Performance Pressure and Comparison in Relational Category Learning

John D. Patterson (jpatter4@binghamton.edu)
Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, 4400 Vestal Parkway East
Binghamton, NY 13905 USA

Abstract

An important objective in higher-order cognition research is to understand how relational categories are acquired and applied. Much of the research on relational category learning has investigated the role of within-category comparison opportunities in category acquisition and transfer – guided by predictions from structure mapping theory that alignment leads to highlighting and abstraction of shared relational structure (Gentner, 1983). Recent research has yielded a within-category comparison advantage under the supervised observational learning mode (relative to twice as many single-item trials), but not under the supervised classification mode (Patterson & Kurtz, 2015). In the present study we investigate the role that pressure to succeed at the training task – a critical difference between the two learning modes – plays in the apparent ineffectiveness of learning by comparison within the classification mode. In a 2x2 between-subjects design we crossed two levels of performance pressure (elevated and standard) with two presentation formats (single-item and within-category pairs). The main findings are: (1) a significant interaction showing a negative impact of increased performance pressure for single-item learners, but not for comparison learners; and (2) a theoretically predicted, but empirically elusive effect of comparison over single-item in the classification mode. We conclude that: (1) performance pressure exerts a deleterious effect on relational category learning (in accord with findings in the attribute category literature) that opportunities to compare may compensate for; and (2) pressure to perform does not appear to underlie lackluster comparison + classification performance (relative to observational learning). Further, we offer new evidence on the role that within-category comparison plays in relational category learning.

Keywords: relational categories; structural alignment; comparison; classification learning; transfer; performance pressure

Introduction

Categorization and comparison are two mechanisms that play a central role in human learning, comprehension, and knowledge use. The study of categorization has largely focused on attribute-based categories, or categories whose members belong based on a shared set of intrinsic features. Attribute categories have been studied under a diverse set of circumstances ranging from inference of missing features (Markman & Ross, 2003) to category construction (Ahn & Medin, 1992). By far however, the most prevalent paradigm of study has been the traditional artificial classification learning (TACL) paradigm. In its most common form, the paradigm operates as follows: a single stimulus is presented, the participant is asked to classify the item into one of two category options, a response is selected, and corrective feedback is given. The study of attribute categories has led to a considerable body of knowledge and has provided a viable testing ground to evaluate formal models of categorization (e.g., ALCOVE, Kruschke, 1992; DIVA, Kurtz, 2007; SUSTAIN, Love, Medin, & Gureckis, 2004).

However, attribute-based understandings alone cannot adequately characterize the richness of human category knowledge. Beyond our attribute-based understanding, we are sensitive to, and knowledgeable about, the ways in which objects and attributes in the world relate to one another. Accordingly, the categorization literature has placed increasing emphasis on the study of relational categories (Gentner & Kurtz, 2005; Markman & Stilwell, 2001). Members belong to relational categories, not based on their attributes, but instead based on shared relational structure. For example, consider the relational noun *barrier*. A thunderstorm, for instance, might be a *barrier* to a baseball game. Just the same, a sick child could be a *barrier* to you going to work. Considering the attributes of these two instances of *barrier* reveals that they are greatly disparate; they would be poor candidates for shared membership based on their attributes. However, their overlapping relational structure – that they both occupy an obstructing role between two other things – grants them membership in a shared relational category.

A key question in the study of relational categories is how we come to learn them. A strong candidate mechanism is comparison. A substantial body of evidence from the relational reasoning literature has found that comparison promotes learning and transfer of shared relational structure (Alfieri, Nokes-Malach, & Schunn, 2013, Gick & Holyoak, 1983; Kurtz, Miao, & Gentner, 2001; Loewenstein, Thompson, & Gentner, 1999; see Loewenstein, 2010 for a review). The benefits of comparison can be explained through the process of structural alignment (Markman & Gentner, 1993). Comparison encourages the alignment of instances' relational predicates, serving to preferentially highlight shared relational structure that is not readily salient when considering either instance alone. Importantly, comparison fosters abstraction of shared relational structure, thereby promoting later transfer. Thus, it is predicted that comparison of instances belonging to the same relational category should produce greater highlighting and learning of shared relations than sequential presentations. Indeed, the study of relational categories has shown same-category comparison to produce these effects in children (Gentner & Namy, 1999; Son, Smith, & Goldstone, 2011) and in adults (Kurtz, Boukrina, & Gentner, 2013; Patterson & Kurtz, 2015).

Though the support for comparison's central role in relational learning is extensive, research in relational category learning has exposed possible boundary conditions of its potency. Borrowing from the tradition of the attribute category literature, the study of relational categories has frequently used the TACL paradigm. However, when a straightforward pairing of pure within-category comparison with TACL has been used, it has failed to produce the predicted comparison advantage. An attempt by Kurtz and Gentner (1998) found that a straightforward merging of same-category comparison with classification required twice as many stimulus exposures to produce an effect of comparison over sequential presentations. Moreover, we conducted a series of follow-up attempts that failed to find an effect of comparison when exposure was equated – despite efforts to enhance the invitation to compare by eliciting similarity ratings or correspondences between elements.

Recent research, however, has shown that learning mode can influence the impact of comparison. Patterson and Kurtz (2015) found robust comparison advantages relative to equivalent single-item presentation on within- and across-domain tests of knowledge when learning under a supervised observational mode (passive study of labeled examples). However, no effect was found in the supervised classification mode – suggesting that some facet(s) of the classification mode attenuates the power of comparison. A key difference between the two modes is that task performance – tied to the guess-and-correct cycle of TACL – is emphasized in the classification mode. When combining comparison and classification, this may elicit competition for resources or a strategic element that favors one component or the other. By receiving accuracy feedback on each trial, it is possible that participants' theory of task shifts relative importance to the guess-and-correct cycle over comparison. We can refer to this as a *weighting effect.*

Alternatively, the apparent incompatibility between classification and comparison may be explained by a *distraction effect* (Beilock & Carr, 2005; Markman, Maddox, & Worthy, 2006). According to the *distraction hypothesis*, peoples' available cognitive resources decline as pressure to perform increases – a 'choking effect' that is accredited to the occupation of executive attention with task-irrelevant thoughts and performance-related worries. Given the greater performance demands of the classification mode, it is possible that the pressure to perform results in the occupation of cognitive resources needed for comparison. Indeed, working memory resources have been marked as important to the structural alignment process (Waltz, Lau, Grewal, & Holyoak, 2000). Waltz et al. (2000) found that when participants were under load that occupied either phonological or executive working memory, noticing of relational correspondences was significantly attenuated and noticing of simpler-to-detect attribute correspondences increased.

In the present study, we sought to determine whether the ineffectuality of comparison in the classification mode can be traced to *weighting* or *distraction* effects by varying the degree of performance pressure in the classification task. The global factor, termed 'performance pressure', was manipulated by testing an elevated pressure condition operationalized by: (1) an instruction asserting that achieving the highest possible performance was the goal, (2) trial feedback in red/green colors (incorrect/correct) to underscore evaluation, and (3) presentation of a running accuracy percentage at the end of each trial. The standard pressure group did not receive (1) or (3) and received feedback in black text. The manipulation was intended to further accentuate the guess-and-correct cycle and increase perceived pressure to perform. It would also be helpful to make the classification task *less* performance-centric, but reducing pressure would seem to require changing the essential nature of the task or employing a relatively weak manipulation. If comparison in the classification mode is ineffective due to either a *weighting* or *distraction* effect elevated performance pressure should lead to poorer learning outcomes than standard pressure. To be clear, both possible effects would be predicted to result in the same performance outcome – differentiating between the two is not of immediate concern.

In addition to exploring the comparison + classification incompatibility, the present study addresses the influence of performance pressure on relational category learning more generally by looking at its effect on single-item learning. Though addressed in the attribute category literature, this issue has not been studied in the relational category realm. Markman, Maddox, and Worthy (2006) showed that participants under very high pressure to perform (i.e., contingent monetary incentives) were hindered, by a posited *distraction effect*, in their ability to learn sequentially presented rule-based attribute categories, but were facilitated in learning information integration categories. Relational categories are generally considered to be verbalizable and rule-like in nature (Gentner & Kurtz, 2005). To the extent that the rule-like nature of attribute and relational categories is similar, a *distraction* effect marked by reduced learning outcomes should be expected for single-item relational category learning under elevated pressure.

In a 2x2 design, the effect of performance pressure (standard, elevated) on relational category learning was assessed across two presentation formats (single-item, within-category comparison). Additionally, we measured participants' regulatory focus – their degree of approach (promotion) and avoid (prevention) motivation using Higgins, Friedman, Harlow, Idson, Ayduk, and Taylor's (2001) regulatory focus questionnaire (RFQ). The match between a person's regulatory focus and their perceived nature of the reward structure has been shown to be important to cognitive tasks, including category learning (see Maddox, Baldwin, & Markman, 2006). To control for how individual differences in participants' regulatory focus might interact with the pressure manipulation, we incorporated their RFQ data into our statistical models.

Method

Participants

93 undergraduates from Binghamton University participated toward partial fulfillment of a course requirement.

Materials

The RFQ contained 11 questions, six measuring participants' level of promotion focus and five measuring their level of prevention focus. A five-point Likert scale was used to collect participants' RFQ responses. The training and within-domain test phase stimuli consisted of 36 unique, Stonehenge-like rock arrangements (see Figure 1). Rocks varied in their size, shape, color, and spatial location. The stimuli comprised three relational categories: *monotonicity* – defined by a monotonic decrease in height of the arrangement from left to right, *support* – characterized by the presence of one rock being supported by two other rocks, and *mirrored stack* – consisting of two same color rocks of similar size and shape arranged in a stack. Each arrangement belonged to only one of the three categories. A 24 item subset of 36 stimuli was used for training and 12 were reserved for use at test; subsets were balanced by category. The subsets matched those used in Kurtz, Boukrina, and Gentner (2013) and the subset was held constant across participants. For comparison conditions, training stimuli were presented in same-category pairs. Pairs were randomly generated for each participant on each pass through the training set.

To evaluate participants' ability to transfer category knowledge to a novel domain, 15 mobile-like stimuli were used (see Figure 1). Each mobile conformed to one of the three categories from training, five per category. The mobiles differed considerably from the training/test stimuli in their surface characteristics (color/shape of objects) as well as in the spatial orientation of the category-defining core which was reflected over the X-axis.

Design and Procedure

In a between-subjects design, participants were randomly assigned to one of four training conditions using the supervised classification task. Two conditions presented pairs of items for same-category comparison: elevated performance pressure comparison (n = 23) and standard performance pressure comparison (n = 23). The other two conditions trained with twice as many single-item trials: elevated performance pressure single-item (n = 24) and standard performance pressure single-item (n = 23). Prior to the training instructions, all participants completed the RFQ (Higgins et al., 2001). Following the RFQ, all participants received an archeological cover story and instructions indicating that their goal was to figure out what makes a rock arrangement belong to a particular type and that they would be tested on their knowledge later. Participants in comparison conditions received an additional instruction indicating that looking at the two arrangements together can be helpful for learning. A further instruction was given to the elevated performance pressure groups, indicating (1) that a performance tracker would show them the percentage of trials they had gotten correct after each trial, and (2) that their job was to achieve the highest percentage they could.

Training – Comparison Conditions Training consisted of two passes through 12 paired stimulus trials, totaling 48 stimulus exposures. Each trial showed two stimuli, side-by-side, that remained visible until the trial was complete. Participants were asked to make a joint category decision for the presented items and selected their response with the mouse. The joint decision represents a deviation from Patterson and Kurtz (2015), however preliminary work showed performance under joint and independent responses to be nearly identical. Time was unconstrained for the category decision. Following the response, participants were given feedback for two seconds indicating: (1) whether they were correct and (2) the correct category of the items. In the elevated performance pressure condition, participants received feedback in either green (correct) or red (incorrect) and at the end of the trial participants saw a screen for two seconds showing the percentage of trials correct to that point before advancing to the next trial.

Training – Single-item Conditions Training consisted of two passes through 24 randomized, single-item trials, totaling 48 stimulus exposures. Outside of the number of trials and language adjusted to accommodate single, instead of paired, items in the decision query, the single-item conditions were conducted identically to the comparison conditions.

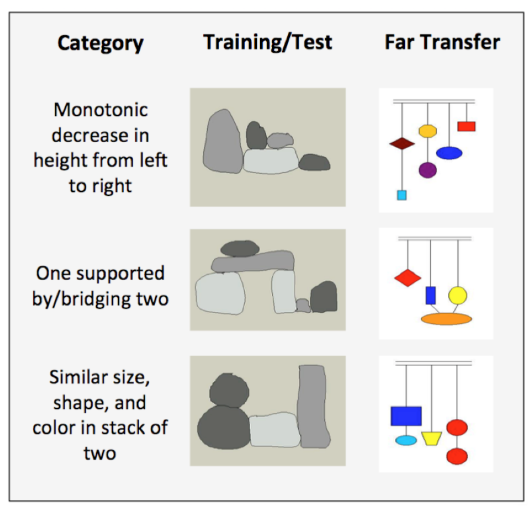

Figure 1: Sample stimuli for each category in each phase.

Assessment After the training, all conditions performed an identical assessment sequence. Participants first received a within-domain test consisting of the 24 'old' rock arrangements from training and 12 previously unseen arrangements. Old and new items were interspersed randomly. The transfer assessment, consisting of 15 randomly ordered mobile stimuli, was presented after the within-domain test. Both the within-domain and transfer assessments were conducted using a category endorsement task: on each trial an item was presented along with a query asking whether the item belonged to a given category and participants responded with either “yes” or “no.” Given our primary interest in how well acquired category knowledge could be extended to new cases, old test items were presented once each while the new within-domain and far transfer items were each presented twice – once with an accurate category label and once with an incorrect label.

Results

The trial-wise accuracy data were modeled using binomial generalized linear mixed effect regressions via the lme4 (Bates, Maechler, Bolker, & Walker, 2015) and lmerTest (Kuznetsova, Brockhoff, & Christensen, 2015) packages for the R environment (R Core Team, 2015). Models examining main effects included trial number and the effect of interest as fixed effects. Models examining interactions included trial number, presentation condition, performance pressure condition, and the presentation condition by performance pressure interaction as fixed effects. All models controlled for participants’ regulatory focus in the random effects structure by including random intercepts for the 16 levels of prevention focus and 15 levels of promotion focus that resulted from the sample’s RFQ data. In the interest of brevity, results of the RFQ will not be discussed. Adjusted means and standard errors for accuracy data can be seen in Table 1.

Training

Modeling of the performance data revealed three effects in the training phase. First, trial number was found to be significantly predictive of accuracy ($\beta = 0.04$, $SE = 0.003$, Wald $Z = 11.86$, $p < .0001$), indicating that participants’ accuracy increased as they progressed through training. Second, a main effect of presentation condition was observed ($\beta = 0.53$, $SE = 0.14$, Wald $Z = 3.84$, $p = .0001$), showing that comparison learners were more accurate than their single-item counterparts. Third, the analysis revealed a marginal presentation condition by performance pressure interaction ($\beta = 0.34$, $SE = 0.18$, Wald $Z = 1.90$, $p = .057$). The interaction was driven by an effect of performance pressure for the single-item condition (standard pressure > elevated pressure; $\beta = -0.22$, $SE = 0.11$, Wald $Z = -2.06$, $p = .04$), but not for the comparison condition ($\beta = -0.11$, $SE = 0.14$, Wald $Z = -0.77$, $p = .44$). In a follow-up test to the interaction, both comparison conditions outperformed their single-item counterparts (elevated pressure: $\beta = 0.87$, $SE = 0.14$, Wald $Z = 6.20$, $p < .0001$; standard pressure: $\beta = 0.53$, $SE = 0.14$, Wald $Z = 3.84$, $p < .001$).

Old Within-domain Items

Modeling of the old-item accuracy data revealed only a significant effect of trial number ($\beta = -0.01$, SE = 0.01, Wald Z = -2.76, p = .006), indicating that participants’ accuracy was slightly lower for items encountered later in the queue compared to those encountered earlier. This effect may be associated with old items being interspersed with new ones. There was also a marginal main effect of performance pressure ($\beta = -0.22$, SE = 0.13, Wald Z = -1.80, p = .07), marking higher accuracy for the standard pressure group compared to the elevated pressure group. In addition, there was a marginal interaction mirroring that in the training phase ($\beta = 0.41$, SE = 0.24, Wald Z = 1.69, p = .09) – showing a significant difference between levels of performance pressure for the single-item condition (standard pressure > elevated pressure; $\beta = -0.43$, SE = 0.17, Wald Z = -2.44, p = .01), but not for the comparison condition ($\beta = 0.02$, SE = 0.17, Wald Z = 0.10, p = .92).

Novel Within-domain Items

Analyses of the new, within-domain items yielded one reliable effect: a presentation condition by performance focus condition interaction ($\beta = 0.73$, SE = 0.22, Wald Z = 3.40, p = .001). Again, the interaction was characterized by a reliable difference between levels of performance pressure for the single-item presentation group (standard pressure > elevated pressure; $\beta = -0.52$, SE = 0.15, Wald Z = -3.41, p = .001), but not for the comparison group ($\beta = 0.21$, SE = 0.15, Wald Z = 1.34, p = .18). Examining the interaction further, a comparison advantage was found for the elevated pressure group ($\beta = 0.56$, SE = 0.18, Wald Z = 3.11, p = .002), but not the standard pressure group ($\beta = -0.17$, SE = 0.18, Wald Z = -0.98, p = .33). Since the comparison conditions did not differ, it suggests that the interaction and comparison advantage were driven by poor performance among elevated pressure single-item learners, rather than exceptional performance among elevated pressure comparison learners.

Table 1: Adjusted condition means and standard errors across all performance phases.

	Single-item		Comparison	
	Standard Pressure	Elevated Pressure	Standard Pressure	Elevated Pressure
Training	.59(.04)	.54(.04)	.71(.04)	.73(.04)
Old test	.86(.02)	.80(.03)	.81(.03)	.81(.03)
New test	.78(.03)	.68(.03)	.75(.03)	.79(.03)
Transfer	.64(.04)	.57(.05)	.79(.03)	.82(.03)

Transfer

The transfer results are visualized in Figure 2. Three reliable effects emerged in the analysis of the transfer data. First, an effect of trial number ($\beta = 0.02$, $SE = 0.005$, Wald $Z = 4.73$, $p < .0001$) demonstrated that participants’ accuracy increased slightly over the course of the transfer assessment. The effect suggests that domain experience helped participants to recognize the deep similarity between the two domains and apply their category knowledge.

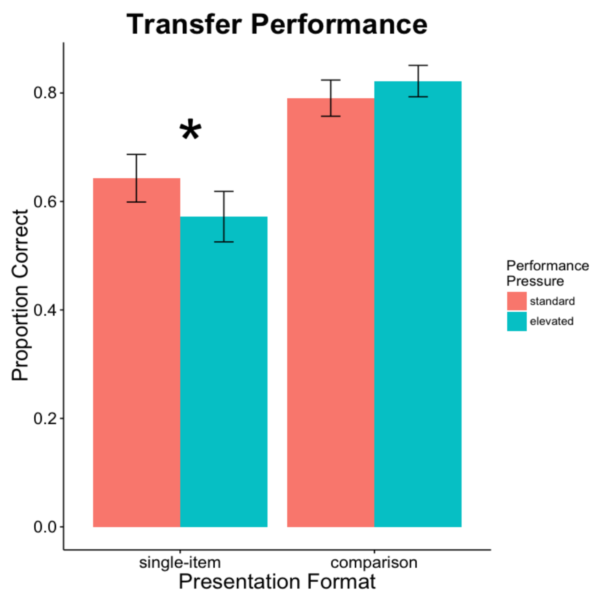

Figure 2: Transfer accuracy by condition. Error bars show +/- 1 SE.

Second, a highly reliable main effect of presentation condition was observed ($\beta = 0.99$, $SE = 0.16$, Wald $Z = 6.08$, $p < .00001$), indicating that participants in the comparison group were better at extending their knowledge to a novel domain than those in the single-item group. Third, a significant presentation condition by performance pressure interaction was observed ($\beta = 0.50$, $SE = 0.19$, Wald $Z = 2.63$, $p = .01$). The interaction was characterized by pressure-related decrements for the single-item group (standard pressure > elevated pressure; $\beta = -0.29$, $SE = 0.14$, Wald $Z = -2.19$, $p = .03$), but not the comparison group ($\beta = 0.20$, $SE = 0.14$, Wald $Z = 1.46$, $p = .14$). Critically, follow-up tests examining the effect of comparison revealed that both standard ($\beta = 0.74$, $SE = 0.19$, Wald $Z = 3.94$, $p = .0001$) and elevated pressure ($\beta = 1.24$, $SE = 0.19$, Wald $Z = 6.56$, $p < .0001$) comparison led to highly reliable advantages over their single-item counterparts.

Discussion

The goal of this study was to evaluate the effect that performance pressure exerts on relational category learning. The results convincingly show that performance pressure has a deleterious effect on relational category learning – leading to learning decrements that were realized across training, within-domain test (both old and new items), and at transfer. However, the pressure-related deficits were restricted to learners who trained via sequential presentations. The effect of performance pressure on single-item learning shows a consistency between the domains of attribute and relational category learning. The similarity between pressure effects for relational and attribute category learning (Markman, Maddox, & Worthy, 2006) suggests that similar systems and/or resources are employed for learning the both category types – perhaps mutually depending on an explicit, verbal system like that posited by the COVIS model (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). Future research should further explore the conditions under which the behavior of attribute and relational categories converges and diverges.

A specific interest embedded in our broader goal was to investigate if *weighting* or *distraction* effects might underlie the ineffectuality of comparison in the classification mode as seen by Patterson and Kurtz (2015). If cognitive resources that are necessary for comparison are retrained on the guess-and-correct cycle of classification *or* if they become occupied with worry under pressure to perform, then our performance pressure manipulation ought to have led to poorer comparisons and poorer learning. However, accentuating performance and increasing pressure was found to have no effect on the learning outcomes of comparison learners; this finding casts tentative doubt on the *weighting* and *distraction* accounts of lackluster comparison performance in the classification mode (relative to observational).

One possible rationale for why pressure did not affect comparison learning is that the manipulation was not sufficiently strong to elicit an effect. Other studies (e.g., Markman, Maddox, & Worthy, 2006) have employed stronger manipulations, such as contingent monetary incentives, to achieve high levels of pressure. It seems likely that, at greater levels of pressure, a *distraction* effect would occur, and learning outcomes under comparison would suffer. However, it is clear that our manipulation was sufficiently strong to elicit consistent performance pressure deficits for the single-item group. This suggests that comparison served a compensatory role to the otherwise negative effects of pressure. By this account, elevated pressure learners may have faced a *weighting* or *distraction* effect, but the resources allocated to the alignment process were sufficient to avoid performance decrements. To further elaborate on this effect, future studies should evaluate how differing levels of performance pressure affect alignment-based relational category learning.

Of considerable note, this study also showed a reliable effect of same-category comparison in the supervised classification mode (under standard pressure) at training and transfer over its single-item counterpart. To our knowledge, this is unprecedented. Finding a comparison effect at training and transfer was puzzling, since comparison advantages were not found in Patterson and Kurtz (2015). One notable difference is that the two studies utilized different analysis techniques. Follow-up analyses revealed that the effects from the present study remained after dropping the random effects structure; this prompted us to reanalyze Patterson and Kurtz (2015) using a logistic regression framework as in the present study. Though comparison in the observational mode was the clear winner, showing advantages over single-item at every phase, the results showed a reliable comparison effect in the classification mode over single-item at training and a marginal effect (p = .09) at transfer. This finding provides further evidence that comparison is an effective means to promote relational category learning. In addition, it echoes the need for researchers to utilize analysis techniques that adequately reflect the richness of the data. For category learning research, analyzing the data trial-wise using logistic regression increases the number of observations available for model fitting, thereby increasing sensitivity over analysis methods that aggregate across trials.

References

Ahn, W., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, *16*(1), 81-121.

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: a meta-analytic review. *Educational Psychologist*, *48*(2), 87-113.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A., & Waldron, E. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442.

Bates, D., Maechler, M., Bolker, B., Walker, S., (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail working memory and "choking under pressure" in math. *Psychological Science*, *16*(2), 101-105.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin*. Washington, DC: American Psychological Association.

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development*, *14*(4), 487-513.

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

Higgins, E. T., Friedman, R. S., Harlow, R. E., Idson, L. C., Ayduk, O. N., Taylor, A. (2001). Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. *European Journal of Social Psychology*, *31*, 3-23.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*, 560–576.

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*(4), 1303–1310.

Kurtz, K. J., & Gentner, D. (1998). Category learning and comparison in the evolution of similarity structure. *In Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kurtz, K. J., Miao, C. H., & Gentner, D. (2001). Learning by analogical bootstrapping. *The Journal of the Learning Sciences*, *10*(4), 417-446.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests for random and fixed effects for linear mixed effect models.

Loewenstein, J. (2010). How one's hook is baited matters for catching an analogy. *Psychology of Learning and Motivation*, *53*, 149-182.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review*, *6*(4), 586-597.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2), 309-332.

Maddox, W. T., Baldwin, G. C., & Markman, A. B. (2006). A test of the regulatory fit hypothesis in perceptual classification learning. *Memory & Cognition*, *34*(7), 1377-1397.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive psychology*, *25*(4), 431-467.

Markman, A. B., Maddox, W. T., & Worthy, D. A. (2006). Choking and excelling under pressure. *Psychological Science, 17*(11), 944-948.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592-613.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*, 329– 358.

Patterson, J.D., & Kurtz, K.J. (2015). Learning mode and comparison in relational category learning. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Son, J. Y., Smith, L. B., & Goldstone, R. L. (2011). Connecting instances to promote children's relational reasoning. *Journal of Experimental Child Psychology*, *108*(2), 260-277.

Waltz, J. A., Lau, A., Grewal, S. K., & Holyoak, K. J. (2000). The role of working memory in analogical mapping. *Memory & Cognition*, *28*(7), 1205-1212.