who received the explanation prompts reported doing more explanation than participants who did not receive the explanation prompts, $F(1, 150) = 26.9$, $p < .001$, $\eta^2 = 0.152$, as intended. In contrast, participants who received the comparison prompts reported doing less explanation than participants who did not receive comparison prompts, $F(1, 150) = 7.15$, $p = .008$, $\eta^2 = 0.046$.

Consistent with the prediction that explanation-generation promotes comparison processing, an equivalent ANOVA found that participants who received explanation prompts

reported doing more comparison than those who did not receive these prompts, $F(1, 151) = 9.34$, $p = .003$, $\eta^2 = 0.058$. However, receiving the comparison prompts did not have a significant effect on the amount of reported comparison processing, $F(1, 151) = 0.17$, $p = .68$, $\eta^2 = 0.001$. Indeed, the amount of self-reported comparison was significantly higher for participants given only the explanation prompts than for those given only the comparison prompts, $F(1, 77) = 5.58$, $p = .021$, $\eta^2 = 0.068$. The relative ineffectiveness of the comparison prompts may have stemmed in part from a high rate of spontaneous comparison, given that the stimuli were highly similar and alignable. Much evidence shows that pairs that are high in overall similarity are likely to be spontaneously compared and are easy to align (Gentner & Toupin, 1987; Sagi, Gentner & Lovett, 2012). In addition, the classification task itself may itself have promoted comparison.

Table 1: Self-reported explanation and comparison in each study condition in Experiment 1. Ratings were made on a 1-7 scale, with higher numbers indicating higher ratings for the amount of explanation / comparison.

Study Condition	**Self-reported Explanation Mean (SD)**	**Self-reported Comparison Mean (SD)**
Control	4.92 (1.86)	4.95 (1.97)
Comparison Only	3.87 (2.18)	4.47 (2.13)
Explanation Only	6.10 (1.32)	5.51 (1.78)
Both Comparison and Explanation	5.63 (1.94)	5.74 (1.91)

2.2.2. Responses to study prompts

Participants' freeform responses to the explanation, comparison, and control prompts varied across many dimensions and were often ambiguous. Therefore, with one exception mentioned in the Experiment 1 Discussion, we did not perform a rigorous coding of these responses. Sample responses to each type of prompt are shown in Table 2.

Table 2: Sample responses to each type of study prompt.

Prompt Type	Sample Responses
Comparison	1. F and H have the same shape but their antennae are different lengths and F has them straight up while H has them sticking out in an angle. Their colors are different. The shapes on their feet are different. 2. Glorp Robots have longer antenna that lean to the left. Drent robots have longer antenna that lean to the right 3. both have pointed feet, both have longer right antennae, both are split in color in the middle, they have different shaped feet, different colors, different body shapes, D has elbows and no knees, C has knees and no elbows, c's antenna are angled, D's are straight up
Control	1. Robot B is purple and yellow. It is square in shape and has diamond shaped feet. It has nothing on its legs. Its left antennae is longer than the right. It has two balls on its arms. 2. Green and gray, 6 sided feet, left antennae longer than right, both pretty short compared to others, round body as most Drents 3.Most Glorps have elbows whereas most Drents have knees.
Explanation	1. Robot C is a Glorp robot because it has pointed feet. 2. Robot C is a Glorp robot because it has a longer antenna on the right side of its head. All Glorp robots have a longer antenna on the right side of their heads. Drent Robots have a longer antenna on the left side of their heads. 3. Robot E is a Drent robot for two reasons, first it has flat bottomed feet like other Drents but unlike Gorps. Secondly, its right side antenna is shorter than its left side antenna.

2.2.3. Discovery of one or more 100% rules

A log-linear analysis of explanation prompts × comparison prompts × discovered at least one 100% rule found that participants who received explanation prompts were more likely to discover at least one 100% rule, $\chi^2(1) = 21.3$, $p < .001$, than those who did not. However, receiving the comparison prompts made participants less likely to discover at least one 100% rule, $\chi^2(1) = 5.48$, $p = .019$ (see Figure 2A).

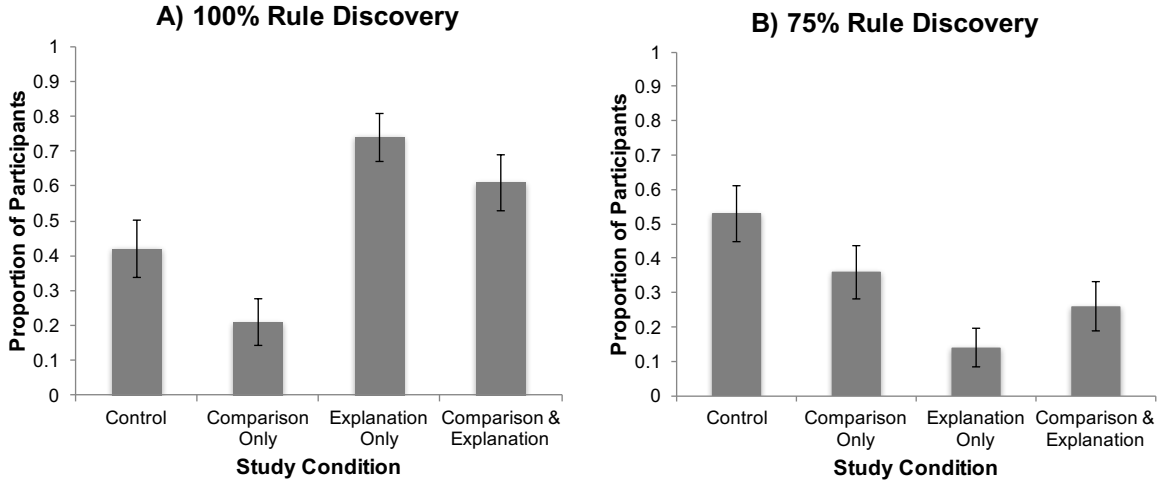

Figure 2: Proportion of participants in each study condition who discovered at least one 100% rule (Fig. 2A) and at least one 75% rule (Fig. 2B) in Experiment 1. Error bars indicate +/- 1 SE.

A simultaneous multiple logistic regression on the amount of self-reported explanation and the amount of self-reported comparison found a significant positive effect of self-reported explanation on discovery of at least one 100% rule, above and beyond self-reported comparison, $W(1) = 7.61$, $p = .006$, $\beta = .332$, $\text{Exp}(\beta) = 1.394$ (constant-only model: $W(1) = 0.059$, $p = .81$, $\beta = -.39$, $\text{Exp}(\beta) = 0.962$). However, there was not a significant effect of self-reported comparison above and beyond that of self-reported explanation, $W(1) = 0.24$, $p = .62$, $\beta = .059$, $\text{Exp}(\beta) = 1.060$. This pattern is consistent with our experimental results and helps validate our self-report

measures: while explanation (prompted or self-reported) was associated with 100% rule discover, pairwise comparison (prompted or self-reported) was not.[4]

2.2.4. Discovery of one or more 75% rules

An equivalent log-linear analysis for whether participants discovered at least one 75% rule found that the explanation prompts made participants less likely to discover at least one 75% rule, $\chi^2(1) = 10.8$, $p = .001$. However, there was no significant effect of whether participants were given the comparison prompts on 75% rule discovery, $\chi^2(1) = 0.09$, $p = .77$ (see Figure 2B). Additionally, participants who reported at least one 100% rule were significantly less likely to report a 75% rule relative to participants who did not report a 100% rule, Fisher's Exact Test: $p = .026$.

2.3. Discussion

The results support one of our key predictions, that instructions to explain category membership would lead participants to engage in comparison. Instructions to explain also led to more explanation. However, instructions to compare category members did not lead to more comparison. We discuss this further below.

Consistent with previous work on explanation in category learning (Williams & Lombrozo, 2010), we found that engaging in explanation improved participants' ability to discover at least one 100% rule, but decreased reporting of 75% rules. Although it is possible that engaging in explanation impaired participants' ability to discover the 75% rules, it is also

[4] We also performed separate logistic regressions on the amount of self-reported explanation and the amount of self-reported comparison. These analyses revealed significant positive effects of both self-reported explanation, $W(1) = 15.0$, $p < .001$, $\beta = .374$, $\text{Exp}(\beta) = 1.454$, and self-reported comparison, $W(1) = 7.28$, $p = .007$, $\beta = .243$, $\text{Exp}(\beta) = 1.275$, on discovery of at least one 100% rule. The latter result is consistent with our intuition that comparison *processing*, potentially in the service of generating explanations, aids in the discovery of a 100% rule.

possible that many participants noticed these rules but did not report them because they judged them inadequate, or had already discovered a "better" 100% rule.

Although we had expected that responding to the comparison prompts would also make participants more likely to discover at least one 100% rule, the comparison prompts seemed to impair performance. These data are surprising given the extensive literature showing that comparison has robust positive effects on category learning. We attribute this result in part to a failure of the intended manipulation: the self-report measures suggested that participants who were prompted to compare engaged in no more comparison, and in less explanation, than those in other conditions. As to why the comparison prompts were so ineffective, we have already noted that the high perceptual similarity and alignability of all the items may have rendered the instruction to compare rather superfluous. Indeed, the comparison group did not differ from the control group in self-reported comparison processing.

Another possible contributor is that the comparison prompts in this study directed participants to compare pairs of robots from the same category. It is possible that participants in the explanation condition performed a broader range of comparisons. In particular, they may have carried out more between-category comparisons than did those in the comparison condition. Indeed, evidence from studies by Higgins and Ross (2011) suggests that between-category comparisons may be more effective than within-category comparisons for learning categories whose members are highly alignable within-category, as in the present case. Finally, those in the explanation condition may also have benefited from the self-directed nature of their own comparisons.

We explored these possibilities in two supplementary experiments, reported in detail in the Supplementary Materials. In brief, Supplemental Experiment A manipulated whether

participants were prompted to perform within-category comparisons versus between-category comparisons (see also Experiment 2 of Edwards, Williams, & Lombrozo, 2013); this did not affect the rate at which participants discovered at least one 100% rule. In Supplemental Experiment B, we asked whether comparison prompts would be more effective if participants were allowed to choose which pairs of robots to compare (e.g., Markant & Gureckis, 2014). One group of participants was assigned to compare specific pairs of robots, whereas a second group of participants was allowed to choose which pairs of robots to compare. These two groups of participants did not differ in their performance on the rule discovery task.

The findings from our Supplementary Experiments led us to explore another hypothesis for why pairwise comparisons may have been ineffective. As we have already suggested, discovering the 100% rule may have required participants to attend to group-level properties in order to arrive at an appropriate representation of the features (e.g., as pointy feet rather than 'feet that are either triangular, heart-shaped, wedge-shaped, or diamond-shaped') and in order to extract maximally diagnostic properties. Simply comparing pairs of robots, either within or between categories, would be insufficient to achieve this, without some further integration across both within- and between-category comparisons. Interestingly, of the 39 Experiment 1 participants who received only pairwise comparison prompts, 19 participants (49%) performed spontaneous group-level comparisons when responding to the first comparison prompt.[5] Accordingly, in Experiments 2 and 3, we explored the effects of two comparison strategies:

[5] Participants were coded as having engaged in group comparison if they either described a category-wide pattern (e.g., "most Glorp robots have balls on their arms") or described a category-level difference between Glorp and Drent robots (e.g., "Glorp robots are the only ones to have pointed feet"). Of the participants who engaged in spontaneous group comparison, 26% (5 of 19) discovered at least one 100% rule, compared to 15% (3 of 20) of participants who did not engage in spontaneous group comparison. This trend was not statistically significant, Fisher's Exact Text: $p = .45$, perhaps due to the small sample size.

engaging in both within- and between-category comparisons, and engaging in group-level comparisons.

3. Experiment 2

The results of Experiment 1 supported our prediction that instructions to generate explanations would lead participants to engage in comparison (as well as in explanation). We also found, as expected, that instructions to explain would lead to better category learning. However, contrary to expectation, we did not find evidence that prompting pairwise comparisons of robots within the same category improved learning. In Experiment 2, we revisited our predictions by considering a broader range of comparison strategies. Specifically, we examined whether (1) prompting both between-category and within-category pairwise comparison (i.e., comparisons between pairs of robots both across different categories and within the same category) would lead to better performance than prompting only within-category pairwise comparison, and (2) whether prompting group comparison (i.e., prompting participants to compare all four robots within each category) would increase categorization rule discovery compared to prompting pairwise comparison. Perhaps the emphasis on a single kind of pairwise comparison in our instructions led participants to "lose the forest for the trees"—that is, to attend to local pairs rather than thinking about the global category structure. Thus, we hypothesized that relative to prompting pairwise comparison, prompting group comparison might place greater emphasis on the overall statistical structure of the categories (versus pairwise relationships), and

thus, improve participants' ability to effectively represent features and discover the global rules underlying category membership.[6]

3.1. Method

3.1.1. Participants

Participants were 497 adults recruited from Amazon Mechanical Turk and tested online. An additional 146 participants were tested, but excluded from analyses. The exclusion criteria were the same as in Experiment 1.

3.1.2. Materials

The stimuli were the eight robots used in Experiment 1.

3.1.3. Procedure

The procedure consisted of a study phase, a rule-reporting phase, self-report questions, and end-of-study questions. In the study phase, each participant was randomly assigned to one of eight study conditions based on a 2 × 2 × 2 design in which participants were prompted to do explanation or comparison, within-category study only or both within-category and between-category study, and individual/pairwise study or group study. The eight resulting conditions are described in Table 3. The total study time, 360 seconds, was matched across all conditions. As in Experiment 1, the picture of the eight study robots was on-screen for the duration of the study phase. Additionally, as in Experiment 1, participants solved simple math problems as catch trials. Each problem was presented between prompts and after each 180 seconds of study.

[6] In an experiment published in conference proceedings, Experiment 3 of Edwards, Williams, and Lombrozo (2013), participants who were prompted to do group comparison were more likely to discover at least one 100% rule than participants who were prompted to do pairwise comparison. Here we present a more systematic investigation of group comparison.

Table 3: Summary of the eight conditions in Experiment 2. Comparison prompts either asked participants to compare specific pairs (pairwise study), or to compare entire groups (group study).

Condition	Study Prompt	Study Order	Prompt Duration	Notes
Explanation, Within-Category—**Individual Study**	"Try to *explain why* robot X is a Glorp [Drent] robot."	A, B, F, H, C, D, E, G	45 s	
Explanation, Within-Category—**Group Study**	"Try to *explain why* robots A-D [E-H] are Glorp [Drent] robots."	A-D, E-H	180 s	
Explanation, Within- and Between-Category—**Individual Study**	"Try to *explain why* robot X is a Glorp [Drent] robot."	A, B, F, H, C, G, D, E or A, H, B, F, C, D, E, G	45 s	Study order counterbalanced across participants
Explanation, Within- and Between-Category—**Group Study**	Within category: "Try to *explain why* robots A-D [E-H] are Glorp [Drent] robots." Between category: "Try to *explain why* robots A-D are Glorp robots and why robots E-H are Drent robots."	A-D, E-H, A-H	90 s (A-D and E-H), 180 s (A-H)	Study position of A-H (first or last) counterbalanced across participants.
Comparison, Within-Category—**Pairwise Study**	"*Compare* Glorp [Drent] robot X and Glorp [Drent] robot Y (i.e., what are the *similarities and differences* between these robots?)."	A and B, F and H, C and D, E and G	45 s	After participants compared all four pairs of robots, participants repeated the same four comparisons in the same order.
Comparison, Within-Category—**Group Study**	"*Compare* the Glorp [Drent] robots (robots A-D [robots E-H]) (i.e., what are the *similarities and differences* between these robots?)."	A-D, E-H	180 s	
Comparison, Within- and Between-Category—**Pairwise Study**	Within category: "*Compare* Glorp [Drent] robot X and Glorp [Drent] robot Y (i.e., what are the *similarities and differences* between these robots?). Between category: "*Compare* Glorp robot X and Drent robot Y (i.e., what are the *similarities and differences* between these robots?).	A and B, F and H, C and D, E and G, A and H, B and F, C and D, E and G	45 s	
Comparison, Within- and Between-Category—**Group Study**	Within category: "*Compare* the Glorp [Drent] robots (robots A-D [robots E-H]) (i.e., what are the *similarities and differences* between these robots?)." Between category: "*Compare* the Glorp robots (robots A-D) and the Drent robots (robots E-H) (i.e., what are the *similarities and differences* between these robots?)."	A-D, E-H, A-H	90 s (A-D and E-H), 180 s (A-H)	Study position of A-H (first or last) counterbalanced across participants.

After completing the study phase, participants completed the rule-reporting phase, followed by the self-report phase and then by a series of end-of-study questions. The rule-reporting phase was the same as in Experiment 1, but the wording of the self-report questions was revised to focus more specifically on the extent to which participants engaged in particular processes while studying the robots. In Experiment 2, we asked: "Whether or not the instructions specifically asked you to do so, to what extent did you engage in the following activities?" This question was followed by three sub-items: (1) "*Explaining* why particular robots are Glorp robots or Drent robots," (2) "*Comparing* pairs of robots from the *same* category (i.e., noting similarities and differences between them)," and (3) "*Comparing* pairs of robots from *different* categories (i.e., noting similarities and differences between them)." As in the previous experiments, participants provided ratings on a 1-7 scale. The remaining end-of-study questions were identical to those used in Experiment 1.

3.2. Results

3.2.1. Self-reports

We first performed a series of ANOVAs to analyze the self-report data (see Table 4). To analyze effects of the study conditions on self-reported explanation, we performed a 2 × 2 × 2 ANOVA with explanation vs. comparison, group study vs. individual/pairwise study, and within-category study only vs. both within- and between-category study as between-subjects factors and the amount of reported explanation as the dependent variable. This analysis found that participants in the explanation conditions reported doing significantly more explanation than participants in the comparison conditions, $F(1, 473) = 68.1$, $p < .001$, $\eta^2 = 0.126$. There were also three significant interactions: between explanation versus comparison and within-category

study only versus both within-and-between category study, $F(1, 473) = 4.69$, $p = .031$, $\eta^2 = 0.010$, between explanation versus comparison and individual/pairwise versus group study, $F(1, 473) = 11.4$, $p = .001$, $\eta^2 = 0.023$, and between within-category study only versus both within-and-between category study and individual/pairwise versus group study, $F(1, 473) = 4.57$, $p = .033$, $\eta^2 = 0.010$. As these interactions were not central to our predictions, we do not pursue them further.

Turning to one of our key predictions, that instructions to explain would foster comparison processing, we next analyzed whether the amount of reported comparison differed across conditions. We calculated the total amount of comparison that each participant reported performing by adding the numerical scores for self-reported within-category comparison and self-reported between-category comparison. Here and elsewhere, we report only the data for the total amount of comparison, except where the data for within- and between-category comparison exhibited distinct patterns. Consistent with the findings from Experiment 1, an equivalent ANOVA found that participants in the explanation conditions reported more total comparison than participants in the comparison conditions, $F(1, 481) = 4.13$, $p = .043$, $\eta^2 = 0.009$. As in Experiment 1, instructions to explain resulted in more comparison processing, as well as more explanation processing, than did instructions to compare.

Table 4: Self-reported explanation and comparison in each study condition in Experiment 2.

Study Condition	Self-reported Explanation Mean (SD)	Self-reported Comparison (Within-Category) Mean (SD)	Self-reported Comparison (Between-Category) Mean (SD)
Individual Explanation (Within)	5.89 (1.52)	5.62 (1.82)	5.72 (1.75)
Group Explanation (Within)	4.91 (1.90)	5.39 (1.83)	5.40 (1.72)
Individual Explanation (Within + Between)	5.20 (1.54)	5.32 (1.65)	5.43 (1.64)
Group Explanation (Within + Between)	4.92 (1.86)	5.60 (1.67)	5.88 (1.46)
Pairwise Comparison (Within)	3.49 (2.16)	5.86 (1.33)	4.43 (2.22)
Group Comparison (Within)	3.65 (2.16)	5.81 (1.18)	5.28 (1.87)
Pairwise Comparison (Within + Between)	3.51 (1.96)	5.02 (1.48)	5.27 (1.40)
Group Comparison (Within + Between)	4.47 (2.15)	5.17 (1.64)	5.44 (1.51)

3.2.2. Discovery of one or more 100% rules

We analyzed whether the proportion of participants who discovered at least one 100% rule varied across study conditions by performing a log-linear analysis of explanation vs. comparison prompts × group vs. individual/pairwise study × within-category only vs. both within- and between-category study × discovered vs. did not discover a 100% rule (see Figure 3A). This analysis revealed that participants in the explanation conditions were significantly more likely to discover a 100% rule than participants in the comparison conditions, $\chi^2(1) = 27.8$, $p < .001$, and that participants given group study instructions were significantly more likely to discover a 100% rule than participants given individual/pairwise instructions, $\chi^2(1) = 4.49$, $p = .034$. Additionally, there was a significant interaction between being given explanation prompts versus comparison prompts and group study versus individual/pairwise study, $\chi^2(1) = 4.56$, $p = .033$.

To better understand this interaction, we conducted separate log-linear analyses for participants who received the explanation and comparison prompts. Of participants in a comparison condition, participants who were prompted to do group comparison were significantly more likely to discover a 100% rule than participants who were prompted to do pairwise comparison, $\chi^2(1) = 9.06$, $p = .003$. This suggests that in our task, prompting group comparison is a more effective way to promote category learning than prompting pairwise comparison.

For participants in the explanation conditions, an equivalent log-linear analysis did not find a significant effect of group study instructions vs. individual study instructions on 100% rule discovery, $\chi^2(1) = 0.06$, $p = .80$, suggesting that the benefit of group study instructions was specific to participants in a comparison condition. Group study prompts may have made

comparison participants more sensitive to the statistical structure of the categories, making participants more likely to search for and discover category-wide patterns. Additionally, among participants who received group study prompts, a log-linear analysis of explanation vs. comparison prompts × discovered vs. did not discover a 100% rule found an advantage for group explanation; group explanation participants were significantly more likely to discover a 100% rule than group comparison participants, $\chi^2(1) = 5.65$, $p = .017$.

So far, we have focused on the effects of being instructed to explain vs. compare (i.e., prompt type) and of whether participants were instructed to process all members of a category, pairs of robots, or individual robots. Next, we analyzed the effects of the kind of processing participants engaged in, as gauged by their self-reports. As in Experiment 1, we performed a series of logistic regressions to examine the relationship between self-reported explanation and self-reported comparison and the discovery of at least one 100% rule. A simultaneous multiple logistic regression predicting discovery of at least one 100% rule from amount of self-reported explanation and the amount of self-reported total comparison found a significant positive effect of self-reported explanation, $W(1) = 14.2$, $p < .001$, $\beta = .177$, Exp(β) = 1.194 (constant-only model: $W(1) = 0.756$, $p = .38$, $\beta = -.080$, Exp(β) = 0.923), and a marginal positive effect of self-reported comparison, $W(1) = 3.67$, $p = .056$, $\beta = .066$, Exp(β) = 1.068. Separate logistic regressions of discovery of a 100% rule on the amounts of self-reported within-category and between-category comparison found a significant positive effect of between-category comparison on discovery of at least one 100% rule, $W(1) = 8.38$, $p = .004$, $\beta = .155$, Exp(β) = 1.167 (constant-only model: $W(1) = 0.983$, $p = .32$, $\beta = -.089$, Exp(β) = 0.914), but not of within-category comparison, $W(1) = 2.00$, $p = .16$, $\beta = .081$, Exp(β) = 1.084 (constant-only model: $W(1) = 1.17$, $p = .28$, $\beta = -.098$, Exp(β) = 0.907).

3.2.3. Discovery of one or more 75% rules

We analyzed whether 75% rule discovery varied across study conditions by performing a log-linear analysis of discovered vs. did not discover a 75% rule equivalent to that for the 100% rule (see Figure 3B). Participants who performed a group study task were significantly more likely to report discovering a 75% rule than participants who performed an individual or pairwise study task, $\chi^2(1) = 5.21$, $p = .022$. However, whether participants performed an explanation task versus a comparison task did not have a significant effect on 75% rule discovery, $\chi^2(1) = 0.56$, $p = .46$. Participants who reported at least one 100% rule were marginally less likely to report a 75% rule than participants who did not report a 100% rule, Fisher's Exact Test: $p = .089$.

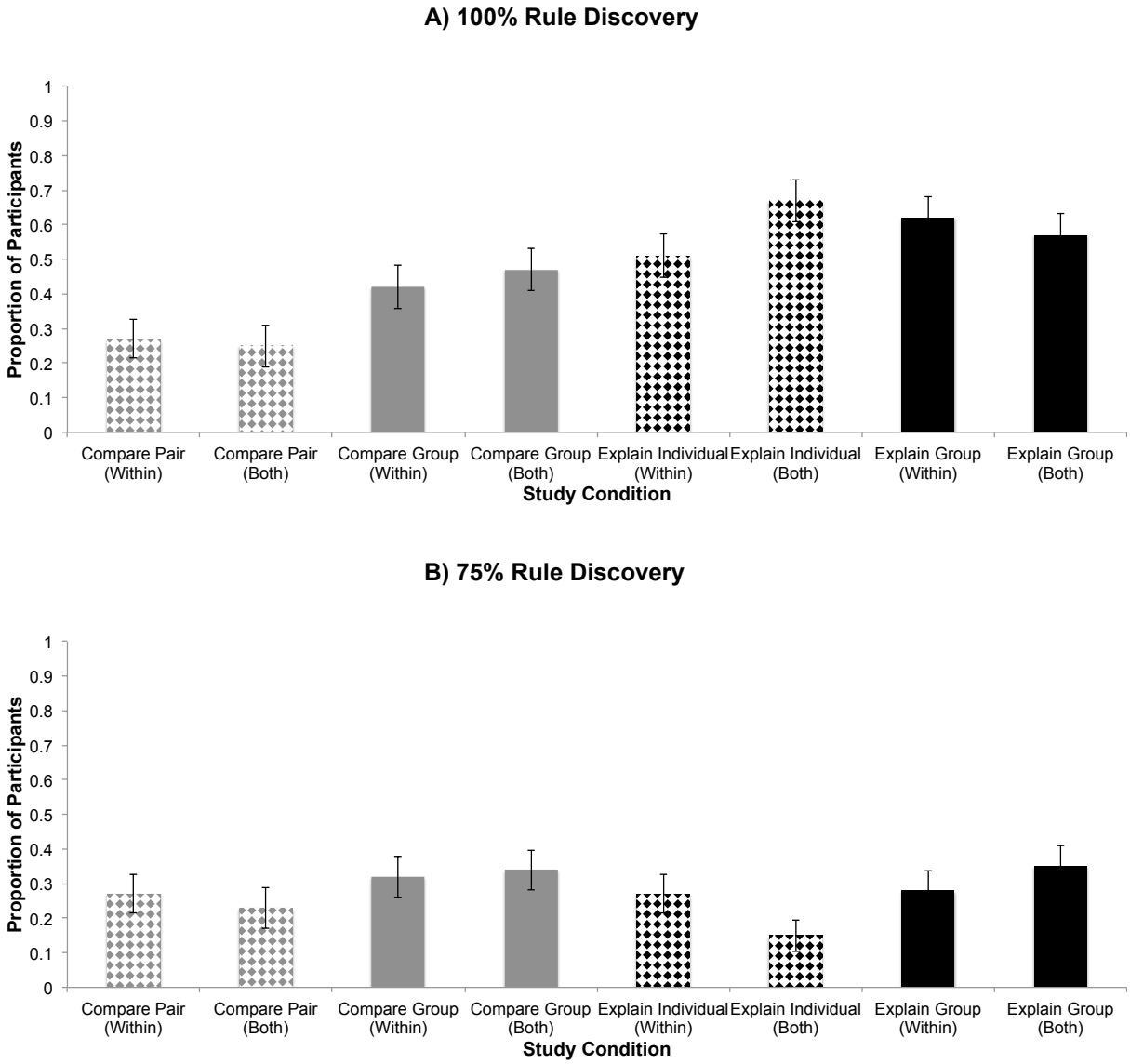

Figure 3: Proportion of participants in each study condition who discovered at least one 100% rule (Fig. 3A) and at least one 75% rule (Fig. 3B) in Experiment 2. "Within" participants only received within-category study prompts, whereas "both" participants received both within-category and between-category study prompts. Error bars indicate +/- 1 SE.

3.3. Discussion

Most importantly, Experiment 2 went beyond Experiment 1 in demonstrating that prompting group comparison was significantly more effective than prompting pairwise comparison for promoting the discovery of at least one 100% rule. This lends credence to the idea that the emphasis on pairwise comparison prompts in Experiment 1 may have decreased participants' attention to the overall structure of the categories. In Experiment 3, we examined whether group comparison might mediate effects of explanation on 100% rule discovery.

Additionally, Experiment 2 replicated our prior finding that, as predicted, prompts to explain led to increased comparison as well as increased explanation (as assessed by self-report). Indeed, in both experiments, prompts to explain resulted in more comparison than did prompts to compare. We also replicated the previous result that an explanation prompt was associated with greater 100% rule discovery than was a pairwise comparison prompt. We found that self-reported explanation was associated with 100% rule discovery, as in Experiment 1, and additionally that self-reported between-category comparison was associated with 100% rule discovery.

4. Experiment 3

Given our findings so far—that prompts to explain fostered comparison processing (Experiments 1 and 2), and that receiving the group comparison prompts (but not pairwise comparison prompts) increased rule discovery for a 100% rule (Experiment 2), we next asked whether group comparison processing mediates the relationship between explanation and 100% rule discovery. In Experiment 3, we used a 3 × 2 design: participants were prompted to generate explanations, make comparisons, or engage in a control task, and within each of these three study

conditions, participants were prompted to engage in either group study or individual/pairwise study. We also included a self-report measure of group comparison to allow us to test for mediation.

4.1. Method

4.1.1. Participants

Participants were 284 adults recruited from Amazon Mechanical Turk and tested online. An additional 145 participants were tested, but excluded from the analyses. The exclusion criteria were the same as in the previous experiments.

4.1.2. Materials

The stimuli were the same eight robots used in the previous experiments.

4.1.3. Procedure

The procedure consisted of a study phase, a rule-reporting phase, a self-report phase, and end-of-study questions. For the study phase, each participant was randomly assigned to one of six conditions. The conditions were based on a 3 × 2 design, in which participants were prompted to do explanation or comparison or responded to control prompts, and independently, were asked to engage in either individual/pairwise study or group study. The total study time in each condition was 360 seconds. In all conditions, the picture of the eight robots remained visible for the entirety of the study phase. The group comparison condition and both explanation conditions were identical to the corresponding within-category study conditions from Experiment 2. The pairwise comparison condition was similar to the corresponding within-category pairwise comparison condition from Experiment 2; however, in Experiment 3, participants studied each pair once for 90 seconds. The two control conditions are described below.

Individual study control condition. Participants responded to prompts of the form "*Write out your thoughts* as you learn to categorize Glorp [Drent] robot X." The study order was A, B, F, H, C, D, E, and G. Participants studied each robot for 45 seconds.

Group study control condition. Participants responded to the prompts "*Write out your thoughts* as you learn to categorize the Glorp [Drent] robots (robots A-D [E-H])." Participants studied the Glorp robots (A-D) followed by the Drent robots (E-H). Participants studied each group of robots for 180 seconds.

As in the previous experiments, participants completed simple math problems every 180 seconds in between prompts as a catch trial.

The rule-reporting phase was identical to the previous experiments. In the self-report questions, participants were asked about the extent to which they engaged in three cognitive processes: explanation, pairwise comparison, and group comparison. The three prompts were "*Explaining* why particular robots are Glorp robots or Drent robots," "*Comparing* pairs of robots (i.e., noting similarities and differences between TWO robots)," and "*Comparing* all the robots from the same category to each other (i.e., noting similarities and differences between ALL FOUR Glorps or Drents)," respectively. After answering these questions on the same 1-7 scale used in the previous experiments, participants were asked to estimate the number of within-category and between-category pairwise comparisons that they performed. The response options were as follows: 0, 1-4, 5-8, 9-12, 13-16, and more than 16. They then answered the same end-of-study questions as in the previous experiments.

4.2. Results

4.2.1. Self-reports

Means for self-reported explanation, pairwise comparison, and group comparison are shown in Table 5. We analyzed these data by performing a series of 3 × 2 ANOVAs with study prompt (control vs. explanation vs. comparison) and whether participants engaged in group study vs. individual/pairwise study as between-subjects factors and the amounts of self-reported explanation, self-reported pairwise comparison, and self-reported group comparison as the dependent variables. Both study prompt, $F(2, 265) = 10.1$, $p < .001$, $\eta^2 = 0.071$, and whether participants were asked to engage in group study vs. individual/pairwise study, $F(1, 265) = 4.52$, $p = .034$, $\eta^2 = 0.017$, affected the amount of self-reported explanation. Participants assigned to do group study reported more explanation than participants assigned to do individual/pairwise study. A Tukey post-hoc analysis found that participants in the explanation condition reported significantly more explanation than participants in the comparison ($p < .001$) and control ($p = .038$) conditions, but found no significant differences in self-reported explanation between the comparison and control conditions.

An equivalent ANOVA for the amount of self-reported pairwise comparison found a significant effect of study prompt, $F(2, 270) = 4.87$, $p = .008$, $\eta^2 = 0.035$. There was also an interaction between the two factors: $F(2, 270) = 7.563$, $p = .001$, $\eta^2 = 0.053$. A Tukey post-hoc analysis showed that participants in the comparison condition reported significantly more pairwise comparison than participants in the control condition ($p = .009$), but found no other significant differences. $F(2, 270) = 7.563$, $p = .001$, $\eta^2 = 0.053$. An equivalent ANOVA for the amount of self-reported group comparison found significant main effects of study prompt, $F(2, 276) = 7.44$, $p = .001$, $\eta^2 = 0.051$, and group study vs. individual/pairwise study, $F(1, 276) = 8.40$, $p = .004$, $\eta^2 = 0.030$, with group study participants reporting significantly more group comparison than individual/pairwise participants, as well as an interaction between these two

factors, $F(2, 276) = 3.32$, $p = .037$, $\eta^2 = 0.024$. A Tukey post-hoc analysis found that participants in the explanation condition performed more group comparison than participants in the control condition ($p = .024$) and comparison condition ($p = .001$).

In sum, the explanation prompts were effective in boosting self-reported explanation, as well as in increasing self-reported *group* comparison. The comparison prompts did succeed in increasing self-reported *pairwise* comparison relative to control prompts, but not relative to explanation prompts.

Table 5: Self-reported explanation and comparison in each study condition in Experiment 3.

Study Condition	**Self-reported Explanation Mean (SD)**	**Self-reported Pairwise Comparison Mean (SD)**	**Self-reported Group Comparison Mean (SD)**
Group Comparison	4.17 (2.27)	4.59 (2.13)	5.61 (1.58)
Pairwise Comparison	3.67 (2.31)	5.82 (1.54)	4.57 (2.13)
Group Explanation	5.14 (1.62)	4.69 (2.05)	5.91 (1.40)
Individual Explanation	5.32 (1.83)	4.78 (2.07)	6.04 (1.29)
Group Control	5.02 (1.67)	4.81 (2.03)	5.70 (1.33)
Individual Control	3.78 (2.20)	3.82 (1.75)	4.90 (2.00)

4.2.2. Discovery of one or more 100% rules

Next, we performed a log-linear analysis of study prompt × group vs. individual/pairwise study × discovered vs. did not discover at least one 100% rule; see Figure 4A. The log-linear analysis showed a significant effect of study prompt on the proportion of participants discovering at least one 100% rule, $\chi^2(2) = 22.3$, $p < .001$, but not a significant effect of group study versus individual/pairwise study, $\chi^2(1) = 0.161$, $p = .69$. Since these analyses did not find any significant effects of group study versus individual/pairwise study,[7] we focus on effects of study prompt. We performed a series of similar log-linear analyses that evaluated the nature of the effect of study prompt across pairs of conditions (e.g., explanation vs. comparison). We found that participants receiving explanation prompts were significantly more likely to discover a 100% rule than those receiving control prompts, $\chi^2(1) = 9.05$, $p = .003$, and those receiving comparison prompts, $\chi^2(1) = 21.8$, $p < .001$, but that the latter two conditions did not differ from each other, $\chi^2(1) = 2.50$, $p = .11$.

We also performed a simultaneous logistic regression of discovered at least one 100% rule (yes vs. no) on the amounts of (1) self-reported explanation, (2) self-reported group comparison, and (3) self-reported pairwise comparison. This analysis found significant positive associations between self-reported explanation and 100% rule discovery, $W(1) = 7.18$, $p = .007$, $\beta = .216$, $\text{Exp}(\beta) = 1.207$ (constant-only model: $W(1) = 6.97$, $p = .008$, $\beta = -.33$, $\text{Exp}(\beta) = 0.719$), and between self-reported group comparison and 100% rule discovery, $W(1) = 3.254$, $p < .001$, β

[7] In contrast to Experiment 2 (reported here) and Experiment 3 of Edwards, Williams, and Lomobrozo (2013), participants who received group comparison prompts were not significantly more likely to discover at least one 100% rule than participants who received pairwise comparison prompts. However, there was a non-significant trend in the same direction, $\chi^2(1) = 0.914$, $p = .34$, as shown in Figure 4.

=.32, Exp(β) = 1.377, and a marginal negative association between self-reported pairwise comparison and 100% rule discovery, $W(1) = 3.25$, $p = .071$, $\beta = -.09$, Exp(β) = .888.

4.2.3. Discovery of one or more 75% rules

A log-linear analysis of discovered at least one 75% rule equivalent to that for the 100% rules found significant effects of study prompt, $\chi^2(2) = 6.56$, $p = .038$, and of group study vs. individual/pairwise study, $\chi^2(1) = 8.55$, $p = .003$, with participants in the group study condition more likely to discover a 75% rule than participants who performed individual/pairwise study (see Figure 4B). We followed up on these significant effects by performing a series of log-linear analyses that evaluated the nature of these effects across pairs of study conditions.

Participants receiving control prompts were significantly more likely to discover a 75% rule than those receiving explanation prompts, $\chi^2(1) = 6.27$, $p = .012$, and across both of these prompt conditions, group study participants were significantly more likely to discover a 75% rule than individual study participants, $\chi^2(1) = 7.86$, $p = .005$. The analysis of explanation and comparison participants found that participants receiving comparison prompts were marginally more likely to discover a 75% rule than those receiving explanation prompts, $\chi^2(1) = 2.96$, $p = .085$, and that group study participants were significantly more likely to discover a 75% rule than individual/pairwise study participants, $\chi^2(1) = 4.62$, $p = .032$. Among control and comparison participants, group study participants were significantly more likely to discover a 75% rule than individual/pairwise study participants, $\chi^2(1) = 4.97$, $p = .026$, but the proportion of participants who discovered at least one 75% rule did not significantly differ across prompt conditions, $\chi^2(1) = 0.731$, $p = .39$.

Participants who reported at least one 100% rule were significantly less likely to report a 75% rule relative to those who did not report a 100% rule, Fisher's Exact Test: $p = .036$.

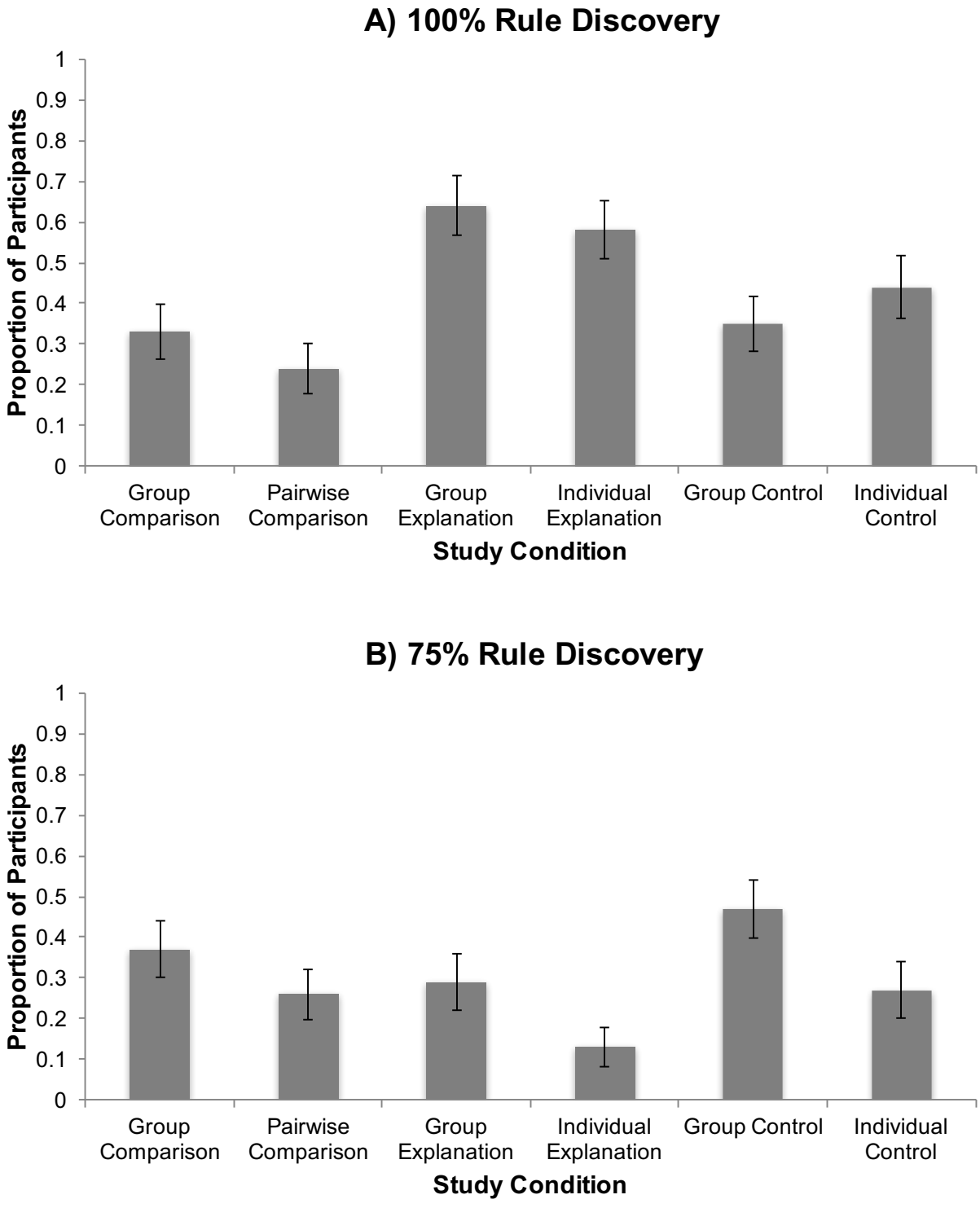

Figure 4: Proportion of participants in each study condition who discovered at least one 100% rule (Fig. 4A) and at least one 75% rule (Fig. 4B) in Experiment 3. Error bars indicate +/- 1 SE.

4.2.4. Mediation analysis

One reason for performing Experiment 3 was to investigate whether group comparison processing mediates the relationship between performing the explanation task (vs. control) and 100% rule discovery. In this analysis, the group explanation and individual explanation conditions were combined, as were the group control and individual control conditions. We excluded participants in the comparison condition from this analysis because we were interested in possible mediation effects of *spontaneous* group comparison processing (i.e., in the absence of a prompt to compare). In order for self-reported group comparison to mediate the relationship between explanation and 100% rule discovery, three conditions must hold (Baron & Kenny, 1986). First, there must be a significant effect of the independent variable (whether participants performed the explanation task vs. the control task) on the potential mediator (the amount of self-reported group comparison). Second, there must be a significant effect of the independent variable on the dependent variable (whether or not participants discovered a 100% rule). Third, there must be a significant effect of the mediator on the dependent variable after controlling for the independent variable. A series of three regressions evaluated whether these conditions were satisfied.

First, a linear regression of the amount of self-reported group comparison on study prompt found a significant correlation between these variables ($r = .21$, $p = .005$), with participants who received explanation prompts engaging in more group comparison than participants who received control prompts. Second, a logistic regression of whether participants discovered a 100% rule on study prompt found that participants who received explanation prompts were significantly more likely than control participants to discover a 100% rule, $W(1) = 8.93$, $p = .003$, $\beta = .901$, $\text{Exp}(\beta) = 2.46$, (constant-only model: $W(1) = 0.005$, $p = .94$,

β = .011, Exp(β) = 1.011). Third, a simultaneous multiple logistic regression of whether participants discovered a 100% rule on study prompt and the amount of self-reported group comparison found a significant positive effect of self-reported group comparison on whether participants discovered a 100% rule even after controlling for the type of study prompt participants received, $W(1) = 7.30$, $p = .007$, $\beta = .295$, Exp(β) = 1.343 (constant-only model: $W(1) = 0.022$, $p = .88$, $\beta = .022$, Exp(β) = 1.022). Thus, the three conditions for mediation are satisfied. Additionally, the third analysis found a significant effect of study prompt on whether participants discovered a 100% rule, with the explanation prompts boosting performance (vs. control) even after controlling for the amount of self-reported group comparison, $W(1) = 5.63$, $p = .018$, $\beta = .738$, Exp(β) = 2.092, indicating partial mediation as opposed to complete mediation.

The mediation analysis suggests that group comparison (as assessed by self-reports) partially mediates the relationship between performing the explanation task and discovering a 100% rule. Interestingly, a comparable analysis involving self-reported *pairwise* comparison, rather than self-reported *group* comparison, suggested that pairwise comparison does not mediate the relationship between explanation and 100% rule discovery.[8] These results thus suggest that group comparison (i.e., comparing all members of a particular category), in particular, is one mechanism by which engaging in explanation promotes category learning, and provide evidence for a relationship between explanation and comparison in a category-learning task.

4.3. Discussion

[8] A linear regression of the amount of self-reported pairwise comparison on study prompt did not find a significant correlation between these variables, $r = .09$, $p = .23$. Thus, the first condition for mediation did not hold.

Experiment 3 replicated several key results from Experiment 2, but also went beyond Experiment 2 in soliciting self-reports for group comparison. This allowed us to show that explanation prompts increased self-reported group comparison, and that engaging in this comparison strategy partially mediated the effect of explanation prompts on discovery of a 100% rule. In the General Discussion we consider these results in the context of our other findings.

5. General discussion

In the Introduction, we posed three questions regarding the nature of the relationship between explanation and comparison in a category-learning task. First, we asked whether engaging in explanation can recruit comparison processing. Second, we asked which comparison strategies explanation recruits. And third, we asked how these comparison strategies affect category learning. We address each question in turn; the results are summarized in Table 6.

In all three experiments, performing the explanation task increased self-reported comparison processing, even above an already-high baseline level. In Experiment 1, participants who received explanation prompts reported doing more comparison than participants who did not receive these prompts, even relative to participants who were explicitly prompted to compare. This striking pattern—that explanation instructions were more effective in inducing participants to engage in comparison processing than were direct comparison instructions—also held for Experiment 2.

Crucially, Experiment 3, in which participants were asked about the specific comparison strategies they performed, found that performing the explanation task increased group comparison, but not pairwise comparison. Furthermore, in Experiment 3, group comparison partially mediated the relationship between performing the explanation task and 100% rule

discovery, while pairwise comparison did not. This analysis indicates that explanation is an effective cognitive strategy for promoting discovery of the 100% categorization rules in part because engaging in explanation leads people to engage in comparison. Moreover, our findings shed light on the kind of comparison strategy that matters for this task: group-level comparison.

Consistent with a Subsumptive Constraints account of explanation (Williams & Lombrozo, 2010, 2013), we hypothesize that performing the explanation task led participants to search preferentially for a simple rule that accounted for all items. Explanation participants' higher rate of spontaneous group comparison suggests that these participants performed category-wide comparisons to identify features shared by all members of the same category, and that when compared across categories, are diagnostic of category membership. Given that comparison can support abstract re-representation (Gentner & Medina, 1998), the spontaneous comparisons may also have helped explanation participants re-represent surface-level differences (e.g., triangular feet, diamond-shaped feet) to discover the abstract commonalities (e.g., all Glorp robots have pointy feet) underlying the 100% rules.

The mediation analyses also suggest that explanation supports category learning in ways that go above and beyond effects of comparison. Specifically, performing the explanation task was positively correlated with 100% rule discovery, even after controlling for the amount of reported group comparison. Previous research on explanation (e.g., Williams & Lombrozo, 2010) suggests one possible reason why: explanation encourages learners to seek broad, consistent patterns underlying what they are trying to explain. Indeed, replicating Williams and Lombrozo (2010, 2013), we found that explanation increased discovery of 100% rules, but decreased (or had no effect on) discovery of the more salient but less satisfying 75% rules. By contrast, group comparison prompts increased discovery of both 100% and 75% rules,

suggesting that these prompts helped participants re-represent features and track global statistics, but did not constrain participants to search selectively for the broadest and simplest patterns available (see also Kon & Lombrozo, in press).

Generating explanations may also promote a search for broad patterns with the goal of identifying a causal regularity that underlies category membership (Rehder, 2007). While the artificial nature of our categories—including the category labels—made it difficult for participants to come up with a causal explanation, it is worth noting that comparable effects of explanation prompts on 100% rule discovery have been found for property-generalization tasks involving causally-meaningful explanations (Kon & Lombrozo, 2017, in prep). In the current task, explanation participants may have been more likely than participants in other conditions to see the properties underlying 75% rules as merely accidental as opposed to causally meaningful, and thus not a basis for a rule-based or family-resemblance categorization scheme.

Table 6: Summary of the experiments.

Experiment	Condition (Prompts)	Effects of Condition on 100% Rule Discovery	Effects of Prompts on Self-Reported Processing	Relationship Between Self-Reported Processing and 100% Rule Discovery
1: What are the relative effects of explanation and comparison prompts on a category-learning task?	Explanation (Y/N) × Within-category pairwise comparison (Y/N)	• Explanation prompt > No explanation prompt • No comparison prompt > Comparison prompt	• Explanation prompt → More explanation processing and more comparison processing • Comparison prompt → Less explanation processing, no effect on comparison processing	• Explanation processing → Greater rule discovery • No effect of comparison processing on rule discovery
2: Is prompting "group comparison" (i.e., comparing all members of the same category) more effective than prompting pairwise comparison?	Explanation vs. Comparison × Individual/Pairwise study vs. Group study × Within-category study only vs. Within-category + between-category study	• Explanation prompt > Comparison prompt • Group comparison prompt > Pairwise comparison prompt • Group comparison prompt = Individual explanation prompt	• Explanation (relative to comparison) prompt → More explanation processing and more comparison processing	• Explanation processing → Greater rule discovery • Between (but not within)-category comparison processing → Greater rule discovery
3: Does comparison mediate effects of explanation on category learning?	Explanation vs. Comparison vs. Control × Individual/Pairwise study vs. Group study	• Explanation prompt > Comparison prompt • Explanation prompt > Control prompt	• Group study prompt → More explanation processing • Explanation (relative to comparison and control) prompts → More explanation processing and more group comparison processing • Comparison (relative to control) prompt → More pairwise comparison processing	• Explanation processing → Greater rule discovery • Group (but not pairwise) comparison processing → Greater rule discovery • Group (but not pairwise) comparison processing partially mediated effects of the explanation task on 100% rule discovery

5.1. Comparison effects and non-effects

Our data with respect to comparison are somewhat mixed. In Experiments 2 and 3, we found that the degree of comparison processing (as assessed by self-reports) was positively related to performance on the rule discovery task. These findings are consistent with prior work showing positive effects of comparison on category learning (Christie & Gentner, 2010; Gentner,

Anggoro, & Klibanoff, 2011; Gentner & Namy, 1999; Higgins & Ross, 2011; Kurtz, Boukrina, & Gentner, 2013; for a review, see Gentner, 2010). We also found that the positive effects of explanation on category learning were partially mediated by group comparison processing.

However, despite the positive effects of comparison *processing*, we mostly failed to find an effect of comparison *instructions*. In particular, in all but Experiment 3, pairwise comparison prompts failed to promote comparison processing above the level of other conditions, including the control group (as assessed by self-report). As noted earlier, the ineffectiveness of comparison prompts probably resulted in part from the high level of spontaneous comparison processing across conditions, even in the control condition. Recall that the study robots were all very similar to each other, and all eight were displayed throughout the study phase. This made it difficult for comparison instructions to increase the level of comparison processing above baseline. In addition, prompting a series of independent pairwise comparisons, as in Experiments 1 and 2, may have encouraged suboptimal comparison strategies that encouraged focusing on individual details. Indeed, many participants in Experiment 1 engaged in at least a limited form of group comparison despite being prompted to engage in pairwise comparisons. Moreover, the pairwise comparison prompt in Experiment 1 led participants to engage in less explanation, and this may also have contributed to the relatively low performance in the comparison conditions. In contrast, prompting group comparison does appear to foster discovering category-wide patterns.

Prompting comparison is likely to be most effective in situations in which participants are unlikely to compare spontaneously. Accordingly, we predict that comparison instructions would be more useful when the relevant comparisons are less obvious (e.g., Kurtz, Miao, & Gentner, 2001). Additionally, comparison prompts may be more beneficial to children, who are less experienced at using comparisons to further an explanatory goal. These reflections suggest a

complementary claim for explanation: that explanation instructions may have been effective in the current task in part because our artificial materials were only minimally connected to prior knowledge, and hence spontaneous explanations may have been relatively low compared to what we would find with richer materials.

5.2. Why were group comparison prompts more effective?

In Experiment 2 (but not in Experiment 3), and also in Experiment 3 of Edwards, Williams, and Lombrozo (2013), participants who were prompted to engage in group comparison were significantly more likely to discover a 100% rule than those prompted to engage in pairwise comparison. This could have come about in two ways. First, the group comparison strategy may have made participants more likely to notice category-wide structure across the exemplars than did a series of pairwise comparisons. Second, group comparison may have favored re-representing exemplar features in a more abstract way that revealed diagnostic properties. Interestingly, in both Experiments 2 and 3, group study also tended to increase 75% rule discovery. Thus, like explanation, group study improved participants' ability to discover the 100% rules. But unlike explanation, group comparison did not seem to constrain participants to discover or report only 100% rules.

This raises the question of what cognitive processes these group comparison participants were engaging in when studying the robots. Very few studies have included group comparison prompts, and one study that did do so found them do be less effective than a sequence of paired comparisons (Thompson & Opfer, 2010). However, the sequential comparison condition in their study used a progressive alignment order (from highly concrete to more abstract comparisons)—an order that has been found to be optimal in several studies (e.g., Gentner, Anggoro, & Klibanoff, 2011; Goldstone & Son, 1995; Kotovsky & Gentner, 1996).

In general, theories of comparison have treated comparison as a pairwise process of structural alignment and inference (e.g., Gentner, 1983, 2010; Gentner, Holyoak, & Kokinov, 2001; Krawczyk, Holyoak, & Hummel, 2005). One possibility is that group comparison participants engaged in a fundamentally different form of comparison in which they simultaneously compared all four robots in each category. However, a more likely possibility is that participants engaged in pairwise comparison, but did so in service of discovering the overall category structure. One specific version of this proposal is that participants may have carried out a process of repeated comparison and abstraction, as in the SAGE model of category formation (Forbus et al., 2017; see also Kuehne et al, 2000). In this account, an initial abstraction is formed by comparing one pair of items; then further members are sequentially compared with that abstraction, resulting in a progressively more general abstraction that covers the category. Such a process has been proposed to account for infant relational learning (Ferry et al., 2015). Finer-grained measures of what participants were doing when studying the robots (e.g., think-aloud protocols, eye-tracking data) are needed to discriminate between these two accounts.

5.3. Limitations and future directions

In our studies, a clear pattern emerges in which the goal of explaining category membership invokes comparison processes, which aid in discovering a 100% rule. Although we hypothesize that this pattern may be quite general, we must be cautious in generalizing the current results. First, a category-learning task is especially well-suited for comparison. Indeed, it is hard to explain why a robot is a member of a particular category without comparing it to other robots. Second, in our studies, the members of both categories were visually available throughout the task, making it easy for participants to compare both within and across categories. This could have inflated both the baseline level of comparison and the role of comparison in the explanatory

process. While this is clearly a concern, the structure of the robot categories—family resemblance with a defining feature—is typical of many real-world categories. Further, our finding that people prefer 100% rules over more obvious 75% regularities is consistent with much past work in category construction, including work involving categories that, like ours, were analyzable in terms of prototypes instead of unidimensional rules. Third, we used artificial materials in which participants lack background knowledge. While this likely had the advantage of reducing the baseline level of explanation processing, making it easier to manipulate the amount of explanation that participants performed when studying the robots, future work could evaluate the nature of effects of explanation and comparison on category learning using content-rich categories. It may also be beneficial to explore similar research questions using materials with different types of category structures and different methods of presentation.

As discussed above, the high baseline level of comparison probably reduced effects of comparison instructions. Thus, one direction for future research is to conduct studies using materials that are less apt to be spontaneously compared. Relatedly, future work should explore the exact mechanisms at work in group comparison—for example, whether it involves sequential comparison and abstraction vs. some other kind of comparison process. Another limitation of the present studies concerns our reliance on retrospective self-report measures as an index of the extent to which participants actually engaged in explanation and comparison. Future work could track processing during learning using additional measures, such as eye tracking or coded think-aloud protocols, which could potentially provide convergent support for our interpretation.

Future research might also explore the roles of explanation and comparison in a category-construction task in which participants are not told the category to which each item belongs, and must instead sort a set of uncategorized items into various categories (e.g., Medin, Wattenmaker,

& Hampson, 1987). Prior research using such tasks has found that people often prefer a unidimensional rule for dividing items into categories, even when the stimuli are constructed to favor prototype sorting instead (Medin, Wattenmaker, & Hampson, 1987; Regehr & Brooks, 1995; see also Murphy (2002), pp. 127-133). However, performing a property induction task that emphasizes the relationship between different properties (e.g., "Given that this animal has a short tail, what kind of teeth would you expect it to have?") can make participants more likely to generate family resemblance categories (Lassaline & Murphy, 1996). Background knowledge also has important effects on the kind of category structure that participants construct. For example, Ahn (1990) found that telling participants why some feature values were more appropriate for one category than the other resulted in a higher proportion of family resemblance structures. Additionally, Spalding and Murphy (1996) found that participants were able to use their existing prior knowledge to link features to categories in this way, similarly resulting in a higher proportion of family resemblance structures. Importantly, these family resemblance structures involved features that fit into a common explanation or "theme" – for example, the features "made in Norway" and "heavily insulated" were associated with the theme of being an arctic vehicle (see also Murphy & Allopenna, 1994; Williams, Lombrozo, & Rehder, 2013). So while the resulting category structures were not based on a single defining feature, they were unified by a single theme.

How might explanation and comparison affect the way this learning unfolds? We hypothesize that in the absence of well-defined categories, participants would initially need to rely extensively on comparisons to discover similarities across items that can potentially be used as a basis for categorization. Given the task of discovering categories, we would expect participants to engage in sequential abstraction or some other form of group comparison.

Consistent with this idea, Spalding and Murphy (1996) found that participants were more likely to generate family resemblance structures when they were encouraged to first "preview" all items, which may have prompted such comparisons across the groups. They also suggest that people may approach category construction by "attempting to find underlying reasons or explanations for the categories" (Spalding & Murphy, 1996, p. 527), which could provide a simple and broad rule that underlies category membership at the level of the "theme" (e.g., "arctic vehicle" or not) even if the resulting rule is not unidimensional when it comes to individual features (e.g., being "made in Norway" or not). Given that explanation has been shown to recruit prior knowledge (Williams & Lombrozo, 2013), encourage the discovery and use of such themes (Williams, Lombrozo, & Rehder, 2013), and promote classification on the basis of simple and broad patterns (Lombrozo, 2016), we speculate that spontaneous explanation could underlie some of these findings regarding category construction.

Additionally, explanation is likely to involve many different comparisons, not all of them visual or perceptual, suggesting that the present findings could generalize quite broadly—including beyond the categorization domain. In particular, as mentioned in the Introduction, explanations often involve an implicit contrast. In some cases the contrast may be with a default or counterfactual possibility (e.g., Why did the fire start as opposed to not having started?), which could initiate a comparison between actual and counterfactual objects or events. Such comparisons will rarely be supported by simultaneously presented visual images, but could nonetheless play an important role in generating or evaluating explanations. Accordingly, explanation may recruit spontaneous comparison when generating or evaluating causal explanations. A contrast with a counterfactual alternative such as "If Mike hadn't partied the night before the exam, then he wouldn't have failed" suggests partying as a candidate

explanation for why Mike failed the exam. Likewise, when evaluating a proposed causal explanation, we often spontaneously imagine what would have happened in a specific counterfactual situation and compare this situation with the actual state of events in order to decide whether to accept or reject the proposed explanation.

5.4. Conclusion

The central contributions of the present work are twofold. First, our studies indicate that making comparisons is one mechanism by which engaging in explanation supports category learning. Second, the present study identifies the type of explanation-induced comparison strategy that results in a positive learning outcome in this domain: group-level comparison. By investigating explanation and comparison together, future studies can achieve a greater understanding of both processes and provide insights into how they combine to support learning.

Acknowledgements

We thank David Rapp and members of the Northwestern University Language and Cognition Laboratory and the UC Berkeley Concepts and Cognition Lab for valuable feedback on this work, and Lena Lam for assistance with coding. This work was supported by an NSF Graduate Research Fellowship (BJE), SILC grant SMA1041707 (DG), ONR grant N00014-13-1-0470 (DG), NSF grant DRL-1056712 (TL), and the McDonnell Foundation (TL).

References

Ahn, D., & Yo, N. (2011). Social science for pennies. *Science*, *334*, 307.

Ahn, W. (1990). Effects of background knowledge on family resemblance sorting. *Proceedings of the 12th Annual Conference of the Cognitive Science Society* (pp. 149-156). Hillsdale, NJ: Erlbaum.

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

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). *Learning through case comparisons: A meta-analytic review*. *Educational Psychologist*, *48*, 87-113.

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

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173-1182.

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, *30*, 703-725.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*, 1147-1156.

Chi, M. T. H. (2000). Self-explaining expository tests: 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. W., 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.

Chin-Parker, S., & Bradner, A. (2017). A contrastive account of explanation generation. *Psychonomic Bulletin and Review*, *24*, 1387-1397.

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, *11*, 356-373.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*, 1-43.

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

Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science*, *25*, 245-286.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*, 1-63.

Ferry, A. L., Hespos, S. J., & Gentner, D. (2015). Prelinguistic relational concepts: Investigating analogical processing in infants. *Child Development*, *86*, 1386-1405.

Fonseca, B. A., & Chi, M. T. H. (2011). Instruction based on self-explanation. In R. Mayer & P. Alexander (Eds.), *The handbook of research on learning and instruction* (pp. 296-321). New York: Routledge Press.

Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*, *41*, 1152-1201.

Gadgil, S., Nokes-Malach, T. J., & Chi, M. T. H. (2012). Effectiveness of holistic mental model confrontation in driving conceptual change. *Learning and Instruction*, *22*, 47-61.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought.* New York: Oxford University Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155-170. (Reprinted in A. Collins & E. E. Smith (Eds.), *Readings in cognitive science: A perspective from psychology and artificial intelligence*. Palo Alto, CA: Kaufmann).

Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), Language in mind: Advances in the study of language and thought (pp. 195-235). Cambridge, MA: MIT Press.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34*, 752-775.

Gentner, D., Anggoro, F. K., & Klibanoff, R. S. (2011). Structure mapping and relational language support children's learning of relational categories. *Child Development*, *82*, 1173-1188.

Gentner, D., Holyoak, K. J., & Kokinov, B. (Eds.). (2001). *The analogical mind: Perspectives from cognitive science*. Cambridge, MA: MIT Press.

Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. *Cognitive Science*, *33*, 1343-1382.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*, 45-56.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, *65*, 263-297.

Gentner, D., & Namy, L. (1999). Comparison in the development of categories. *Cognitive Development*, *14*, 497-513.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, *25*, 524-575.

Gentner, D., & Toupin C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, *10*, 277-300.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1-38.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*, 125-157.

Goldstone, R. L., & Son, J. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of Learning Sciences*, *14*, 69-110.

Goldwater, M. B., & Gentner, D. (2015). On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. Cognition, *137*, 137-153.

Higgins, E. J., & Ross, B. H. (2011). Comparisons in category learning: How best to compare for what. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Hoyos, C., & Gentner, D. (2017). Generating explanations via analogical comparison. *Psychonomic Bulletin and Review*, *24*, 1364-1374.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427-466.

Hummel, J. E., Landy, D. H., & Devnich. D. (2008). Toward a process model of explanation with implications for the binding problem. In *Naturally Inspired AI: Papers from the AAAI Fall Symposium*. Technical Report FS-08-06, 79-86.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, *111*, 640-661.

Kon, E., & Lombrozo, T. (2017). Explaining guides learners towards perfect patterns, not perfect prediction. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 682-687). Austin, TX: Cognitive Science Society.

Kon., E., & Lombrozo, T. (in press). Scientific discovery and the human drive to explain. In Richard Samuels & Daniel Wilkenfeld (Eds.), *Advances in Experimental Philosophy of Science*. New York, NY: Bloomsbury Press.

Kon, E., & Lombrozo, T. (in prep). Why explainers take exception to exceptions. *Manuscript in Preparation*.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*, 2797-2822.

Krawczyk, D. C., Holyoak, K. J., & Hummel, J. E. (2005). The one-to-one constraint in analogical mapping and inference. *Cognitive Science*, *29*, 797-806.

Kuehne, S.E., Forbus, K. D., Gentner, D., & Quinn B. (2000). SEQL-Category learning as progressive alignment using structure mapping. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 770-775.

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

Kurtz, K. J., Boukrina, O., & Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1303-1310.

Kurtz, K. J., Miao, C., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences*, *10,* 417-446.

Lassaline, M. E., & Murphy, G. L. (1996). Induction and category coherence. *Psychonomic Bulletin & Review*, *3*, 95-99.

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

Loewenstein, J., Thompson, L., & Gentner, D. (2003). Analogical learning in negotiation teams: Comparing cases promotes learning and transfer. *Academy of Management Learning and Education*, *2*, 119-127.

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

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

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*, 748-759.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*, 94-122.

Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, *32*, 517-535.

Markman, A. B., & Wisniewski, E. J. (1997). Similar and different: The differentiation of basic-level categories. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 54-70.

McEldoon, K. L., Durkin, K. L., & Rittle-Johnson, B. (2013). Is self-explanation worth the time? A comparison to additional practice. *British Journal of Experimental Psychology*, *83*, 615-632.

McGill, A. L. (2002). Alignable and nonalignable differences in causal explanations. *Memory and Cognition*, *30*, 456-468.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*, 242-279.

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

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

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

Namy, L. L., & Gentner D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General*, *131*, 5-15.

Nokes-Malach, T. J., VanLehn, K., Belenky, D., Lichtenstein, M., & Cox, G. (2013). Coordinating principles and examples through analogy and self-explanation. *European Journal of Education of Psychology*, *28*, 1237-1263.

Oppenheimer, D. M., Meyvisb, T., & Davidenkoc, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867-872.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353-363.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382-407.

Regehr, G., & Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 347-363.

Rehder, B. (2007). Essentialism as a generative theory of classification. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. pp. 190-207. Oxfore, UK: Oxford University Press.

Richey, J. E., Zepeda, C. D., & Nokes-Malach, T. J. (2015). Transfer effects of prompted and self-reported analogical comparison and self-explanation. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.),

Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the classroom. *Science*, *316*, 1128-1129.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogy* (pp. 21-59). Cambridge: Cambridge University Press.

Rittle-Johnson, B., Loehr, A., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM*, 49, 599-611.

Rittle-Johnson, B., & Star, J. R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology, 101,* 529-544.

Rittle-Johnson, B., & Star, J. R. (2011). The power of comparison in learning and instruction: Learning outcomes supported by different types of comparisons. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of Learning and Motivation: Cognition in Education (Vol. 55)*. Elsevier.

Rosch, E., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573-605.

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research*, *77*, 534-574.

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

Rozenblit, L. R., & Keil, F. C. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science, 26,* 521-562.

Sagi, E., Gentner, D. & Lovett, A. (2012). What difference reveals about similarity. *Cognitive Science*, *36*, 1019-1050.

Sidney, P. G., Hattikudur, S., & Alibali, M. W. (2015). How do contrasting cases and self-explanation promote learning? Evidence from fraction division. *Learning and Instruction*, *40*, 29-38.

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

Sloman, S. A., & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition*, *65*, 87-101.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

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

Spalding, T. L., & Ross, B. H. (1994). Comparison-based learning: Effects of comparing instances during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1251.

Thompson, C. A., & Opfer, J. E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development*, *81*, 1768-1786.

van Fraassen, B. C. (1980). The scientific image. New York: Oxford University Press.

Walker, C., Bonawitz, E., & Lombrozo, T. (2017). Effects of explaining on children's preference for simpler hypotheses. *Psychonomic Bulletin and Review,* 1538-1547.

Walker, C. & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition,* 167, 266-281.

Walker, C.M., Lombrozo, T., Legare, C., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition, 133*, 343-357.

Walker, C., Lombrozo, T., Williams, J. J., Rafferty, A., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child Development, 88*, 229-246.

Wellman, H. M., & Liu, D. (2007). Causal reasoning as informed by the early development of explanations. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 261-279). New York: Oxford University Press.

Wilkenfeld, D. A., & Lombrozo, T. (2015). Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science and Education*, 1-19.

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

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

Williams J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General, 142*, 1006-1014.

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 and Instruction*, *12*, 233-262.