Supplemental Experiment A

In Experiment 1 in the paper, the within-category pairwise comparison prompts made participants less likely to discover a 100% categorization rule. One possible reason for the ineffectiveness of comparison prompts was that items were highly similar, and comparison prompts were restricted to comparisons of items *within* the same category. In Supplemental Experiment A, we examined whether a between-category pairwise comparison task would lead to higher rates of 100% rule discovery. Previous work has suggested that between-category comparison may be more helpful than within-category comparison for learning categories whose within-category members are highly similar (Higgins & Ross, 2011, see also Andrews, Livingston, & Kurtz, 2011; Ankowski, Vlach, & Sandhofer, 2013).

Additionally, Supplemental Experiment A included an assessment of participants' memory for the studied robots. Prior work suggests that explaining can impair memory for individual items relative to description (Williams & Lombrozo, 2010) and thinking aloud (Williams, Lombrozo, & Rehder, 2013), but to our knowledge differential effects of explanation and comparison on memory for item features have not been investigated.

Method

Participants. Participants were 159 adults recruited from Amazon Mechanical Turk and tested online. An additional 56 participants were tested, but excluded from the analyses. The exclusion criteria were the same as in Experiment 1.

Materials. The stimuli were the same eight robots used in Experiment 1. Additionally, eight new robots were used as part of a memory assessment, as described below.

Procedure. The procedure consisted of a *study phase*, a *rule-reporting phase*, a *memory assessment,* and *end-of-study questions*.

In the *study phase*, each participant was randomly assigned to respond to (1) within-category pairwise comparison prompts, (2) between-category pairwise comparison prompts, (3) explanation prompts with the study order matched to the within-category comparison condition, or (4) explanation prompts with the study order matched to the between-category comparison condition. The between-category comparison prompts were of the form "What are the *similarities and differences* between Glorp robot X and Drent robot Y?" All other prompts were identical to those in Experiment 1.

As in Experiment 1, participants studied the robots that were consistent with respect to both 75% rules first. Thus, the study order for the between-category comparison condition was A and H, B and F, C and G, and D and E. The study order for the matched explanation condition was identical, except that participants studied the robots one at a time. The study orders for the within-category comparison condition and matching explanation condition were identical to the respective conditions in Experiment 1.

In all conditions, the total study time was reduced to 360 seconds. Participants in the explanation conditions were allotted 45 seconds for each explanation and participants in the comparison conditions were allotted 90 seconds for each comparison. As in Experiment 1, participants automatically advanced to the next explanation or comparison once the time elapsed and could not advance before then. The picture of the eight study robots remained visible for the entire study time. Similar to Experiment 1, participants received a simple math problem in-between prompts every 90 seconds as a "catch trial" to determine whether participants were paying attention.

The *rule-reporting phase* was identical to Experiment 1. After completing the rule-reporting phase, but before responding to the end-of-study questions, participants performed a *recognition memory assessment*. In this task, participants were presented with eight pairs of robots; each pair tested participants' memory for a different study robot. The order of these pairs was randomized across participants. In each pair, one of the robots was from the original eight in the study phase and the other robot was novel. Each novel robot was identical to its corresponding study robot except for a single feature change (e.g., pentagonal feet instead of heart-shaped feet, or elbows instead of knees). For each pair of robots, participants were asked which of the two robots appeared in the study phase. Participants then rated their confidence in their response on a 1-7 scale with one-point intervals, where 1 = "not at all confident" and 7 = "extremely confident."

The *end-of-study questions* again contained self-report questions asking participants to report the extent to which they engaged in each of several cognitive processes when studying the robots. The self-report questions were as follows: "Regardless of the task instructions, to what extent did you notice yourself […] when the image of the eight robots was on-screen?" We asked separate questions for each of the following processes: explanation, within-category pairwise comparison, between-category pairwise comparison, and description. The respective prompts were "explaining what makes particular robots Glorp robots or Drent robots," "making comparisons between pairs of Glorp robots and/or pairs of Drent robots," "making comparisons between specific Glorp robots and specific Drent robots," and "describing the features of specific Glorp robots and specific Drent robots." The remaining end-of-study questions (age, gender, previous participation, catch trial) were identical to those in Experiment 1.

Results

As in Experiment 1, we analyzed (1) the self-report data, (2) whether participants discovered at least one 100% rule, and (3) whether participants discovered at least one 75% rule, and additionally, we analyzed (4) performance on the memory assessment. The results are shown in Figure S1 and Table S1.

Self-reports. We analyzed the self-report data by performing a series of 2 × 2 ANOVAs with *study prompt* (explanation vs. comparison) and *study order* (between-category vs. within-category) as between-subjects factors and *self-reported explanation*, *self-reported between-category comparison*, *self-reported within-category comparison*, *self-reported total comparison* (self-reported between-category comparison + self-reported within-category comparison), and *self-reported description* as the dependent measures (see Table S1).

These analyses replicated the results of the corresponding analyses from Experiment 1, including the rather counterintuitive finding that explanation prompts were apparently more effective at stimulating comparison processing than were comparison prompts. Participants in the explanation conditions reported doing more explanation, $F(1, 151) = 27.2$, $p < .001$, and more comparison – including between-category comparison, $F(1, 152) = 4.37$, $p = .038$, within-category comparison, $F(1, 152) = 7.25$, $p = .008$, and total comparison, $F(1, 151) = 7.47$, $p = .007$) – as well as marginally more description, $F(1, 152) = 3.07$, $p = .082$, than participants in the comparison conditions. There were no significant effects of study order on any of the dependent measures. In subsequent analyses, unless there were differential results for between-category and within-category comparison, we only report the analyses for the total amount of self-reported (between-category + within-category) comparison.

Study Condition	Self-reported Explanation Mean (SD)	Self-reported Comparison (Between-Category) Mean (SD)	Self-reported Comparison (Within-Category) Mean (SD)	Self-reported Description Mean (SD)
Comparison (Between)	3.42 (2.10)	4.36 (1.97)	4.15 (2.06)	3.92 (2.18)
Comparison (Within)	4.18 (1.68)	4.15 (1.69)	3.90 (1.77)	4.60 (1.52)
Explanation (Between)	5.41 (1.36)	4.98 (1.62)	4.60 (2.10)	4.78 (2.04)
Explanation (Within)	5.14 (1.84)	5.00 (1.45)	4.76 (1.83)	4.81 (1.76)

Table S1: Self-reported explanation, comparison, and description by participants in each study condition in Supplemental Experiment A.

Discovery of one or more 100% rules. We performed a log-linear analysis of *study prompt* (explanation vs. comparison) × *study order* (between-category vs. within-category) × *discovered at least one 100% rule* (yes vs. no). Participants who received an explanation prompt were significantly more likely to discover at least one 100% rule than participants who received a comparison prompt, $\chi^2(1) = 10.1$, $p = .001$. However, there was not a significant effect of whether participants did between-category or within-category study on discovery of at least one 100% rule, $\chi^2(1) = 0.02$, $p = .89$.

We also analyzed whether self-reported explanation processing, comparison processing, and description were positively correlated with the discovery of at least one 100% rule. As in Experiment 1, a logistic regression of *discovered at least one 100% rule* (yes vs. no) on the *amount of self-reported explanation* showed that the amount of self-reported explanation was positively associated with discovery of at least one 100% rule, $W(1) = 8.16$, $p = .004$. Similarly, we performed separate logistic regressions of *discovered at least one 100% rule* (yes vs. no) on the *amounts of self-reported (1) between-category comparison, (2) within-category comparison, (3) total comparison, and (4) description*; however, these analyses found no significant effects: between-category comparison: $W(1) = 1.10$, $p = .29$; within-category comparison: $W(1) = 0.42$, $p = .52$; total comparison: $W(1) = 0.04$, $p = .85$; description: $W(1) = 0.92$, $p = .34$.

These data replicate the results of Experiment 1 in that participants who were given explanation prompts were more likely to discover at least one 100% rule than participants who were given comparison prompts. Additionally, we did not find any significant effects of whether participants were prompted to do between-category pairwise comparison vs. within-category pairwise comparison. In Experiment 2, we asked whether prompting *both* within- and between-comparisons could be more effective.

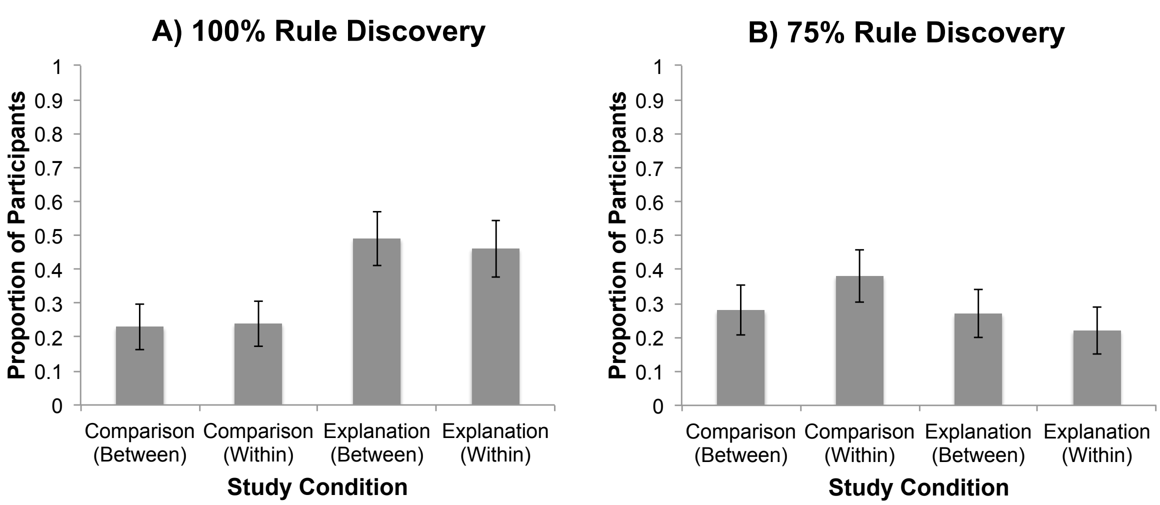

Figure S1: Proportion of participants in each study condition who discovered at least one 100% rule (Fig. S1A) and at least one 75% rule (Fig. S1B) in Supplemental Experiment A. Error bars indicate +/- 1 SE.

Discovery of one or more 75% rules. A log-linear analysis of *study prompt* (explanation vs. comparison) × *study order* (between-category vs. within-category) × *discovered at least one 75% rule* (yes vs. no) found no significant effects of study task, $\chi^2(1) = 1.52$, $p = .22$, or study order, $\chi^2(1) = 0.12$, $p = .73$.

Memory assessment. We evaluated whether performance on the memory assessment varied across conditions. A t-test with *study prompt* (explanation vs. comparison) as the independent variable and *number of memory items correct* as the dependent variable found that memory for the eight study robots did not significantly differ across conditions, $t(155) = 1.80$, $p = .18$. This may be in part because performance was quite low in all conditions ($M = 4.89$ of 8, $SD = 1.32$).

Discussion

In sum, Supplemental Experiment A found that explanation prompts once again promoted comparison (as assessed by self-reports) as well as discovery of 100% rules, and that the learning benefits of explanation prompts exceeded those of comparison prompts. Again, we found that comparison prompts were ineffective in elevating levels of self-reported comparison, and that explanation prompts were more effective at doing so. Supplemental Experiment A also went beyond Experiment 1 to test the relative effects of within-category versus between-category pairwise comparisons, finding no reliable differences on our task.

Supplemental Experiment B

Supplemental Experiment A found that prompting between-category pairwise comparison was no more effective than prompting within-category pairwise comparison; explanation led to performance that surpassed both comparison conditions, which did not differ from each other. Further, the self-report results showed that, as in Experiment 1, explanation prompts were more effective at increasing comparison processing than were comparison prompts. In Supplemental Experiment B, we asked whether the comparison prompts in Experiment 1 and Supplemental Experiment A were ineffective because they restricted participants' comparisons to the particular robots mentioned in the prompts. Accordingly, Supplemental Experiment B explored whether allowing participants to choose which pairs of robots to compare would increase categorization rule discovery.

In addition, Supplemental Experiment B added a dependent measure in which participants categorized novel robots. We included this measure to ensure that effects of study condition on our primary measure of categorization rule discovery (i.e., explicit report) truly reflected learning, not only participants' ability to *verbalize* what they learned.

Method

Participants. Participants were 243 adults recruited from Amazon Mechanical Turk and tested online. An additional 84 participants were tested, but excluded from the analyses. The exclusion criteria were the same as in the previous experiments.

Materials. The stimuli were the eight robots used in the previous experiments. An additional 16 robots were used in a novel robot categorization task, as described below.

Procedure. The procedure consisted of a *study phase*, a *rule-reporting phase*, a *novel robot categorization task*, and *end-of-study questions*.

In the *study phase*, each participant was randomly assigned to study the robots in one of four ways: (1) pairwise comparison with the study order chosen by the experimenter, (2) pairwise comparison with the study order chosen by the participant, (3) explanation with the study order chosen by the experimenter, or (4) explanation with the study order chosen by the participant.

Participants in all conditions began by engaging in free study for 60 seconds. The purpose of the free-study period was to expose participants to the robot stimuli to help participants in the "study order chosen by the participant" conditions later devise a strategy regarding which robots to study and in which order. During this time, participants viewed the picture of the eight robots with the following text below the picture: "Study the eight robots that are on-screen and think about how you might categorize robots as either Glorp robots or Drent robots." Participants did not provide a written response during the free-study period.

After completing the 60 seconds of free study, participants engaged in 360 seconds of condition-specific study as follows. At the beginning of the condition-specific study, participants in all conditions were told that "You will now have six additional minutes to study the robots. This time will be divided into eight 45-second intervals." This was followed by condition-specific instructions, which are described below. In all conditions, the picture of the eight study robots was on-screen for the duration of the study phase.

Pairwise comparison with study order chosen by experimenter condition. Participants received the following instructions: "In each of the eight intervals, you will be asked to compare pairs of robots (i.e., list similarities and differences between these robots). In each interval, the specific pairs of robots that you will compare will be chosen for you and identified on-screen." After reading these instructions, participants studied the robots by responding to a combination of between-category and within-category pairwise comparison prompts. The between-category and within-category pairwise comparison prompts were the same prompts used in Supplemental Experiment A. The order of the comparisons was four within-category comparisons (A and B, F and H, C and D, E and G) followed by four between-category comparisons (A and H, B and F, C and G, D and E).

Pairwise comparison with study order chosen by participant condition. Participants received the following instructions for choosing comparison pairs. “In each of the eight intervals, you will be asked to compare pairs of robots (i.e., list similarities and differences between these robots). In each interval, you will choose which specific pair of robots to compare. In subsequent intervals, you may choose to compare one or more robots that you have already compared, or to compare robots that you have not yet compared. In each interval, begin by writing the letters of the robots that you will compare in the space provided. Then respond to the comparison prompt in the larger text box below.” The specific comparison prompt was “*Compare* the two robots listed above (i.e., what are the *similarities and differences* between these robots?).”

Explanation with study order chosen by experimenter condition. Participants received the following instructions: “In each of the eight intervals, you will be asked to try to explain why individual robots are members of a particular category. In each interval, the specific robots that you will explain will be chosen for you and identified on-screen.” Participants studied the robots by responding to the explanation prompts used in Experiment 1 and Supplemental Experiment A. The study order was matched to the within-category comparison order (A, B, F, H, C, D, E and G).

Explanation with study order chosen by participant condition. Participants received the following instructions: “In each of the eight intervals, you will be asked to try to explain why individual robots are members of a particular category. In each interval, you will choose the specific robot that you will explain. In subsequent intervals, you may choose to explain robots that you have already explained, or to explain robots that you have not yet explained. In each interval, begin by writing the letter of the robot that you wish to explain in the space provided. Then respond to the explanation prompt in the larger text box below." The specific explanation prompt was "Try to *explain why* the robot listed above is a member of its particular category."

As in the previous experiments, simple math problems were inserted as catch trials in-between prompts after each 180 seconds of study.

After participants studied the robots in one of these four ways, they moved on to the *rule-reporting phase*. The rule-reporting phase was the same as in Experiment 1 and Supplemental Experiment A.

After the rule-reporting phase, participants completed a *novel robot categorization task*. In this task, participants were presented with eight pairs of novel robots, with the order of the pairs randomized across participants. Participants were told that each pair consisted of one Glorp robot and one Drent robot. For each pair, participants were randomly asked either to identify which robot was a Glorp or which robot was a Drent.

Each pair of robots tested whether participants had learned one of the four categorization rules, with two pairs testing each rule. In each pair, the two robots were identical except for the feature corresponding to the specific categorization rule being tested. For example, each pair testing whether participants learned the antenna rule included one robot with a right antenna that was taller than the left antenna and one robot with a left antenna that was taller than the right antenna. Apart from this, the two robots were identical; that is, they had the same color patterns, foot shape, body shape, and both elbows and knees. The robots were given values of these features that were neutral with respect to the eight study examples (Foot shape: feet with curved bottoms, Relative antenna length: antennae of equal lengths, Body shape: trapezoidal bodies, Elbows vs. knees: both elbows and knees or neither elbows nor knees, Color pattern: color combinations not used in any of the study robots). We used neutral values to avoid giving Glorp robots features that were associated with Drent robots and vice versa.

Finally, participants completed a series of *end-of-study questions*. As in Experiment 1 and Supplemental Experiment A, in the *self-report* questions we asked participants to report how much explanation and comparison they had performed; however, the wording of these questions was revised slightly. In Supplemental Experiment B, 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 the previous experiments.

Results

We analyzed (1) the self-report data, (2) whether participants discovered at least one 100% rule, (3) whether participants discovered at least one 75% rule, (4) performance on the novel robot categorization task, and (5) the types of comparisons performed. The results are shown in Figure S2 and Table S2.

Self-reports. We performed a series of 2 × 2 ANOVAs with *study prompt* and *robot selection method* as between-subjects factors and (1) self-reported explanation, (2) self-reported between-category pairwise comparison, and (3) self-reported within-category pairwise comparison as the dependent variables. These analyses evaluated the effects of each study condition on the extent to which participants engaged in each of these cognitive processes when studying the robots (see Table S2). Participants who were given explanation prompts reported doing more explanation when studying the robots than participants who were given comparison prompts, $F(1, 236) = 17.0$, $p < .001$; however, there were no other significant effects. In particular, the amount of reported comparison processing did not differ across the explanation and comparison conditions, $F(1, 241) = 0.49$, $p = 0.49$. Next, we analyzed whether the proportion of participants discovering each rule type varied across conditions (see Figure S2).

Study Condition	**Self-reported Explanation Mean (SD)**	**Self-reported Within-Category Comparison Mean (SD)**	**Self-reported Between-Category Comparison Mean (SD)**
Explanation: Study Order Chosen by Experimenter	5.52 (1.39)	5.36 (1.53)	5.92 (1.23)
Explanation: Study Order Chosen by Participant	5.37 (1.72)	4.97 (2.04)	5.55 (1.57)
Comparison: Study Order Chosen by Experimenter	4.69 (2.13)	5.03 (1.79)	5.76 (1.43)
Comparison: Study Order Chosen by Participant	4.25 (1.99)	4.87 (1.80)	5.76 (1.18)

Table S2: Self-reported explanation and comparison in each study condition in Supplemental Experiment B.

Discovery of one or more 100% rules. A log-linear analysis of *study prompt* (explanation vs. comparison) × *robot selection method* (chosen by experimenter vs. chosen by participant) × *discovered a 100% rule* (yes vs. no) found that explanation participants were significantly more likely to discover a 100% rule than comparison participants, $\chi^2(1) = 8.20$, $p =$ 0.004. There was no significant effect of robot selection method on 100% rule discovery, $\chi^2(1)=$ 0.00, $p = .98$.

These data replicate the finding in Experiment 1 and Supplemental Experiment A that performing the explanation task increases discovery of at least one 100% rule. The data also argue against the possibility that the comparison instructions in Experiment 1 or Supplemental Experiment A led to low 100% rule discovery because participants constrained themselves to the specific comparison pairs mentioned in the study prompts; there were no effects of whether the comparison pairs were selected by participants or by the experimenter on categorization rule discovery.

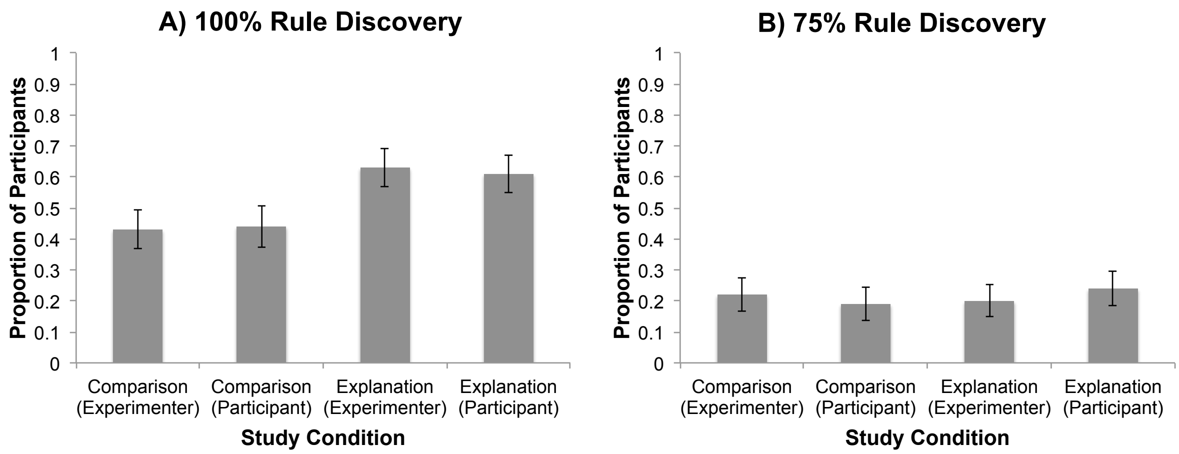

Figure S2: Proportion of participants in each study condition who discovered at least one 100% rule (Fig. S2A) and at least one 75% rule (Fig. S2B) in Supplemental Experiment B. Error bars indicate +/- 1 SE.

As in Experiment 1 and Supplemental Experiment A, we examined whether self-reported explanation and comparison processing were correlated with discovery of at least one 100% rule with separate logistic regressions of *discovered at least one 100% rule* (yes vs. no) on the amounts of (1) *self-reported explanation* and (2) *self-reported comparison*. In this experiment, there were no significant effects of reported explanation processing, $W(1) = 0.81$, $p = .37$, in contrast to the results of Experiment 1 and Supplemental Experiment A. There were also no effects of reported comparison processing on the discovery of at least one 100% rule, $W(1) = 0.69$, $p = .41$, in contrast to the results of Experiment 1.

Discovery of one or more 75% rules. A log-linear analysis of *study prompt* (explanation vs. comparison) × *robot selection method* (chosen by experimenter vs. chosen by participant) × *discovered a 75% rule* (yes vs. no) did not find a significant effect of either study prompt, $\chi^2(1) = 0.10$, $p = .75$, or robot selection method, $\chi^2(1) = 0.00$, $p = .96$.

Novel robot categorization task. We analyzed performance on the novel robot categorization task by seeing whether participants who articulated each categorization rule in the rule-reporting phase were more likely to accurately categorize novel robots that could only be distinguished using that rule. For both 100% rules, participants who articulated the categorization rule correctly classified more novel robots (out of two) than participants who did not discover that rule (Foot rule: Participants who articulated rule: $M = 1.79$, $SD = 0.48$, Participants who did not articulate rule: $M = 0.84$, $SD = 0.69$, $t(241) = 109$, $p < .001$; Antenna rule: Participants who articulated rule: $M = 1.77$, $SD = 0.56$, Participants who did not articulate rule: $M = 1.17$, $SD = 0.74$, $t(241) = 44.7$, $p < .001$). Performance on the rule-reporting task and the novel robot categorization task were also associated for the 75% body shape rule (participants who articulated rule: $M = 1.50$, $SD = 0.75$, participants who did not articulate rule: $M = 1.13$, $SD = 0.74$), $t(241) = 9.08$, $p = .003$, but not for the 75% elbows and knees rule (participants who articulated rule: $M = 1.36$, $SD = 0.85$, participants who did not articulate rule: $M = 1.12$, $SD = 0.78$), $t(241) = 1.89$, $p = .29$. The trends may have been weaker for the 75% rules because the novel robot categorization task measured rule *generalization* as opposed to rule *discovery*, and participants may have been less like to generalize from the 75% rules since some of the study robots were anomalous with respect to these rules. Importantly, however, these data provide evidence that performance on the primary dependent measure used in these experiments, the ability to articulate categorization rules, is correlated with performance on a second key measure of category learning (and one that doesn't require explicit verbalization), the ability to transfer their learning to classify novel robots.

Types of comparisons performed. Finally, we examined the kinds of pairs that participants compared when allowed to select robot pairs themselves.[1] Participants showed a significant preference for between-category comparison, with participants performing an average of 7.07 between-category comparisons and 0.72 within-category comparisons (of 8 total), $t(53) = 15.0$, $p < .001$. In fact, only 13 out of 54 participants performed even a single within-category comparison. This trend was especially strong for the first comparison pair, in which 46 participants chose to compare a between-category pair, whereas only three participants chose to compare a within-category pair. These data suggest that for this task, most people spontaneously adopt a between-category comparison strategy instead of a within-category comparison strategy or a mixed strategy.

Discussion

Across the studies so far (Experiment 1, Supplemental Experiment A, Supplemental Experiment B), we have evidence that prompting participants to explain can increase the extent to which they engage in both explanation and comparison (as assessed by self reports), and that for our particular task configuration, explanation prompts are highly effective in promoting discovery of 100% rules. In contrast, comparison prompts have been strikingly ineffective here. First, participants given comparison prompts have shown relatively poor ability to discover 100% rules. Second, whereas prompting people to explain increased the amount of self-reported explanation processing (*and* self-reported comparison processing, in Experiment 1 and Supplemental Experiment A), prompting people to compare did neither, regardless of whether they were instructed to compare within-category pairs or between-category pairs.

---

[1] Because some participants did not specify which robots they were comparing, the sample sizes vary slightly across these analyses.

Of course, as noted in the main paper, some aspects of the task may have led to a very high level of baseline comparison processing. All the robots were highly similar perceptually, and all eight robots were displayed on the screen together during the study phase. These factors may have increased spontaneous comparison, making it difficult to detect effects of the prompt to compare. However, we do find that explanation prompts increase the amount of self-reported comparison, suggesting that a baseline ceiling effect is not the whole answer. In Experiment 2 (in the main paper), we consider another possibility. Perhaps the emphasis on pairwise comparison in our instructions led participants to focus on local pairs, rather than think about the global category structure. To test this possibility, Experiment 2 investigated whether prompting group comparison (i.e., asking participants to compare *all four* category members) would be more effective than prompting pairwise comparison.

References

Andrews, J. K., Livingston, K. R., & Kurtz, K. J. (2011). Category learning in the context of co-presented items. *Cognitive Processing*, *12*, 161-175.

Ankowski, A. A., Vlach, H. A., & Sandhofer, C. M. (2013). Comparison versus contrast: Task specifics affect category acquisition. *Infant and Child Development*, *22*, 1-23.

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.

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

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