Effects of Explanation and Comparison on Category Learning

Brian J. Edwards (Brian.Edwards@U.Northwestern.Edu)
Department of Psychology, Northwestern University, 2029 Sheridan Road-102 Swift Hall
Evanston, IL 60208 USA

Joseph J. Williams (Joseph_Williams@Berkeley.Edu), Tania Lombrozo (Lombrozo@Berkeley.edu)
Department of Psychology, University of California, Berkeley, 3210 Tolman Hall
Berkeley, CA 94720 USA

Abstract

Generating explanations and making comparisons have both been shown to improve learning. While each process has been studied individually, the relationship between explanation and comparison is not well understood. Three experiments evaluated the effectiveness of explanation and comparison prompts in learning novel categories. In Experiment 1, participants explained items' category membership, performed pairwise comparisons between items (listed similarities and differences), did both, or did a control task. The explanation task increased the discovery of rules underlying category membership; however, the comparison task decreased rule discovery. Experiments 2 and 3 showed that (1) comparing all four category exemplars was more effective than either within-category or between-category pairwise comparisons, and that (2) "explain" participants reported higher levels of both spontaneous explanation and comparison than "compare" participants. This work provides insights into when explanation and comparison are most effective, and how these processes can work together to maximize learning.

Keywords: Explanation; comparison; categorization; learning.

Introduction

Explanation (i.e., answering "why" questions) and comparison (i.e., describing the similarities and differences between entities) are both powerful learning processes. Although they have typically been studied independently, they are often interconnected. Asking people to generate explanations can invite implicit comparison, and the patterns that people discover by comparing can motivate a search for explanations. For example, explaining why someone prefers coffee versus tea might lead one to identify similarities and differences between the two beverages, and comparing coffee and tea might provide insights into why a person would prefer one over the other. Explanation and comparison can also support similar ends: both promote abstraction and generalization, and both facilitate the discovery of patterns that are deep in a system's underlying structure (for reviews, see Gentner, 2010, on analogy and comparison; Lombrozo, 2012, on explanation).

Although explanation and comparison can generate similar effects, these two processes might rely on different cognitive mechanisms and exert different constraints on learning. Explanation has been hypothesized to improve learning through a variety of mechanisms, including an increase in metacognitive awareness (Chi, 2010) and an increase in attention and engagement (e.g., Siegler, 2002), among others. In the context of category learning, generating explanations also enables learners to generalize beyond a specific set of observed data. In particular, Williams and Lombrozo (2010, 2013) proposed a *subsumptive constraints* account of how explanation impacts learning, whereby explaining leads people to interpret individual cases as part of a general pattern. As a result, explanation can help people unify multiple observations and focus on patterns with broader scope, increasing the discovery of rules that account for 100% of the data versus only 75% (Williams & Lombrozo, 2010).

One mechanism by which comparison has been hypothesized to support learning is by promoting explicit structural alignment, leading people to focus on *alignable differences* between two entities (i.e., differences that are embedded in a common relational structure) (Gentner, 1983; Gentner & Markman, 1997). Since comparison causes people to analyze these differences in the context of the common structure, comparison can illuminate deeper similarities and support the formation of an abstract relational schema, even (and especially) when the items being compared have surface differences (Gentner et al., 2009). For example, the analogy "an atom is like a solar system" highlights the fact that an atom consists of electrons orbiting around a nucleus, whereas a solar system consists of planets orbiting around the sun. Across a number of domains, comparing two examples that are superficially dissimilar but share a common relational structure supports transfer more effectively than studying the same examples separately (e.g., Kurtz, Miao, & Gentner, 2001; Loewenstein, Thompson, & Gentner, 2003).

Despite the abundance of research showing that explanation and comparison can (individually) enhance learning, few studies have investigated the effects of both explanation and comparison *on the same experimental task*. Kurtz, Miao, and Gentner (2001) found that comparing two analogous examples of heat flow helped participants discover similarities between the two examples more effectively than describing and explaining the same examples sequentially. Additionally, comparison was most effective when participants performed a task that involved listing which elements of the second scenario corresponded to specific elements of the first scenario. In another study, Nokes-Malach et al. (2012) found that introductory physics

students who explained the solutions to worked examples of physics problems achieved greater "near" transfer than participants who compared pairs of problems, but both groups performed similarly on "far" transfer and outperformed participants in a control condition. While these studies provide valuable insights into the conditions under which explanation and comparison are most effective, many questions remain open.

The present studies examine whether and how explanation and comparison interact to support learning novel categories. Previous work using similar materials (alien robots) has found that relative to control conditions, participants prompted to explain why individual robots belong to particular categories are more likely to discover a categorization rule that accounts for all cases (Williams & Lombrozo, 2010, 2013). The present studies extend this work by investigating whether having participants compare robots also facilitates category learning, and whether participants who perform *both* explanation and comparison tasks are more likely to discover categorization rules than participants who perform only one of these tasks.

We hypothesize that comparison and explanation play complementary roles in category learning. Comparison may be crucial for identifying *similarities* among members of the same category and *differences* between members of different categories. In contrast, explanation should encourage learners to seek *broad patterns* within and across categories, potentially drawing upon the similarities and differences identified through comparison. Very broadly, these hypotheses predict that participants should be more effective in discovering categorization rules to the extent that they both compare and explain, with explanation being especially important in discovering broad patterns.

Three experiments evaluated these predictions. In Experiment 1, participants were randomly assigned to study the robots in one of four ways: (1) explain why individual robots are members of a particular category, (2) compare pairs of robots that belong to the same category, (3) perform both the explanation and comparison tasks, or (4) engage in a "free study" control task. Experiments 2 and 3 evaluated the effectiveness of different types of comparison prompts: between-category pairwise comparison and "group" comparison, respectively. We included a "group" comparison prompt to see whether it would be more effective at improving participants' ability to integrate pairwise comparisons and detect broad patterns.

Experiment 1

Method

Participants One-hundred-sixty-one adults participated through the Amazon Mechanical Turk marketplace. An additional 56 participants were tested, but excluded because they failed a catch trial or had previously completed a similar experiment. Participants were paid for participation.

Materials The stimuli (see Fig. 1) were eight robots adapted from Williams and Lombrozo (2010, 2013). Four robots (A-D) were classified as Glorp robots and the other four robots (E-H) were classified as Drent robots.

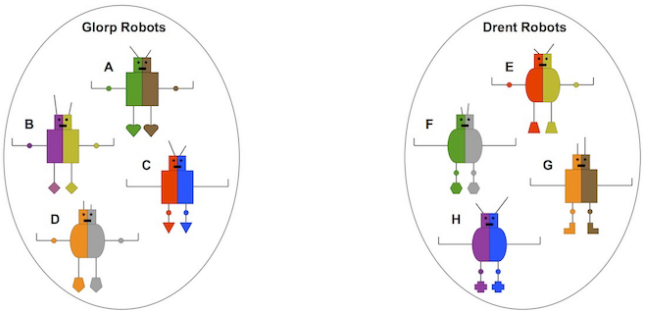

Figure 1: Robots used in Exp. 1-3

Four rules could be used to categorize robots as either Glorp robots or Drent robots. Two rules were "100% rules" that could be used to categorize all eight robots and two rules were "75% rules" that could be used to categorize six of the eight robots (i.e., two robots were 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; all Drent robots have feet with flat bottoms.

(2) Antenna rule (100%): All Glorp robots have a right antenna (from the robot's perspective) that is longer than the left antenna; all Drent robots have a left antenna that is longer than the right antenna.

(3) Elbows/knees rule (75%): Three out of four Glorp robots (A, B, D) have elbows but no knees; three out of four Drent robots (F, G, H) have knees but no elbows. One Glorp robot (C) has knees but no elbows and one Drent robot (E) has elbows but no knees.

(4) Body shape rule (75%): Three out of four Glorp robots (A, B, C) have a rectangular body; three out of four Drent robots (E, F, H) have a round body. One Glorp robot (D) has a round body and one Drent robot (G) has a rectangular body.

The robots also differed in body color; however, there were no systematic category differences in body color.

Procedure The procedure consisted of a *study phase* followed by a *rule-reporting phase*.

In the *study phase*, each participant was assigned to one of four study conditions: (1) comparison only, (2) explanation only, (3) both explanation and comparison, or (4) free study. In every condition, all eight robots appeared on screen for the duration of the study phase, as shown in Figure 1. The total study time (640 seconds) was equal across conditions. The study prompts and procedures for each condition were as follows.

Comparison only condition: "What are the *similarities and differences* between Glorp [Drent] robot X and Glorp [Drent] robot Y?" Participants were given 160 seconds to perform each comparison. The comparisons were presented

in the following order: A and B, F and H, C and D, E and G. This order was chosen so that the four robots that were consistent with respect to both 75% rules were studied before the four robots that were anomalous with respect to one of those rules, making it more likely that participants would learn the 75% rules in addition to the 100% rules.

Explanation only condition: "Try to explain *why* robot X is a Glorp [Drent] robot." Participants were given 80 seconds to provide an explanation. The explanations were requested in the following order: A, B, F, H, C, D, E, G. This matched the order in the comparison condition.

Both explanation and comparison condition: Participants responded to both the explanation and comparison prompts above. To ensure that all the conditions were matched for study time, participants were given 40 seconds to respond to each explanation prompt and 80 seconds to respond to each comparison prompt. The order of the explanation and comparison prompts was counterbalanced across participants. Participants performed both tasks sequentially for each pair (e.g., explain A, explain B, compare A and B) before moving on to study the next pair of robots. The study order was otherwise the same as in the other conditions.

Free study condition: "*Write out your thoughts* below as you learn to categorize Glorp [Drent] robot X." Participants were given 80 seconds to study each robot. The study order was the same as in the other conditions.

At the end of each study period, the screen automatically advanced to the next robot or pair of robots. Participants could not advance before the study period had elapsed.

After each 160 seconds, participants solved a simple math exercise (e.g., "9 + 7"). These exercises were included as a "catch trial" to verify that participants' attention was not diverted to other tasks. Response time was recorded and participants who took more than one minute to answer a question were excluded from analysis.

In the *rule-reporting phase*, participants listed the patterns they noticed "that might help differentiate Glorps and Drents." These responses were classified by a coder who was blind to experimental condition. Twenty-five percent of the data was independently coded for reliability by a second blind coder; agreement for each experiment exceeded 95%. For each pattern that participants discovered, they also indicated (1) how many of the eight study robots could be categorized using that pattern and (2) how many new Glorp and Drent robots (out of 100) could be categorized using that pattern. Because answers to these two questions were contingent on the participant having discovered a particular rule, the sample sizes were relatively small and these data are not discussed further.

After completing the rule-reporting phase, participants answered debriefing questions regarding the extent to which they (1) generated explanations and (2) made comparisons, regardless of the task instructions, using a numerical response on a 1-7 scale, where 1 indicated "not at all" and 7 indicated "all of the time." Participants were then asked whether they had previously completed a similar study and answered a "catch trial" adapted from Oppenheimer, Meyvisb, and Davidenkoc (2009) to find out whether they were reading the instructions. Participants who reported previously doing a similar study and participants who failed the catch trial were excluded from analysis.

Results and Discussion

We first considered whether study task influenced the total number of rules discovered. A 2 × 2 ANOVA with *explain prompt* (yes/no) and *compare prompt* (yes/no) as between-subjects factors and total number of rules discovered (0-4) as a dependent measure revealed no effects of condition (*p*s > .15). We thus considered whether discovery of the 100% and 75% rules varied across study conditions (see Fig. 2).

A log-linear analysis of *explain prompt (yes/no)* × *compare prompt (yes/no)* × *discovered a 100% rule (yes/no)* revealed that performing the explanation task made participants significantly more likely to discover at least one of the two 100% rules, $\chi^2(1) = 21.4$, $p < .001$. Performing the comparison task had the opposite effect: participants were *less* likely to discover a 100% rule, $\chi^2(1) = 5.90$, $p = .015$. There was no significant interaction ($p = .67$). A comparable analysis on discovery of a 75% rule (yes/no) found that performing the explanation task made participants *less* likely to report a 75% rule, $\chi^2(1) = 11.3$, $p < .001$, with no effect of the comparison task, $p = .75$.

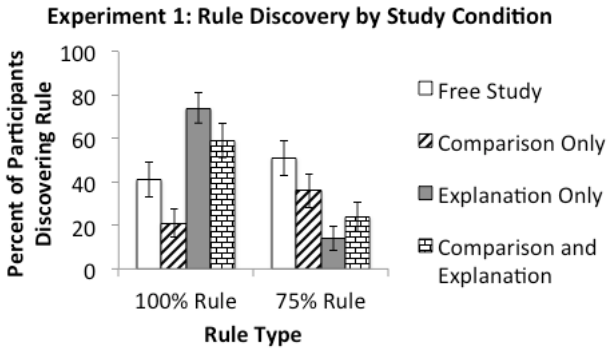

Figure 2: Rule discovery by study condition in Exp. 1, showing the percent of participants discovering at least one rule of each type.

These findings challenge our predictions in that a comparison prompt actually *impaired* 100% rule discovery, and that explanation and comparison did not have additive benefits. The findings do support the idea that explanation and comparison exert distinct constraints on learning, but raise an important puzzle: why didn't comparison – which has been shown to have robust and beneficial effects in other domains – improve performance on this task? We analyzed participants' self-reported explanation and comparison to better understand why the comparison task impaired performance, and in particular, whether the study prompts were effective at promoting explanation and comparison processes as intended.

A 2 × 2 ANOVA with *explanation task (yes/no)* and *comparison task (yes/no)* as between-subjects factors and

amount of self-reported explanation as the dependent variable showed that participants who performed the explanation task reported more explanation (M = 5.83, SD = 1.46) than participants who did not (M = 4.36, SD = 2.09), $F(1, 154) = 27.7$, $p < .001$. Additionally, participants who performed the comparison task reported doing *less* explanation (M = 4.75, SD = 2.06) than participants who did not (M = 5.49, SD = 1.73), $F(1, 154) = 7.26$, $p = .008$. Self-reported explanation was positively correlated with the number of 100% rules discovered, $r = .34$, $p < .001$.

A 2 × 2 ANOVA with *explanation task (yes/no)* and *comparison task (yes/no)* as between-subjects factors and *amount of self-reported comparison* as the dependent variable showed that participants who performed the explanation task reported doing *more* comparison (M = 5.62, SD = 1.64) than participants who did not (M = 4.69, SD = 2.05), $F(1, 155) = 10.2$, $p = .002$. However, performing the comparison task did not affect the amount of reported comparison (Comparison: M = 5.13, SD = 1.92; No comparison: M = 5.21, SD = 1.89). Self-reported comparison was positively correlated with the number of 100% rules discovered, $r = .22$, $p = .006$, but the effect was not significant after controlling for reported explanation.

Two factors might help explain why the comparison task did not support discovery of the 100% rules. First, the comparison prompt failed to boost overall comparison (as reflected in self-reports), and additionally decreased self-reported explanation, which was beneficial to learning. Second, the comparison prompt may have constrained the particular *types* of comparisons that participants performed in unhelpful ways, restricting them to within-category, pairwise comparisons at the expense of between-category comparisons or category-wide comparisons. In particular, previous work has shown that between-category pairwise comparison can be more effective than within-category pairwise comparison for learning feature-based categories (Higgins & Ross, 2011). The subsequent experiments evaluated these hypotheses by investigating whether *between-category* comparison (Experiment 2) or "*group*" comparison (Experiment 3) would support greater rule discovery than within-category pairwise comparison.

Experiment 2

Method

Participants One-hundred-sixty-one adults participated in the study through the Amazon Mechanical Turk marketplace. An additional 54 participants were tested, but were excluded because they failed a catch trial or because they had previously completed a similar experiment. Participants were paid for their participation.

Materials The stimuli were those in Experiment 1.

Procedure As in Experiment 1, the procedure consisted of a *study phase* followed by a *rule-reporting phase*.

The *study phase* was identical to Experiment 1 with the following changes. First, the total study time was reduced from 640 seconds to 360 seconds, with the time allotted for each study prompt reduced proportionally. Second, each participant was assigned to one of four study conditions: (1) the Experiment 1 explanation task, (2) the Experiment 1 within-category pairwise comparison task, (3) a between-category pairwise comparison task, or (4) an explanation task in which participants alternated explaining Glorp and Drent robots. Conditions (3) and (4) are described below.

Between-category pairwise comparison task: "What are the *similarities and differences* between Glorp robot X and Drent robot Y?" The comparisons were performed in the following order: A and H, B and F, C and G, D and E.

Between-category explanation task: This task was identical to the Experiment 1 explanation task except that the robot study order matched the between-category pairwise comparison task.

The *rule-reporting phase* was identical to Experiment 1. After the rule-reporting phase, but before the debriefing questions, participants completed a recognition memory task. However, performance was very poor and did not differ across conditions; this task is not discussed further.

After completing the memory task, participants answered debriefing questions regarding the extent to which they (1) generated explanations, (2) made within-category comparisons, (3) made between-category comparisons, and (4) described the features of individual robots, all regardless of the task instructions. As in Experiment 1, participants were asked if they had previously completed a similar experiment and answered a "catch trial" question.

Results and Discussion

We first analyzed the total number of rules discovered (0-4) in a 2 × 2 ANOVA with *study task (explain/compare)* and *study order (between/within)* as between-subjects factors. The explanation task resulted in a marginal increase in the total number of rules discovered, $F(1, 157) = 3.62$, $p = .059$.

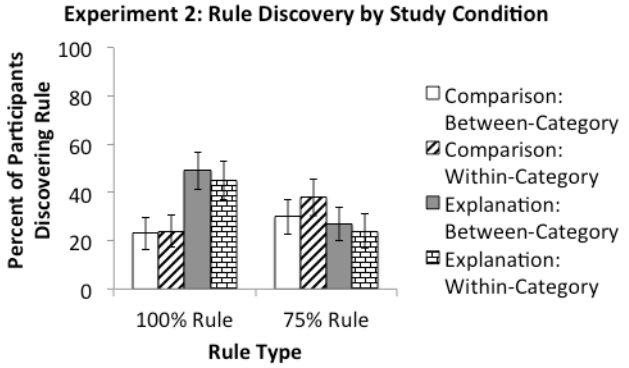

Figure 3: Rule Discovery by Condition in Exp. 2

A log-linear analysis of *study task (explain/compare)* × *study order (between/within)* × *discovered a 100% rule (yes/no)* found that participants who performed the explanation task were more likely to discover a 100% rule

than participants who performed the comparison task, $\chi^2(1) = 10.0$, $p = .002$ (see Fig. 3), with no effect of study order. An equivalent analysis on discovery of a 75% rule (yes/no) found no effect of condition, $\chi^2(1) = .57$, $p = .45$.

As in Experiment 1, we found that the explain prompt was successful in boosting self-reported explanation (relative to compare), $F(1, 149) = 26.9$, $p < .001$, but that the compare prompt was not effective in boosting self-reported comparison (between-category comparison + within-category comparison). In fact, participants prompted to explain reported significantly higher levels of total comparison that participants prompted to compare, $p = .005$.

These results suggest that the poor performance of participants prompted to compare in Experiment 1 was not due to the restriction to *within-category* comparisons. Experiment 3 thus considers whether a broader within-category comparison, one that focuses on all four items at once, might lead to better learning.

Experiment 3

Method

Participants One-hundred-ninety-three adults participated in the study through the Amazon Mechanical Turk marketplace. An additional 60 participants were tested, but were excluded because they failed a catch trial or because they had previously completed a similar experiment. Participants were paid for their participation.

Materials The stimuli were those in Experiments 1-2.

Procedure As in Experiments 1-2, the procedure consisted of a *study phase* followed by a *rule-reporting phase*.

The *study phase* was identical to Experiment 2 except that the four study conditions were as follows: (1) the explanation task from Experiments 1-2, (2) the within-category pairwise comparison task from Experiments 1-2, (3) a group comparison task in which participants simultaneously compared all four robots in each category, or (4) a group explanation task. Conditions (3) and (4) are described below. As in Experiment 2, the total study time in each condition was 360 seconds.

Group comparison task: "What are the *similarities and differences* between the Glorp robots (Robots A-D)?" After participants responded to this prompt, they received a similar prompt for the Drent robots.

Group explanation task: "Try to explain *why* robots A-D are Glorp robots." After participants responded to this prompt, they received a similar prompt for the Drent robots.

The *rule-reporting phase* was identical to Experiments 1 and 2. After completing the rule-reporting phase, participants received the same debriefing questions as in Experiment 2. No memory task was included in this study.

Results and Discussion

We first analyzed the total number of rules discovered (0-4) across each of the four study conditions. A one-way ANOVA revealed a significant difference in number of rules discovered, $F(3, 189) = 4.74$, $p = .003$. A Tukey post-hoc analysis showed that participants who performed pairwise comparisons discovered significantly fewer rules than participants who performed individual explanations ($p = .013$) or group explanations ($p = .005$), and marginally fewer rules than participants who performed group comparisons ($p = .068$).

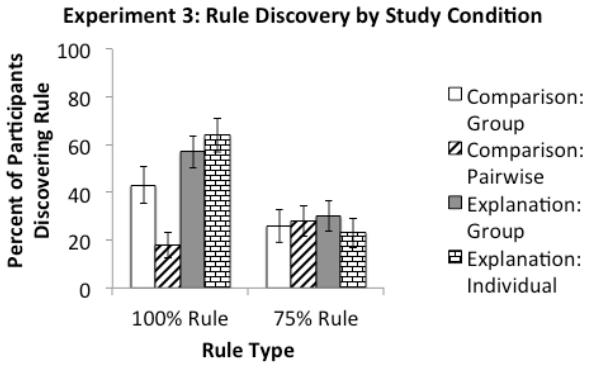

Figure 4: Rule Discovery by Study Condition in Exp. 3

We next analyzed whether the proportion of participants who discovered at least one 100% rule varied across conditions (see Fig. 4). A log-linear analysis of *study task* × *discovery of at least one 100% rule (yes/no)* found a significant effect of study task on whether participants discovered a 100% rule, $\chi^2(3) = 26.4$, $p < .001$. Additional log-linear analyses found no difference in performance between the group comparison, group explanation, and individual explanation conditions, $\chi^2(1) = 4.12$, $p = .13$; however, the pairwise-comparison prompt was significantly less effective than the other three, $\chi^2(1) = 22.3$, $p < .001$, including the group-comparison condition, $\chi^2(1) = 6.86$, $p = .009$. A log-linear analysis of *study task* × *discovered a 75% categorization rule (yes/no)* found that the study task did not affect whether participants discovered a 75% rule, $\chi^2(3) = .54$, $p = .91$.

These results suggest that the pairwise comparison condition was relatively ineffective not because comparison is an ineffective category learning strategy more generally, but instead because participants in the pairwise comparison condition focused on a prescribed set of comparisons involving two items at a time. When it comes to category learning, it may be important to consider the global structure of categories to effectively assess the cue and category validities of different features.

General Discussion

The present study investigated whether generating explanations and making comparisons would improve people's ability to discover rules that could be used to categorize a set of novel objects. All three experiments found that performing an explanation task enhanced discovery of categorization rules that could account for all cases; however, the effects of the comparison tasks were

more varied. Performing either within-category or between-category pairwise comparisons did not support rule discovery. However, comparing all the category exemplars in each group did increase 100% rule discovery.

Our results are consistent with previous work demonstrating that engaging in explanation supports learning. In particular, we replicate the results of studies that have used similar materials (Williams & Lombrozo, 2010, 2013). In the present study, the explanation task succeeded in helping participants discover abstract patterns that unified each of the categories. Furthermore, the explanation task stimulated spontaneous comparison, allowing participants to reap the benefits of comparison even if they were not explicitly asked to compare.

Surprisingly, we find that under some conditions engaging in a pairwise comparison task can impair learning. However, other types of comparison, such as comparing all the exemplars in each category, did promote learning, suggesting that comparison can be an effective strategy for learning novel categories. But importantly, some comparison prompts are more effective than others (see also Rittle-Johnson & Star, 2009), and comparison *prompts* may be most effective when they stimulate a broad range of comparison *processes*. One question for future research is whether the *combination* of within-category and between-category pairwise comparisons can in fact be beneficial, or whether "group" comparison provides unique advantages.

It is also worth pointing out some of the limitations of this study. Overall, the eight robots were highly similar and easily alignable. This might explain why spontaneous comparison was so common among participants who completed the explanation task. The high rates of spontaneous comparison make it difficult to differentiate effects of explanation from effects of comparison; the question of whether explanation and comparison exert unique constraints on learning may be easier to address with a task that more effectively isolates each process.

In future work, we hope to explore whether explanation and comparison have additive effects in more difficult learning tasks, where we also anticipate benefits to comparing (to align features) *before* explaining (to identify patterns). More research is needed, but the present studies provide important steps towards understanding the relationship between explanation and comparison and how these processes can most effectively support learning.

Acknowledgments

We thank Dedre Gentner and members of the Northwestern University Language and Cognition Laboratory for valuable feedback on this work and Lena Lam for assistance with coding. BJE was supported by an NSF Graduate Research Fellowship; TL was supported by an NSF grant (DRL-1056712) and the McDonnell Foundation.

References

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.

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

Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K. (2009) Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. *Cognitive Science*, *3,* 1343-1382.

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

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.

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

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review*, *6,* 586-597.

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.

Nokes-Malach, T. J., VanLehn, K., Belenky, D., Lichtenstein, M., & Cox, G. (2012). Coordinating principles and examples through analogy and self-explanation. *European Journal of Education of Psychology*.

Oppenheimer, D. M., Meyvisb, T., & Davidenkoc, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45(4)*, 867-872.

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.

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.

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.