

Comparison-based Learning of Relational Categories (You'll Never Guess)

John D. Patterson & Kenneth J. Kurtz

Binghamton University

The authors would like to extend a special thank you to Garrett Honke and Sean Snoddy, as well as to current and former members of the Learning and Representation in Cognition (LaRC) Lab at Binghamton University (SUNY), for their valuable consult. Data uploaded on OSF and can be accessed here:

https://osf.io/w8gue/?view_only=7c1c0b80a9604bed9934b24c5b699b49

Correspondence concerning this article should be addressed to John D. (JD) Patterson, Department of Psychology, Binghamton University, Binghamton, NY 13902. Contact: jpatter4@binghamton.edu

Abstract

In accord with structural alignment theory, same-category comparison opportunities within a classification learning task should promote relational category acquisition. However, a straightforward merging of the classification paradigm with co-presentation of same-category item pairs does not yield an advantage relative to an equal number of single-item exposures. In three experiments, we explore the hypothesis that the traditional classification learning mode (guess-and-correct) and comparison have a previously unforeseen incompatibility. In Experiment 1, we test this hypothesis by contrasting classification with supervised observational learning (passive study of labeled examples) under three presentation formats: same-category pairs, mixed pairs, and single-item. We find an observational advantage with same-category pairs and produce the elusive advantage over single-item exposures. In Experiment 2, we assess the generality of the learning mode effect by testing both same- and different-category comparison. The observational advantage replicates and extends to different-category comparison – although, we do not find a significant difference between the two types of comparison. In Experiment 3, relative to the classification mode, we find enhanced performance in an intermediate learning mode between classification and observation in which participants are instructed to make a covert category guess (without making an actual response) before seeing the correct category label. Implications and interpretations – including our interpretation that the performance emphasis inherent in classification learning undermines the benefits that arise from comparison opportunities – are discussed.

Keywords: relational categories; comparison; classification learning; observational learning; transfer

Comparison-based Learning of Relational Categories (You'll Never Guess)

Categorization and comparison are mechanisms that play a central role in human learning, comprehension, and knowledge use. The processes of comparison and categorization interface in a number of critical ways in current psychological understanding: from similarity as the basis for determining category membership to categories based on analogies among their members. From a theoretical perspective, we address the broad question of how the joint processing of multiple examples (noting that such juxtapositions are myriad and arise from spatial, temporal, or symbolic origins; Gentner, 1989) impacts concept formation. From an applied perspective, educators need every advantage in promoting the speed and quality of concept acquisition – we need to know how techniques grounded in core cognitive processes can be formulated to achieve better classroom impacts and outcomes. The question of how to productively integrate two such broad and powerful mechanisms (comparison and supervised inductive categorization) is a direct and compelling concern. As Goldwater and Schalk (2016) lay out, relational categories are heavily embedded in formal education, so investigating relational category acquisition helps in the important work of advancing toward psychological explanation that treats natural categories acquired in real learning situations.

Attribute-based vs Relational Categories

In the study of human category learning, researchers have generally turned to well-controlled, artificial stimuli that belong to categories based on their independent attributes. These attribute categories have been studied under a diverse set of task circumstances (see Kurtz, 2015) including learning by inference of missing features (see Markman & Ross, 2003), unsupervised category construction (e.g., Ahn & Medin, 1992; Pothos et al., 2011) and passive, observational study of labeled examples (e.g., Estes, 1994; Levering & Kurtz, 2015). However, the dominant

approach has been the classification learning paradigm (Markman & Ross, 2003). In its canonical form, classification consists of a guess-and-correct cycle in which a single stimulus is presented, the learner makes a guess from provided category label options, and corrective feedback is provided. The study of attribute/feature-based categories via classification has unequivocally advanced our understanding of human category acquisition and provided a useful basis for comparing different formal models of categorization (e.g., ALCOVE, Kruschke, 1992; DIVA, Kurtz, 2007, 2015; SUSTAIN, Love, Medin, & Gureckis, 2004).

Much of the category knowledge we possess, however, is not reducible to lists of independent attributes (Murphy & Medin, 1985). Rather, relational categories like *revenge* are more abstract and depend on core relationships rather than on the presence of particular attributes or objects (Gentner & Kurtz, 2005). We possess a wealth of knowledge about the ways objects and attributes in the world relate to one another (see Gentner, 1981, 1982) – and this structural knowledge meaningfully picks out kinds in the world. Accordingly, an increased emphasis has emerged addressing the nature and acquisition of relational categories (e.g., Asmuth & Gentner, 2017; Corral & Jones, 2014; Gentner & Kurtz, 2005; Goldwater & Markman, 2011; Goldwater, Markman, & Stilwell, 2011; Goldwater & Schalk, 2016; Higgins, 2017; Markman & Stilwell, 2001; Rehder & Ross, 2001; Rottman, Gentner, & Goldwater, 2012). Relational categories (e.g., gift, robbery, sibling, barrier, reciprocity) are categories whose members share a common set of relations between objects and/or attributes (i.e., a shared relational structure). Since relational categories need only share a common relational structure, the members tend to be quite disparate in their surface attributes. In fact, what holds members together is more analogical similarity than literal similarity. Given the all-or-none constraint placed on membership, relational concepts can be interpreted as rule-like knowledge structures (see Gentner & Medina, 1998). That is,

relational criteria may be used to categorize members relatively unambiguously – in a way akin to the classical, definitional view of categorization (Gentner & Kurtz, 2005).

Comparison: A Possible Mechanism for Relational Category Learning

A key question that bears both theoretical and applied import is: how do we acquire relational category knowledge? Drawing on the analogical transfer literature, a promising candidate mechanism is learning via comparison. Considerable research has shown that simultaneous comparison of cases facilitates the acquisition and transfer of relational concepts (see Alfieri, Nokes-Malach, & Schunn, 2013 for a review and meta-analysis; see also Loewenstein, 2010). The benefits of comparison to relational learning have been observed across a wide array of learning domains – mathematics (Ming, 2009; Rittle-Johnson & Star, 2007, 2009), science (Kurtz & Gentner, 2013), negotiation (Loewenstein, Thompson, & Gentner, 1999, 2003), and engineering (Gentner et al., 2016) – as well as with stimuli both verbal (e.g., Catrambone & Holyoak, 1989; Gick & Holyoak, 1983; Gick & Paterson, 1992) and visual-perceptual (e.g., Gentner & Namy, 1999; Kotovsky & Gentner, 1996; Kurtz, Boukrina, & Gentner, 2013; Namy & Clepper, 2010; Namy & Gentner, 2002). Prevailing theory holds that comparison benefits arise through structural alignment (Gentner & Markman, 1997; Markman & Gentner, 1993a) in which the shared relational predicates of compared instances are brought into alignment – serving to highlight shared relational structure, as well as alignable differences (Markman & Gentner, 1993b). Importantly, this alignment process paves the way for abstraction and renders the common relational structure more portable for transfer to surface-dissimilar domains.

Despite the wealth of evidence supporting the benefits of comparison to learning and transfer, the evidence in the relational categorization literature has been somewhat mixed.

Comparison has been found to promote relational categorizations in the relational match-to-sample task – a task where an exemplar is given and participants must match it to one of two samples: one that matches relationally and another that matches based on common theme or perceptual characteristics. Relational match-to-sample task studies using both children and adults have shown a boost in relational categorization when comparison opportunities are provided (Christie & Gentner, 2010; Gentner, Anggoro, & Klibanoff, 2011; Gentner & Namy, 1999; Goldwater & Markman, 2011; Son, Smith, & Goldstone, 2011). The increased rate of relational categorization seen across these studies strongly suggests that participants were engaging in structural alignment – allowing them to discover and use common relational structure in their categorizations. The evidence from these studies, as well as a host of others in the analogical transfer literature, indicate that engaging in same-category comparison should yield the familiar benefits of structural alignment and promote not just increased relational responding, but also enhanced learning and transfer.

Failures of Same-Category Comparison: Previous and Preliminary work

When the acquisition and transfer of relational knowledge within an inductive category learning task has been the target of study, the picture has been markedly different. The natural approach is a paired-exemplar variant of the traditional classification paradigm such that on each trial the learner is presented with two exemplars, queried for their category membership, and provided with corrective feedback for both items. Guided by predictions that follow from the structural alignment view, providing learners with same-category comparison opportunities during training should produce enhanced category acquisition (via training and transfer measures) relative to providing learners with an equal number of one-at-a-time stimulus exposures (single-item learning). The logic behind this prediction is straightforward: (1)

comparison has been widely shown to benefit the acquisition of core relations through induction; (2) the classification learning mode is a reliable platform for learning novel categories through induction; therefore (3) combining comparison with the classification learning mode should provide a potent platform for learning novel relational categories through induction. Curiously though, we know of no successful realization of this formula. One attempt by Kurtz and Gentner (1998) revealed that twice as many stimulus exposures could produce a reliable advantage over single-item learning, but when controlling for the number of stimulus exposures between learning conditions, no such comparison advantage was found (Kurtz & Boukrina, 2004).

The comparison manipulations utilized in these studies were relatively non-intensive – therefore, a plausible account of these failures is that the comparison engine was simply not engaged to a sufficient degree (see Kurtz, Miao, & Gentner, 2001). In preliminary work for the present investigation, we explored this possibility in a series of attempts to boost joint consideration and alignment. In one case we simply elicited similarity ratings from learners prior to making classification decisions on each learning trial – a technique known to encourage alignment (e.g., Goldwater & Markman, 2011; Markman & Gentner, 1993a). When this did not work, we attempted a more heavy-handed approach: learners were instructed to produce explicit element-to-element alignments (e.g., Dumas & Hummel, 2013; Kurtz et al., 2001) by drawing correspondence lines between elements that played similar roles in their exemplars before making classifications on each trial. Once again, an established method for encouraging alignment failed to elicit a comparison advantage.

The Guess-and-Correct Cycle as an Obstructing Factor

One account of this string of failures is that same-category comparison simply does not benefit the acquisition of relational kinds, but this is difficult to reconcile with the surfeit of

evidence for comparison-based learning (see Corral, Kurtz, & Jones, 2018). The impetus for the present investigation is the possibility that some aspect of the task environment is acting in opposition to the expected benefits of comparison. There is a key difference in task environment between the failures and the plethora of studies showing comparison benefits: the integration of comparison with the classification learning mode. Thus, a plausible account of these failures is that the classification mode in some way obstructs the benefits of same-category comparison.

We consider several factors that come into play. First, classification is based on an extended series of trials (i.e. training blocks) providing a large number of varied comparison opportunities. Presumably this would boost the potential impact of comparison (though one could also imagine a mitigating fatigue factor). The lightweight nature of these comparison opportunities is also a notable difference, but our preliminary studies showed no impact of employing a more robust comparison task. Secondly, classification brings with it a performance emphasis (i.e., learners strive to produce a high rate of correct answers and reach a stopping criterion). Given it is performance on the classification task, rather than concept learning per se, that is emphasized, strategies that allow participants to perform well with relatively low effort during training tend to be favored (see Levering & Kurtz, 2015). One strategy that is readily available, due to the trial structure in a uniform same-category comparison learning format, is a ‘pick-the-best’ approach. Under this strategy, participants make classification decisions for both stimuli based on whichever stimulus they are most confident about. As such, joint consideration of the co-presented items may be engaged merely as a tool to ascertain which item is more clearly associated with a particular category, rather than as a vehicle for alignment, highlighting, and abstraction (see Kurtz et al., 2013). It follows clearly how this would undermine a predicted comparison-based learning effect.

A ‘pick-the-best’ account alone, however, is seemingly insufficient to explain the apparent incompatibility between comparison and classification learning. If this strategy were the only barrier, our attempts at enhancing comparison engagement ought to have counteracted the negative effects of it. Why did these engagement attempts not improve the effectiveness of comparison? An additional dynamic is that there is quite a bit going on in a task with repeated trials of two items being presented, classified, and evaluated for correctness. We consider the learner’s *theory of task* to be their derived interpretation of what they need to accomplish in the learning setting and how best to achieve it. Learners are afforded with degrees of freedom in deciding if, and to what extent, they engage in or rely upon each element of the learning experience. We posit that the supervisory feedback of classification may bias learners to perceive the guess-and-correct cycle to be the most central component. Without any direction or supervision of the comparison component, and without a definitive indication as to whether or how comparison is linked to their performance in the classification task, the comparison component may be seen as an inconsequential facet of the task. Thus, participants may prioritize the guess-and-correct cycle and allocate insufficient resources to comparison – even with additional task components meant to encourage the comparison process.

The Observational Mode: Removing the Guess-and-Correct Cycle

These proposed explanations are presumably not exhaustive – other accounts related to available resources or the integration of the two task components may be plausible. What is clear is that the stark contrast in the effect of comparison between studies where comparison is embedded in the classification learning mode versus the rest of the literature strongly implicates the classification learning mode as an obstruction to making fruitful comparisons. Accordingly, we predicted that removing the guess-and-correct cycle from the task environment would lead to

the expected benefits of same-category comparison over single-item learning. To this end, in Experiment 1, we use the supervised observational mode – a task that is informationally-equivalent to the classification mode (i.e., both tasks provide a complete example and a class label), but not procedurally-equivalent, in that it requires only the passive study of items with their category labels instead of active prediction. Despite matching on information provided, the key procedural difference (i.e., the guess-and-correct cycle) was expected to lead to differences in what learners would be able to extract from comparison opportunities.

We compare the observational mode to the classification learning mode under three presentation formats (same-category pairs, mixed pairs, and single-item). This design allows a number of potentially fruitful comparisons highlighted by allowing us to test the prediction of more successful category acquisition through same-category comparison opportunities in the observational learning mode relative to: (1) the same comparison opportunities in classification mode; (2) a baseline of single-item learning; and (3) the current best practice for promoting relational category learning – an even mixture of same- and different-category comparison opportunities (Kurtz et al., 2013).

Overview of Experiments

In these experiments, we investigate how learning mode and comparison influence the acquisition and transfer of relational category knowledge. In Experiment 1, we evaluate how these two learning modes affect learning from same-category pairs, mixed pairs, and single-item formats – showing that learning mode plays a vital role in the success of same-category comparison. In Experiment 2, we replicate the effect of learning mode on same-category comparison and extend it to contrastive comparisons. We find that both types of comparison are aided by the removal of the guess-and-correct cycle, but that comparison type does not exert an

effect on overall learning outcomes. In Experiment 3, we conduct a replication with extension and, further, attempt to explicate the reason(s) for the incompatibility between classification and comparison. To this end, we construct two novel learning conditions, intermediate on a classification-observational continuum, and find evidence that deficits associated with the classification mode are linked to the task's emphasis on performance.

Experiment 1

The purpose of this experiment was to assess the impact of learning mode on relational category learning. We employed a 3 (presentation type: same-category pairs, mixed pairs, single-item) x 2 (learning mode: classification, observational) design. Our core prediction was that same-category comparison would result in better learning under the observational mode than in the classification mode and would produce an advantage over single-item learning with equivalent item exposures. We also sought a conceptual replication of the advantage of mixed pairs comparison over single-item in the classification learning mode (seen in Kurtz et al., 2013) – including substantial alterations to the procedure (described below) and an extension to evaluate the impact of the observational learning mode. Finally, we included the single-item learning control in the observational mode as an exploratory component since there has been no previous test (to our knowledge) of classification versus observational single-item learning of relational categories (see Higgins, 2017 for related work).

Method

Participants. This study and all subsequent studies were cleared by the Institutional Review Board at Binghamton University. 184 Binghamton University undergraduates participated for partial course credit. It was unclear ahead of time what the mode effect size for same-category comparison would be, and consequently how we would calculate an appropriate

sample size, for two reasons. First, no studies prior to the current study have examined the effect of mode on comparison-based relational category learning; while a mode effect size from the feature-based category learning literature could be used, this would be arbitrary and likely inaccurate, given the emphasis on comparison and relational categories in the present work. Second, our intention was to analyze the data with a statistically rigorous method (generalized mixed effect regression), which accounts for random effects of participant and item. Estimating power and sample size requirements in this case requires an estimate of both participant and item variability; thus, a previous effect size alone would be insufficient to estimate an appropriate sample size.

In these cases, an accepted approach is to generate/simulate data from a model that is fitted to data that is representative of the target effect and random effects, fit the model again to the newly simulated data, and then run the intended statistical test across many different runs (Arnold, Hogan, Colford, & Hubbard, 2011; Green & MacLeod, 2016). The degree to which the test finds an effect where one exists, across different numbers of simulated subjects, is then used to calculate power and estimate an appropriate sample size. We report these simulations below. The sample size we used ($n_{\text{subjects}} \approx 31$, $n_{\text{observations}} \approx 2418$) was hoped to provide sufficient power. However, either way, the data in this experiment is essential for constructing an appropriately powered experiment.

Materials. The training and testing phase stimuli consisted of 36 unique, Stonehenge-like arrangements of rocks – examples can be seen in Figure 1. Rocks varied in their size, shape, and color. As in our previous studies, the stimuli comprised three relational categories (category labels in brackets): monotonicity [Besod] – defined by a monotonic decrease in height of the arrangement from left to right, support [Makif] – characterized by the presence of a rock being

supported by two other rocks, forming a sort of bridge, and symmetry [Tolar] – captured by the presence of two same color rocks of similar size and shape, one stacked atop the other. Each arrangement belonged to only one of the three categories. Of the 36 stimuli, a subset of 24 was utilized as the training set (eight per category) and 12 were reserved for use at test (four per category). The subsets matched those used in Kurtz et al. (2013), and subsets were held constant across participants (though the order was randomized for each participant). For comparison conditions, training stimuli were presented in pairs. The pairings were randomly generated for each participant according to their condition – all same-category pairs or a fifty-fifty blend of same- and different-category pairs.

In order to assess far transfer of category knowledge, a set of 15 mobile-like stimuli (colorful, geometric objects connected with vertical lines, as if hanging down from a platform; see Figure 1) was used. Each mobile conformed to one of the three relational categories from training, five mobiles per category. Compared to the training and testing stimuli, the mobiles were dissimilar in their surface characteristics (in color and shape of objects) and the orientation of the category-defining core in each item was reflected over the X-axis.

Procedure. In a between-subjects design, participants were randomly assigned to one of six conditions. Four conditions employed comparison learning: same-category classification ($n = 30$), mixed-pairs classification ($n = 31$), same-category observational ($n = 31$), and mixed-pairs observational ($n = 32$); and two conditions used the traditional one item per trial: single-item classification ($n = 31$) and single-item observational ($n = 29$). All participants received an archeological cover story and the following instructions: “Your overall goal is to figure out what makes a given rock arrangement belong to one of the three types: Besods, Makifs, or Tolars. You will be tested on your knowledge of each type later.” The following instructions were given to

comparison conditions (the same instructions were stripped of two-item and comparison language in the single-item version): “On each learning trial, you will see two rock arrangements. [Observational: You will be shown the correct type for each arrangement to help you learn; Classification: Try to figure out the correct type for each arrangement. Use the mouse to select your response. A box will appear around the arrangement that you should respond to. You will be given feedback at the end of each trial to help you learn]. At first you will not understand what makes them belong to a type, but before long you should become quite good at recognizing the different types. Remember that there are three different styles for arranging the rocks into configurations. Looking at the two arrangements together can help you learn these types. Try your best to gain mastery of the names of each type and what makes an arrangement belong to those types. Learn as much as you can before the test!”

Comparison conditions – training. Training consisted of two cycles of 12 paired stimulus trials, totaling 48 stimulus exposures. At the beginning of each trial, two side-by-side stimuli were presented and remained visible until the trial was complete. In the classification conditions, the stimuli were presented on the screen for 500ms before a box appeared that randomly queried one of them. At the same time, response buttons appeared below the stimuli and participants were asked for the category of the queried item. They selected their response using the mouse, and were shown visual confirmation of their selection. Participants were then queried about the other item, made a response, and were shown visual confirmation of their selection. Following both responses, participants were shown simultaneous feedback for each item indicating: (1) whether or not their response was correct, (2) the correct category of the item (in green), and (3) if incorrect, the category they responded with (in red). In the observational conditions, the correct category labels appeared below the stimuli 500ms after stimulus onset and

remained on screen for the duration of the trial. When the participant finished studying an item pair they continued to the next trial with a mouse click. Participants in both classification and observational conditions had as much time to engage with each trial as they wished. To be clear, the same-category and mixed-category conditions followed the same procedure with the only difference being whether the two items were from the same category for every pair or for half of the pairs. Note that we did not employ an orienting task to encourage comparison before each trial (as in Kurtz et al., 2013) in part because it is a somewhat awkward component of the task and in part due to evidence that enhancing the invitation to compare had little impact.

Single-item conditions – training. Training consisted of two cycles of 24 randomized, single-item trials, totaling 48 stimulus exposures (equal to the comparison conditions). A single stimulus was presented at the start of each trial and remained visible until the trial was complete. The classification and observational conditions paralleled their comparison counterparts. In the classification condition, participants were asked for the category of the item, made a response, and were given visual confirmation of their selection. Following their response, they were presented feedback identical in nature to the comparison classification conditions. In the observational condition, participants were presented with a single labeled item. Onset times for the membership query and response buttons (classification) or label presentation (observational) were the same as the comparison group. As in the comparison conditions, single-item conditions were permitted as much time as desired for each trial.

Assessment. Following training, all conditions performed an identical assessment sequence. The sequence consisted of a within-domain test followed by a far transfer assessment. The within-domain test presented the 24 “old” rock arrangements from the training set and 12 new arrangements in a randomly intermixed order for each participant. After the within-domain

test, the 15 mobile stimuli were presented in random order for each participant in the far transfer phase. Both the test and transfer trials employed an endorsement format: on each trial a single item was presented, the participant was asked if the item belonged to a given category, and participants gave a yes/no response. This measure of categorization performance is similar to, but distinct from, both the classification and observational learning tasks. The endorsement task reduces any transfer appropriate processing advantages (Morris, Bransford, & Franks, 1977) that might result from a perfect task match between training and testing phases. As our primary interest was in how well knowledge could be extended beyond the exemplars encountered during the learning phase, old test items were each presented once while the new test and transfer items were each presented twice (once with the accurate and once with an inaccurate category label).

Results and Discussion

Learning phase. Although our primary interest concerns performance at test (where we have common dependent measures across conditions), we first report on accuracy and time on task data from the learning phase. Accuracy data from the classification groups were modeled trial-wise using generalized linear mixed effect regression with the lme4 (Bates, Maechler, Bolker, & Walker, 2015) and lmerTest (Kuznetsova, Brockhoff, & Christensen, 2015) packages in the R environment (R Core Team, 2015). Trial number and presentation format were used as fixed effects. Subject was included as a random effect, however the random effect for item was omitted, as it was unclear how to adequately represent both one and two concurrent items (depending on condition) within the model. All classification conditions demonstrated evidence of learning with all significantly performing above the chance level of .33 ($p_s < .05$). A reliable advantage in training accuracy was found for mixed pairs comparison ($M = 0.65$, $SE = 0.03$) over single-item ($M = 0.54$, $SE = 0.04$; $\beta = 0.50$, $SE = 0.20$, Wald $Z = 2.46$, $p = .014$), which is

consistent with the findings of Kurtz et al. (2013). Also in keeping with previous and preliminary work, same-category comparison ($M = 0.59$, $SE = 0.03$) once again failed to show an advantage over the single-item control, $p = .28$. The observational groups have no accuracy data during the training phase.

It is also useful to evaluate the learning modes for differences in time on task since it is conceivable that an advantage could be attributable merely to increased study time. Time on task included the entire amount of time from stimulus onset until the participant advanced to the next trial. Median time on task values were extracted for each participant and these data were subjected to a linear regression. Descriptive statistics reflect adjusted means and standard errors. The analysis showed that, across presentation formats, observational learners ($M = 3.80$, $SE = 0.31$) spent significantly less time on each trial relative to classification learners ($M = 4.74$, $SE = 0.31$); $p < .05$. Probing this difference further, we found it was underpinned by a reliable effect of mode for the comparison groups ($ps < .05$), but not the single-item group. As such, we can conclude that an advantage for the observational mode (as we predicted) would not be attributable to more study time since there was in fact less.

We also looked at the effect of presentation format on time on task. Mixed pairs learners (Observational: $M = 5.08$, $SE = 0.46$; Classification: $M = 6.59$, $SE = 0.47$) spent marginally more time ($ps = .07$) than same pairs learners (Observational: $M = 3.89$, $SE = 0.47$; Classification: $M = 5.37$, $SE = 0.47$), and both types of comparison spent more time than single-item learners (Observational: $M = 2.30$, $SE = 0.48$; Classification: $M = 2.27$, $SE = 0.47$) which is unsurprising given there are two stimuli and classification decisions per trial; $ps < .05$.

Test phase. Adjusted means and standard errors for all test item types can be seen in Table 1. See Figures 2-4 for old item, new item, and transfer performance, respectively. Like the

training data, the accuracy data for the test and transfer phases were modeled trial-wise using binomial generalized linear mixed effect regression models. These models included trial number, presentation format, learning mode, and the presentation format by learning mode interaction as fixed effects. The random effects structure was determined through model comparison via Akaike's Information Criterion starting with the maximal model. The model failed to converge with the inclusion of random slopes under the random effect for item, so the final random effects structure included only random intercepts for participant and item. As we have specific questions/hypotheses we seek to test in the present work, the model employs treatment contrasts, the reference group of which is shifted to address these different hypotheses. The initial model included same-category comparison under the observational mode; this speaks to the mode effect for same-class pairs, whether same-class pairs provides an advantage over one-at-a-time presentations in the observational mode, and the interaction between these factors. The model was then relevelled such that the classification mode was in the reference to assess for a corresponding comparison between same-category classification and single-item classification. The same process was repeated, but with mixed pairs in the reference to address our questions relevant to the mixed pairs group.

Another important consideration – for both the current and subsequent experiments – is the issue of power. How many participants are required to achieve appropriate power for each item type individually (old, new, and transfer) and across item types (i.e., all item types collapsed)? Our core interest is in whether the classification mode is harmful to same-category comparison. Using the SIMR package (Green & MacLeod, 2016) for R (R Core Team, 2015), we fit four models to the data from the same-category pairs group – one each for the old, new, and transfer item subsets as well as one to the full assessment-phase data, collapsed across item

type. Each model included mode as a fixed effect and participant and item as random effects. We applied the SIMR powerCurve function to each of these models, setting the number of simulations to 500. The function: (1) simulates new data from the original model, (2) fits a new model on the simulated data; and, (3) conducts a test for significance. The function repeats this process 500 times at each of ten levels of n per cell (3, 9, 16, 22, 29, 35, 42, 48, 55, and 61) to determine the power at each sample size. The simulations showed adequate power ($> 80\%$) for the collapsed data at the sample size we used, but indicated the individual item types were somewhat underpowered (70%, 76%, and 69% on old, new, and transfer items respectively; see Appendix A-D for power curves). To achieve 80% power on the individual item types, approximately 42, 36, and 39 subjects per cell are needed for old, new, and transfer items respectively (which we satisfy in Experiment 2).

However, given our prediction is that the classification mode is disruptive to same-class comparison learning – and not that the effect should manifest focally on one or another item type – the collapsed data are sufficient. Though somewhat underpowered, we offer the item type analyses as well – though a degree of caution should be taken in their interpretation.

Effects of learning mode. Consistent with our predictions, the same-category comparison group performed reliably better in observational mode than in classification mode – i.e., with the guess-and-correct cycle removed. We observed advantages for the collapsed data (observational $>$ classification; $\beta = 0.86$, $SE = 0.24$, Wald $Z = 3.53$, $p = .0004$), as well as for the old items (observational $>$ classification; $\beta = 1.15$, $SE = 0.31$, Wald $Z = 3.67$, $p = .0002$), new items (observational $>$ classification; $\beta = 0.89$, $SE = 0.25$, Wald $Z = 3.55$, $p = .0004$), and far transfer items (observational $>$ classification; $\beta = 0.85$, $SE = 0.28$, Wald $Z = 3.01$, $p = .003$). This finding strongly suggests that the guess-and-correct cycle obstructs the benefits of same-category

comparison. In evaluating same-category comparison against the single-item control, the enhanced performance for the observational same-category comparison learners led to a significant learning mode (classification, observational) by presentation format (single-item, same-cat comparison) interaction for the collapsed data ($\beta = 0.88$, $SE = 0.35$, Wald $Z = 2.51$, $p = .01$), and also for old items ($\beta = 1.22$, $SE = 0.44$, Wald $Z = 2.77$, $p = .006$), new items ($\beta = 1.05$, $SE = 0.35$, Wald $Z = 3.00$, $p = .003$), and transfer items ($\beta = 0.82$, $SE = 0.40$, Wald $Z = 2.03$, $p = .04$). Across each dependent measure, the interaction was characterized by a significant comparison advantage under the observational mode (collapsed: $\beta = 0.55$, $SE = 0.25$, Wald $Z = -2.23$, $p = .03$; old items: $\beta = 0.75$, $SE = 0.32$, Wald $Z = -2.32$, $p = .02$; new items: $\beta = 0.79$, $SE = 0.25$, Wald $Z = 3.11$, $p = .002$; transfer items: $\beta = 0.65$, $SE = 0.29$, Wald $Z = 2.26$, $p = .024$), but not under the classification mode ($ps > .11$).

These results are compelling for several reasons. First, these findings clearly show that same-category comparison is an effective way to learn relational categories – as expected under the theoretical framework of comparison developed in the study of analogy (Gentner, 1983, 2010; Gentner & Markman, 1997). By bringing shared relational structure between examples into alignment, learners were better able to discover and become knowledgeable about deep, relational properties that defined each category; this greater knowledge led to substantial performance advantages on all of our dependent measures. As we have made clear, learning mode critically determines the impact of same-category comparison. The present results support the interpretation that the guess-and-correct cycle of classification learning is disruptive to fruitful same-category comparison, and this clarifies previous failures to find comparison effects in the relational category learning literature (Kurtz & Boukrina, 2004; Kurtz & Gentner, 1998). Second, previous work with attribute categories has shown pure observational/passive learning to

be either equivalent or disadvantaged relative to classification/feedback learning when category membership is the target of assessment (Ashby, Maddox, & Bohil, 2002; Edmunds, Milton, & Wills., 2015; Estes, 1994; Levering & Kurtz, 2015; Thai, Krasne, & Kellman, 2015). Although a mixture of feedback and passive learning has been found to produce learning enhancements over either mode independently (Thai et al., 2015), we demonstrate for the first time that learning solely in the observational mode has outperformed feedback training on a test of category membership. The observational mode did not provide a generally stronger platform for relational category learning, as the single-item groups did not differ as a function of learning mode on any of the dependent measures; $ps > .4$. This shows that comparison processes, specifically, are interrupted by the classification mode and that the observational mode benefits comparison by removing this interruption.

Effects of mixed-pairs comparison. With regard to the conceptual replication of prior work using mixed pairs, we found that the mixed pairs learners did not differ from single-item learners in either mode on any of the item types individually or collapsed across item types ($ps > .17$). As for the extension to learning mode, we observed only a marginal advantage of the observational mode for the mixed pairs group on old items (observational > classification; $\beta = 0.56$, $SE = 0.30$, Wald $Z = 1.83$, $p = .07$), however this did not carry through to the new or transfer items – nor did the collapsed data reflect this difference. The failure to replicate the findings of Kurtz et al. (2013) suggests that the mixed-pairs effect is sensitive to variation in operationalization (while another possibility is that one or the other result is anomalous). We consider two notable procedural differences that may underlie the failure to replicate. First, to avoid transfer appropriate processing advantages for classification conditions in the current work, we utilized a category endorsement task. This contrasts with the classification assessment

used by Kurtz et al. (2013). Endorsement is highly comparable to classification; however, it is possible that the previously observed mixed pairs effect hinged critically on similarity between the training and testing tasks. This seems unlikely since the training classification trials included two items and the test classification trials included just one.

The other notable difference is that the comparison orienting task instructions used on each trial in Kurtz et al. were omitted in the current study. Specifically, on each trial, the instructions told participants to find a rock in one of the arrangements, consider the role it played within the arrangement, and to find a rock that played a corresponding role in the other arrangement. To address the concern that the orienting instructions might be instigating an object/attribute bias under same-category comparison, we dropped these instructions from the current design and opted for a subtler and more generic invitation to compare items in the pre-training instructions. It is plausible that without these trial-to-trial orienting instructions the power of mixed pairs comparison was lost – particularly if the instructions help to orient learners to the type of comparison they should engage in (i.e., same- or different-class comparison), given the random alternation between the two types. While we do not have a direct experimental test at hand, the present results suggest that the mixed-pairs approach may be effective as long as it is combined with a direct invitation to compare. Given the impressive results under the observational mode, we were less interested in tracking down the circumstances under which the mixed-pair approach succeeds. Instead we elected to pursue two goals in the subsequent study: (1) replication of the observational advantage and (2) a more direct examination of the critical component that varies between same-category and mixed-category learning: the impact of contrasting or different-category pairs.

Experiment 2

The observational learning mode played a vital role in the success of same-category comparison, but it did not exert a reliable effect on mixed-pairs learning. This is somewhat surprising given that half of the trials were same-category comparisons. The present study is designed as a replication of the critical finding of Experiment 1 (observational advantage for same-category pairs) and, additionally, we look to understand the isolated impact of the different-category pairs that constitute the other half of the mixed-pair trials. This is expected to provide an important foundation for future research addressing when and why a combination of same- and different-category pairs is effective.

As such, we utilized a 2 (learning mode: classification, observational) x 2 (comparison type: same-category, different-category) design. This design allows three things: (1) a replication of the observational advantage for same-category comparison; (2) a contribution to an emerging literature investigating same-category versus different-category comparison for relational category learning (e.g., Corral et al., 2018; Higgins, 2017; Namy & Clepper, 2010); and (3) an evaluation of how learning mode specifically impacts different-category comparison.

Co-presentation of examples from different categories does not provide a pathway to highlighting and abstraction of common structure. Instead, per Structure-Mapping theory, we can expect highlighting of alignable differences. Alignable differences (Markman & Gentner, 1993b) are differences that are tied to a common relational structure shared by two compared cases. In the categories we use, there are likely many alignable differences between any two exemplars from different categories. However, very few of these differences are tied to the structure that defines each category. Given this, detection of alignable differences between categories should not aid the discovery of the relational core defining each class. This suggests that different-category comparison should produce similar performance levels to same-category comparison

under the classification mode and single-item learning; and further that there is no reason to expect different-category comparison to become effective with the removal of the guess-and-correct cycle. On the other hand, if relational categories are importantly rule-like in nature, one might reasonably expect repeated opportunities to directly compare one positive and one negative example of each rule to be highly effective. Rather than identification of similar structure within a category, the benefit would accrue from identification of differentiated structure between categories – which can become a source of hypotheses (and a basis for hypothesis testing) about the relational basis underlying the categories. Further, if such category contrasts offer a separate path for comparison-based relational category learning, two further expectations arise: (1) once again, the observational mode ought to produce an advantage by keeping open the door to comparative processing that we believe gets closed off by the presence of the guess-and-correct cycle; and (2) it is possible to interpret the relatively poor performance observed in the mixed-pairs group as arising from the two pathways working poorly in conjunction – perhaps being undermined by the limited opportunities for each or the switching costs between the two (a question for future research).

Method

Participants. 200 undergraduates from Binghamton University participated for partial course credit.

Materials and procedure. The materials and procedure matched the comparison conditions of Experiment 1 with the only difference being that the different-category comparison condition used random pairs constrained to be from different categories. The design included four conditions based on crossing two factors: same-category classification ($n = 49$), same-category observational ($n = 48$), different-category classification ($n = 51$), and different-category

observational ($n = 52$). These cell counts provide sufficient power to both the collapsed data (94%) and each item type individually (between 86.2% and 91%).

Results and Discussion

Learning phase. While the primary focus is again on performance at the test phase, we first address performance and time on task data from the learning phase for classification learners. Accuracy data were modeled trial-wise using linear mixed effects regression. Keeping consistent with Experiment 1, trial number and presentation format were included as fixed effects and subject was included as a random effect. Accuracy data indicated that both classification learning conditions led to levels of performance that exceeded chance ($ps < .05$). However, the same-category ($M = 0.60$, $SE = 0.02$) and different-category ($M = 0.59$, $SE = 0.03$) conditions did not differ in learning accuracy. Time on task was operationalized and analyzed as in Experiment 1 and a similar pattern was found with observational learners ($M = 5.32$, $SE = 0.42$) allocating a reliably smaller amount of time to each trial relative to classification learners ($M = 6.90$, $SE = 0.43$); $p < .01$. This was driven by a reliable difference on time on task for different-category pairs (observational: $M = 5.45$, $SE = 0.59$; classification: $M = 7.23$, $SE = 0.61$; $p < .05$) and a trend in the same direction for same-category pairs (observational: $M = 5.18$, $SE = 0.59$; classification: $M = 6.57$, $SE = 0.60$; $p = .10$). As such, and as in Experiment 1, any performance advantage for the observational group cannot be attributed to greater processing time. Further, comparison type did not lead to reliable differences in time-on-task ($ps > .4$).

Test phase. Modeling and analyses for the test phase mirrored those used in Experiment 1. Adjusted means and standard errors can be seen in Table 2 and plots for old item, new item, and transfer performance are shown in Figures 5-7. In accord with our first goal of replicating the critical finding of Experiment 1, a significant advantage for observational learning was once

again found over classification learning for the combined data (observational > classification; $\beta = 0.61$, $SE = 0.19$, Wald $Z = 3.22$, $p = .001$) and across old items (observational > classification; $\beta = 0.55$, $SE = 0.22$, Wald $Z = 2.56$, $p = .01$), new items (observational > classification; $\beta = 0.53$, $SE = 0.20$, Wald $Z = 2.65$, $p = .008$), and transfer items (observational > classification; $\beta = 0.64$, $SE = 0.22$, Wald $Z = 2.87$, $p = .004$).

The second goal of this design was to evaluate learning from co-presented pairs uniformly in the same category versus uniformly in different categories. Across all dependent measures, no reliable differences were found under either learning mode ($ps \geq .2$). This suggests that it was not the different-category trials per se that negatively impacted the performance of mixed pairs in Experiment 1. Instead, it seems likely that the mixed-pair learning phase represents a case in which doing some of two equally beneficial things is less beneficial than sticking with one.

While we do not wish to make too much of a null finding, we note that the lack of a difference between the two types of comparison is consistent with pilot data in our laboratory – however, these findings do not fit cleanly in the existing literature. On a theoretical level, an emphasis on schema abstraction via comparison-based learning would perhaps suggest an advantage for same-category over different-category comparison. However, on an empirical level, a recent set of results showed the exact opposite – consistent advantages for different- over same-category comparison in a variety of relational category learning domains (Corral et al., 2018).

The difference in results between Corral et al. and the current work can perhaps be understood as reflecting important differences in the set-up of the category learning tasks. One difference is that the rock arrangement domain is not clearly alignable across categories. Given that contrastive comparison has been shown to facilitate noticing of key alignable differences

with highly alignable categories (e.g., positive feedback loop vs. negative feedback loop; Smith & Gentner, 2014), contrastive comparison may have facilitated the acquisition of these key differences in Corral et al. (2018). A second notable difference is our use of a three-way classification task. In two-way classification, the different-category pair on each trial always makes the same distinction between the same two classes – and further, when the learner knows the category of one item in a pair, the category of the other item can be directly inferred. In the three-class case, these advantages do not hold – thereby limiting the power of contrastive comparison-based learning. Future work will help to determine the role of these factors and to solidify a theoretical basis for understanding the benefits of contrastive comparison in relational category learning (particularly in the non-alignable case; though see Corral et al., 2018, for discussion).

We have seen a replication of the observational advantage for same-category comparison and no effect of type of comparison. The remaining question is how learning mode affects different-category comparison. We found that the different-category observational group reliably outperformed its classification counterpart group across all dependent measures: collapsed data ($\beta = 0.60$, $SE = 0.19$, Wald $Z = 3.20$, $p = .001$), old items ($\beta = 0.71$, $SE = 0.22$, Wald $Z = 3.30$, $p = .001$), new items ($\beta = 0.59$, $SE = 0.20$, Wald $Z = 2.98$, $p = .003$), and transfer items ($\beta = 0.52$, $SE = 0.22$, Wald $Z = 2.37$, $p = .02$). This is consistent with our primary theme: the guess-and-correct cycle restricts the potential to take advantage of comparison opportunities – whether the comparison benefits that arise are from identifying structural commonalities within a category or identifying structural differences between categories.

Experiment 3

Across two experiments, we have developed evidence that the classification mode

disrupts the benefits of comparison-based learning. In this experiment, we ask: what exactly drives the considerable difference in test performance between the classification and observational learning modes? The time-on-task data tell us it is not a matter of exposure or processing time. One evident difference between the modes is that classification is less passive since it requires a response and outcome (i.e., the guess-and-correct cycle), while the observational task just involves studying the provided information. However, we know that the two learning modes are informationally equivalent and equally effective in the single-item case (without comparison opportunities), so there is little reason to believe it is something intrinsic to the modes themselves. Instead, the action is in the interplay between the category learning mode and the comparison opportunity.

The core hypothesis of these experiments has been that the guess-and-correct cycle of the classification mode creates a performance orientation that directs learners to focus more on generating their guesses, the correctness of their responses, and their overall success level – at the expense of a more broad-based effort to learn about the categories and draw upon comparison opportunities (for more on learning mode and generative versus discriminative human category learning, see Kurtz, 2015; Levering & Kurtz, 2015). Besides performance orientation though, there is another simple difference between the modes that may explain greater utilization of comparison opportunities under the observational mode: the latency of label presentation. With observational learning, the presentation of category labels occurs early in the trial, and this may stand as an invitation through language to compare. A range of empirical work shows that symbolic juxtaposition, i.e., the sharing of a common label, can promote comparison (e.g., Christie & Gentner, 2014; Kotovsky & Gentner, 1996). By contrast, classification learners do not see the correct labels until the feedback period at the end of each trial. Thus, an alternative

hypothesis is that label delay, either solely or in concert with performance orientation, leads to poorer utilization of comparison opportunities and poorer learning.

To assess the impact of these two potential drivers on same-category comparison, we introduce two new learning conditions that are also informationally equivalent but occupy intermediate positions on the continuum between the classification and observational modes. To ask our primary question of interest (whether poorer performance in the classification mode is attributable to the performance orientation of the task) we compare our first novel condition – covert classification (CovClass), or classification without a submitted response – to the classification mode. This condition is just like standard classification learning except learners are instructed to make a private guess about the category membership of the exemplars on each trial rather than submitting an overt response. In this condition, learners are shown the pair of unlabeled exemplars and the category options and are prompted to make an internal guess about the items' membership. After they make their guess, participants click anywhere on the screen to view the correct category labels for a brief, fixed period and are left to self-evaluate their accuracy before the next trial begins. Given that the learner knows that the experiment does not record their category decisions during learning (and they are free to not even make an internal guess if they do not wish to), this CovClass condition should effectively reduce the performance focus. When the labels arrive, it is less in this case about being right or wrong than it is about getting useful information. On the view that the perceived centrality of the guess-and-correct component competes with learners' engagement in comparison, we expected better test performance in the CovClass condition relative to standard classification – and greater similarity to the observational group.

To evaluate the label delay hypothesis for the observational advantage, we compare our

second new condition – observational with delayed labels (ObsDelay) – to standard observational learning. The ObsDelay condition is just like observational learning but has the key difference of the label presentation being delayed until the end of the self-paced trial. In this condition, learners study the pair of unlabeled exemplars in each trial for as long as they wish; when they are finished studying they click to see the category labels for a brief, fixed period before the next trial begins. Thus, this condition has the same label delay as classification learning, but does not have a guess-and-correct cycle, like standard observational learning. If the ObsDelay group is worse at test than its standard observational control group, this would suggest that having the labels early in the trial is an important facet of the mode difference. Given previous literature showing benefits of labels (e.g., Christie & Gentner, 2014; Davidson & Gelman, 1990; Gentner & Namy, 1999; Kotovsky & Gentner, 1996; see Gentner, 2016), we expected that withholding label presentation until the end of the trial might reduce comparison engagement to some extent – evidenced by lower performance relative to standard observational training. While we did expect label delay to exert some effect, we expected a more distinguished effect of response removal (i.e., performance orientation).

Method

Participants. 142 undergraduates from Binghamton University participated for partial course credit.

Materials and procedure. The materials were the same as those of Experiments 1 and 2. Participants were randomly assigned to one of four learning mode conditions in a between-subjects design. All four conditions employed same-category comparison: observational ($n = 34$), observational with delayed labels ($n = 36$), covert classification ($n = 36$), and classification ($n = 36$). Participants received the same archeological cover story and task framing (i.e., that

their overall goal was to figure out what makes a given rock arrangement belong to one of the three categories and that they would be tested on their knowledge of each category later) given in Experiments 1 and 2. In the pre-training instructions, all participants were told they would see two arrangements on each trial. Following this instruction, learners received condition-specific instructions that prepared them for their particular condition. Classification learners were told that they were to figure out the correct type for the arrangements on each trial and to select a response. CovClass learners were told they were to make a guess in their head about the type the examples belonged to; after making their guess they were to click anywhere on the screen to see the exemplars' type. ObsDelay learners were told they were to click when they were ready to continue, but that they would see the correct type for each arrangement before the arrangements were removed. Observational learners were simply told they would see the correct type for each arrangement. Following these condition-specific instructions, all conditions were given the same reminder given in Experiments 1 and 2: "Remember that there are three different styles for arranging the rocks into configurations. Looking at the two arrangements together can help you learn these types. Try your best to gain mastery of the names of each type and what makes an arrangement belong to those types. Learn as much as you can before the test!"

Learning. As in Experiments 1 and 2, the training phase consisted of two cycles of 12 paired stimulus trials, totaling 48 stimulus exposures. At the beginning of each trial, two side-by-side stimuli were presented and these remained visible until the trial was complete (see Figure 8 for a schematic of the procedure for each condition). A series of modifications were made to the procedure of the classification task (relative to Experiments 1 and 2) in order to allow the cleanest comparison across the conditions in the present design. Given our interest in label delay, we presented labels at stimulus/trial onset to the observational condition – instead of at the

500ms delay used in the former experiments – to have a clear distinction between immediate and delayed label presentation groups, rather than groups with differing degrees of delay. Similarly, we presented the query and category response buttons at stimulus/trial onset to the classification and CovClass conditions as well to bring them into correspondence.

In the classification group, participants were queried for a joint response for both stimuli; time to make a decision was unconstrained. The participant selected a response by clicking one of three buttons that corresponded to the categories in the learning domain. The use of a joint response for the two examples represents a deviation from Experiments 1 and 2. Given that learners pick up quickly in a same-category paradigm that the items will be from the same category each time, we shifted to a joint response to remove the annoyance of making two identical responses each trial and to eliminate variability in when participants noticed this regularity; preliminary work showed performance under joint and independent responses to be nearly identical. When classification learners selected their response, the button they clicked showed them visual confirmation of their selection. Following the classification decision, the correct category labels were shown for 1s – leaving participants to self-evaluate the accuracy of their decisions. Unlike the previous experiments, the feedback included only the category label, not whether the response was correct or incorrect, in order to equate the label/feedback experience across conditions as closely as possible without altering the fundamental nature of the task. Given the importance of label presentation timing in this design, we fixed the duration of feedback to 1s.

The procedure for the CovClass group was just like the classification group except that they were instructed to make an internal guess (rather than submitting a response) at each prompt and, then, to click the mouse to see the correct category labels for 1s. The category buttons,

though not functional, were presented on screen as a reminder of the labels and for consistency with the standard classification condition. If a learner did click on the category buttons, nothing happened (i.e., there was no visual confirmation of a selection). We note that the CovClass procedure left participants the option to disregard the instructions and not make a guess on any trial (if they fully disregard the instruction to make a guess, then it is essentially the ObsDelay condition) – we relied on the natural tendency to make a prediction in this type of task as well as the tendency among participants to attempt to follow experimenter instructions that are not especially arduous.

The ObsDelay condition followed exactly the same procedure as the CovClass condition except participants were not asked to make guesses and were neither presented with a category query nor category buttons. Participants were able to study the exemplars in an unlabeled state for as long as they wished and then clicked to see the category labels for 1s.

The observational condition was just like that in Experiments 1 and 2 but with the label-onset exception noted above – i.e., labels were presented with the examples at trial onset instead of at a 500ms delay. The common assessment phase was conducted just as in Experiments 1 and 2.

Results and Discussion

Learning phase. Learning phase accuracy data for the classification learning group (the only group that makes a recorded response) showed a level of performance that was significantly better ($M = .57$, $SE = .02$; $p < .05$) than chance (.33). As in the experiments above, we looked at time-on-task by taking the median time-on-task for each participant and predicting them with condition in a linear regression. We found that all other conditions (Observational: $M = 4.03$, $SE = .35$; ObsDelay: $M = 4.58$, $SE = .37$; CovClass: $M = 3.86$, $SE = .23$) spent significantly more

time on each trial compared to classification learners ($M = 2.64$, $SE = .15$). This is the opposite of Experiments 1 and 2, but this is attributable to the change in task format from two classification decisions to one on each trial for classification learners. It should be noted that this time-on-task reversal is informative. Although a potential concern noted in the previous experiments was that observational learners might spend more time on each trial and that greater processing time might explain the advantage, the opposite concern is also valid – that shorter trial durations could aid learning, perhaps by permitting better comparisons to be made across trials. Thus, if shorter time-on-task values explain the observational advantages observed above, then the classification group ought to outperform the observational group here. However, as seen below, this is not the case.

Test phase. As in the above experiments, trial-wise accuracy data were modeled using binomial generalized linear mixed effect regressions. Models included trial number and mode condition as fixed effects. The random effects structure was fit using performance on the old item test data, starting with a maximal random effects structure (random intercept for participant, random slope and intercept for item). As with Experiments 1 and 2, inclusion of the random slope impeded model convergence; the final random effects structure included only random intercepts for participant and item. Adjusted means and standard errors can be seen in Table 3. Figures 9-11 illustrate performance on old items, new items, and transfer items, respectively. We report on the collapsed data as well as the individual item types below. The simulations derived from Experiment 1 suggest the collapsed data are well-powered (~87%). Item type analyses achieved respectable power, though slightly short of convention in some cases (76.4%, 80.5%, 76.3% for old, new, and transfer items respectively).

The observational advantage. Replication of the core mode effect is of critical importance, given the slight alterations that were made to the procedure (i.e., participants provide only a single guess for both items and participants self-evaluate their accuracy given the correct label). As preliminary work suggested, and as was seen in the above experiments, we again found better performance for learners that received same-category comparison opportunities in the observational mode, relative to those trained with the classification mode. The advantage was evidenced by a reliable effect for the collapsed data ($\beta = 0.67$, $SE = 0.25$, Wald $Z = 2.70$, $p = .007$), as well as for old items ($\beta = 0.65$, $SE = 0.29$, Wald $Z = 2.24$, $p = .03$), a trend for new items ($\beta = 0.41$, $SE = 0.24$, Wald $Z = 1.76$, $p = .08$), and a reliable advantage at transfer ($\beta = 0.70$, $SE = 0.26$, Wald $Z = 2.63$, $p = .009$). These findings show a degree of generalizability of the learning mode effect, taking into account the minor changes made, and provide confidence in the adjusted paradigm for evaluating the two novel conditions.

Effect of delayed label presentation. With confidence in our paradigm, we now turn to the primary goal of this experiment: evaluating the performance-orientation and label-delay accounts of the observational advantage. We first approach the issue of whether the observational advantage can be explained by a greater invitation to compare through language, facilitated by the immediacy of label presentation. If initial labels invite superior comparison engagement relative to final labels, we should see poorer test performance for the ObsDelay group relative to the standard observational group.

Accuracy data at test showed the ObsDelay group did not differ reliably from standard observational on old items ($p = .14$), new items ($p = .34$), or transfer items ($\beta = -0.51$, $SE = 0.27$, Wald $Z = -1.93$, $p = .054$) – nor did it differ on the collapsed set ($\beta = -0.46$, $SE = 0.25$, Wald $Z = -1.85$, $p = .064$). The transfer result could be interpreted as a marginal effect (with higher

accuracy in the standard observational condition). With considerable caution, this could reflect weaker comparison opportunities (and stifled abstraction) as a function of having briefly presented labels presented at the end of the trial.

To get additional clarity, we look at how the ObsDelay group fares against standard classification. Although the standard observational group broadly outperformed standard classification, it's delayed label counterpart did not reliably differ from classification on any of our dependent measures ($ps > .39$). Looking at these data collectively, they suggest that the presentation of class labels later in the trial timeline impairs the effectiveness of the observational mode to some degree – which may in part account for the observational advantage. Though these findings relate to some extent to the labeled comparison literature (e.g., Christie & Gentner, 2014; Davidson & Gelman, 1990; Gentner & Namy, 1999; Kotovsky & Gentner, 1996), they do not speak to the presence/absence of labels, as this literature is typically concerned with. While not wishing to make much of a null result, our results suggest a novel qualification to these findings – that it might not merely be a matter of *if* a label is provided, but also *when* that might dictate whether a label serves as an invitation to compare. We see this as a useful area for future research.

Effect of explicit guessing and response collection. Although we found some evidence suggesting that label delay contributes to the mode difference, the relatively weak evidence makes it difficult to accept as the key driver. How then does reducing the performance emphasis of the classification mode impact comparison-based learning? We speculated that learners may focus on the guess-and-correct component of the trial in standard classification at the expense of the comparison opportunity and that, by softening the performance emphasis through making internal, unforced guesses, the CovClass group would engage in deeper comparisons and

demonstrate better performance. The collapsed data did not show a reliable difference between on the two conditions ($\beta = 0.40$, $SE = 0.24$, Wald $Z = 1.65$, $p = .099$) – though they suggested a possible trend. Looking at the item types individually, we found that the two conditions did not differ on old items ($p = .16$) or new items ($p = .27$) – though the CovClass group showed a numerical advantage on both. However, at transfer the CovClass group exhibited reliably higher accuracy than standard classification ($\beta = 0.54$, $SE = 0.26$, Wald $Z = 2.10$, $p = .035$). This indicates that reducing the performance focus of the task environment could have allowed learners to more deeply engage in and benefit from comparison.

The effect was not seen across the board, but it is quite possible that far transfer is a more sensitive dependent measure, i.e., one that depends more heavily on the degree of successful abstraction. In conjunction with this finding, we also note there were no differences between the CovClass and standard observational groups on any of our test measures ($ps > .39$) which prompted us to conduct two one-sided test (TOST) for equivalence. Using a region of similarity equal to six percent ($\epsilon = .06$), we found a trend for equivalence on both the collapsed item set ($p = .069$) and old items ($p = .10$), as well as reliable equivalence on both new and transfer items ($ps < .05$). With the same parameters, we then tested the equivalence of the observational group to all other groups; it was not found to be equivalent to any other condition on any of our measures ($ps > .13$). To be clear, our intent with this analysis is not to assert that the CovClass group is learning in a similar way to the standard observational group – the predictive focus of classification is almost certainly distinct. Rather, we argue that relaxing the task's emphasis on performance facilitated the making of higher quality comparisons – leading to a degree of abstraction and performance comparable to the observational group.

We must also consider whether there are alternate explanations of the covert classification advantage – could it be better to *not* have to fully formulate, explicitly commit to, and execute a response based on one’s guess? A memory-based perspective suggests that under some circumstances a wrong guess may produce a competing association (Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Vaughn & Rawson, 2012) – and it is conceivable that greater commitment to that guess might strengthen the association. However, if removing inaccurate associations explained the benefit of the observational mode, we would expect the observational mode to *generally* lead to advantages over classification on tests of category membership – this is not seen in our single-item learning results, nor in the attribute category learning literature (Ashby et al., 2002; Edmunds et al., 2014; Estes, 1994; Levering & Kurtz, 2015). Another possibility is that greater time spent on task by CovClass learners might explain the better performance. However, as we have seen across the all three experiments, time-on-task has not provided a strong correspondence with accuracy differences. Rather, what happens during the trial appears paramount.

Finally, we note that the two new learning conditions introduced in the present experiment depended on learners doing what we asked of them. As such, we can neither verify that CovClass learners made predictions, nor verify that ObsDelay learners refrained from making any. However, the pattern of evidence suggests learners largely completed the task as asked. Had ObsDelay learners made spontaneous class predictions of their own accord without instruction to do so, the ObsDelay task would transform into a classification mode, but a covert one. Additionally, had CovClass learners skimmed on their class predictions, the task would effectively reduce to ObsDelay (with the addition of category buttons on the screen). Thus, the effects observed above can be seen to reflect participants engaging in different activities. An

added consideration is that, although the data suggest participants generally complied with the instructions, it is likely some made spontaneous predictions or refrained from making any. To the extent this is true, the performance gains (CovClass) and decrements (ObsDelay) of the manipulations may be under-approximated.

General Discussion

Across three experiments, we explored the role of learning mode in relational category learning. This investigation followed from a series of failed attempts to find a theoretically predicted advantage for same-category comparison over one-at-a-time presentations. In the above studies, we used a learning mode manipulation in order to: (1) examine predictions from the structural alignment view of an advantage for same-category comparison over single-stimulus presentations; (2) evaluate the impact that removing the guess-and-correct cycle has on relational category learning more broadly by testing its effect with or without comparison and under different types of comparison opportunities; and (3) isolate the mechanics of how the guess-and-correct cycle might affect the success of comparison.

Regarding our revamped effort to test predictions from the structural alignment view, we found strong evidence in Experiment 1 that same-category comparison leads to a theoretically-predicted advantage over sequential item presentations. However, we found that learning mode plays a critical role in whether this advantage emerges. Specifically, when same-class comparisons were made under the classification mode, comparison opportunities did not provide any benefit – consistent with previous work (Kurtz & Boukrina, 2004; Kurtz & Gentner, 1998; unpublished pilot data for the present investigation). However, when learners engaged with comparison opportunities in the observational learning mode, we found the elusive advantage

over single-item presentations and unprecedented accuracy levels on both within- and across-domain tests of knowledge.

These findings serve as a resolution to the tension between the straightforward predictions of the structural alignment view and the previous failures to find same-category comparison effects in relational category learning studies. To the best of our knowledge, this is the first time a benefit of pure same-category pairs (over single-item) has been shown within the context of an inductive relational category learning paradigm. This finding thus contributes to existing evidence that the structural alignment process is integral to the learning and transfer of relational categories (Kurtz et al., 2013) and indicates that Structure-Mapping theory (Gentner, 1983, 2003) is a useful framework through which to understand relational category learning. This however comes with an obvious, but notable, caveat: the theory can only hold explanatory value within a context that is hospitable to engaged alignment. We explore this point further below.

To place the present findings in a fuller context, the notion that the guess-and-correct task disrupts successful comparison processing is consistent with other findings in the existing literature. One potential aspect of this is a resource shortfall caused by combining guess-and-correct with comparison. Competition for different types of resources in a dual-task load paradigm (Waltz, Lau, Grewal, & Holyoak, 2000) or a prior anxiety-inducing task (Tohill & Holyoak, 2000) have been shown to lead to diminished relational processing. Our argument is not that relational category learning is *impossible* due to a resource limitation; this follows from comparison benefits under the classification learning mode when explicit instructions to focus on the comparison component are provided (Kurtz et al., 2013). Instead, we contend that when comparison and classification-based category learning are combined – without forcing attention

to the comparison piece – the allocation of resources is not favorable for garnering comparison-based outcomes. An important question embedded in this is whether allocation issues arise from the guess-and-correct cycle generally or, more specifically, guess-and-correct when the task goal is classification. Evidence from several research groups shows facilitated relational category learning under an inference-based learning mode (Erikson, Chin-Parker, & Ross, 2005; Goldwater, Don, Krusche, & Livesey, 2018; Higgins, 2017). The inference task (receiving partially-specified category members and predicting the missing information) is supervised and involves a guess-and-correct cycle, but differs from classification learning in promoting broad-based, generative-style knowledge of each category rather than invoking a discriminative-style, narrow focus on diagnostic features that minimally separate the categories. We have argued that both observational and inference learning tasks are notably more generative in nature than classification learning (Kurtz, 2015; Levering & Kurtz, 2015). Collectively this highlights two competing accounts that should be tested going forward. One is that the guess-and-correct cycle diverts resources from comparison irrespective of the type of knowledge engendered by the task. Another is that the discriminative, diagnostic knowledge encouraged by classification is less useful for comparison. Under this view, guess-and-correct – for the goal of classifying – selectively diverts resources from comparison. Examining performance when comparison is nested in the classification, inference, or observational modes will help deliberate between these accounts.

Despite being born out of an interest to determine the efficacy of same-class comparisons, the present line of inquiry represents a broader effort to understand how relational category learning proceeds under different learning circumstances. An important part of this effort is determining if the selected learning modes differ under standard format: one-at-a-time

presentations. Although learning mode played an important role for the comparison formats, we failed to find an effect of mode on any of our measures under the single-item format. We note that the lack of a mode effect serves as an important qualifier to the observational advantage seen under the comparison formats. An alternative interpretation of the observational advantage might have been that the observational mode is generally better than the classification mode for relational category learning. However, the lack of an observational advantage under single-item indicates that is not the case; the observational mode served not as a general prop to learning but rather a prop to comparison learning specifically.

We also note that, while our investigation of the classification and observational modes under one-at-a-time presentations is the first to do so with relational categories as the target of learning, many comparisons between these learning modes have been made in the attribute category learning literature (e.g., Ashby et al., 2002; Edmunds et al., 2015; Estes, 1994; Levering & Kurtz, 2015; Thai et al., 2015). These comparisons have been made using different assessment measures, different category dimensionality, and different category structures – including rule-based structures (Ashby et al., 2002; Edmunds et al., 2015, Levering & Kurtz, 2015), which are perhaps the *most* relatable to relational categories. This research has been somewhat mixed. Some has shown no difference between observational and classification learning on tests of class membership (Ashby et al., 2002; Levering & Kurtz, 2015), while other research has found an advantage for classification over observational (Edmunds et al., 2015). In the cases of Ashby et al. (2002) and Levering and Kurtz (2015), the lack of difference between the modes was likely driven by a ceiling effect – attributable to the simplicity of the rule (a unidimensional rule) – which is not the case in the current work. However, Edmunds et al. (2015) used a more complex, conjunctive rule as the target of learning (avoiding ceiling performance) and showed an

advantage was tipped in favor of classification learning. Given the inherent differences between even rule-based attribute categories and relational categories (which are thought of as rule-like, Gentner & Kurtz, 2005) and the likely difficulty differences between the category structures used in the current work and previous attribute studies, we do not wish to make much of a cross-literature comparison. However, the available data suggest that which of these modes will lead to the best learning outcome under one-at-a-time presentations may depend on whether attribute or relational categories are the target of learning. This highlights a growing need for research relating and dissociating attribute and relational category learning, in the hopes of identifying a unified mechanistic account of category learning.

Another important facet of this investigation was examining the effects of learning mode on two additional types of comparison: pure different-category pairs and mixed pairs (50%/50%, same- and different-category). We found only a non-reliable, numerical advantage for the mixed pairs observational group over its classification complement. However, mode exerted a pronounced effect under different-category pairs – an observational advantage mirroring that found for same-class pairs. What does this common effect of the observational mode for both same- and different-category pairs say about the role mode is playing in comparison? For one, it indicates the benefit of the observational mode does not arise by affecting the particular mechanics of a specific type of comparison – i.e., it occurs regardless of whether commonality or difference finding is the goal. Instead, this pattern suggests the general mechanics of comparison (engaged, joint consideration) are being affected by the observational mode. This pattern is consistent with the central thesis of this paper – that the additional task of classification detracts from engaged comparison making – and suggests the manipulation worked as intended.

If the observational mode generally encourages joint consideration, the absence of a mode effect under mixed pairs is a seemingly puzzling fact to integrate. Given that the mixed pairs group is made up of equal parts same- and different-category comparison, one might expect mixed pairs to respond to the mode manipulation in a way that is similar to either of its parts alone. However, recall that the goal of the mode manipulation was to restore the primacy of comparison within the task environment. The relationship between the observational mode and improved learning outcomes should only exist in cases where deeper engagement in the underlying form of comparison holds the potential to improve learning – which we elaborate on further below.

We believe there are two related distinctions between mixed pairs learning and either of the uniform comparison learning formats that make mixed pairs a more challenging and less fruitful learning platform. First, there is a clear distinction in the strategies that are required; in either pure format, the learner consistently makes comparisons of a given type. In the mixed format, however, the learner must frequently switch between same- and different-category comparison. It should be expected that, although there are common mechanics in both types of comparison, the strategies in each differ; same-pairs comparisons should foster a commonality finding strategy while different-pairs should encourage a difference finding strategy. Second, it should be expected that the category makeup of a given pair influences the strategy for comparison. In either pure comparison format, there are three subtypes of comparison (AA, BB, CC, and AB, AC, BC, for same- and different-category conditions respectively). Under the mixed pairs format, by contrast, learners must contend with all six of those randomly interspersed subtypes during training, each of which requires its own strategies – i.e., which objects and relations are relevant to attend to, and which are not. In sum, mixed pairs learning –

in the flavor we investigate in this study – requires much more task switching than either pure variety of comparison. Given the costs to category learning known to arise from strategy switching (e.g., Crossley, Roeder, Helie, & Ashby, 2018; Erickson, 2008), we believe having to randomly switch between different comparison strategies incurs a comparable cost during mixed pairs learning. In further support of this point, previous work from our lab has shown that learning from a 75%/25% ratio of same- to different-category comparison did not differ from another condition with the same ratio flipped on far transfer items. However, both weighted conditions outperformed a condition that had a 50%/50% ratio (Patterson & Kurtz, 2014; see Figure 12). Collectively this work suggests that having greater consistency in the comparison strategies used during learning leads to better learning outcomes.

The present work also speaks to the relative efficacy of same-category and different-category comparisons for learning relational kinds. We did not find the type of comparison to exert an effect in the present work – comparable learning and transfer was found to arise from both types. In the case of between-category comparison, the advantage does not arise from aligning common structure between co-presented examples, given the many relevant but non-alignable relational differences between categories. There are several possibilities for what mechanism is at play. The first is a natural extension of structure-mapping theory in which the alignable differences are highlighted relative to common structure at the domain level (rather than category level). Another possibility consistent with Structure Mapping Theory is the idea that productive alignments are made between each item in a different-category comparison opportunity and its activated generic category representation. A third possibility, indirectly explained by the alignment view, is that different-category comparison may be effective through a *failure of alignment* (see Corral et al., 2018 for discussion). Under this view, contrast is an

eliminative process in which relations present in both co-presented exemplars are ruled out while relations contained by one exemplar, and not the other (a failure of alignment), are considered as potentially category defining. In all these possibilities, the observational mode benefits both different-category and same-category comparison because it is the comparison process that allows productive use of item co-presentation and the guess-and-correct cycle stands in the way of its engagement. There remains much theoretical and empirical work to be done on the mechanisms behind the efficacy of contrastive comparisons; deliberation between these, and other, possibilities represents an important area for further work.

Contrasting the present results with other work, Corral et al. (2018) found a marked advantage for different-category comparison across a series of experiments addressing two-choice relational category learning. While the relative efficacies of presentation type differ between these two investigations, we do not view the findings as conflicting. Instead, we find it likely that which type of comparison is best will be driven by the nature of the learning problem. First, we note that three-way classification creates a drive for positive (generative-style) category knowledge, while two-way classification can be more about finding a boundary (discriminative-style). With three-way classification there are as many boundaries as there are categories (unlike two-way classification). Further, two-way classification can succeed by mastering one category without requiring any knowledge at all of the alternative. The other major difference in the learning conditions is that the categories used by Corral et al. were quite similar to one another – with much overlapping, alignable structure between categories. Categories A and B were as follows across three experiments: “bigger object on the right” versus “smaller object on the right,” containment versus support, and “objects of same color also match in shape” versus “objects of same color also match in size.” In all cases, one or a few alignable differences

distinguish the categories (note that Corral et al., 2018, also offer a theoretical analysis in which relational learning tasks operate more like feature-based category learning if relational information is encoded as flat, rather than structured, representational content), and in this light, it is unsurprising that different-category comparisons facilitated better learning than same-category comparisons as they would better serve to highlight the small number of category-relevant differences (Carvalho & Goldstone, 2014; Markman & Gentner, 1993b; Smith & Gentner, 2014); learning purely through same-category comparisons, on the other hand, would make these category-relevant differences more difficult to detect.

In the present work, there is very little category-relevant structural overlap between categories – i.e., there are many relevant but non-alignable differences (i.e., not connected to common structure) between the relational structures that define each category. Under these circumstances, different-category comparison should be expected to be less effective, while same-category comparison should be more effective. This perhaps general principle of comparison-based learning is also reflected in feature-based category learning, where interleaving exemplars of different classes has been shown to benefit the learning under high between-category similarity while blocking exemplars of a given class aids the learning when there is high within-category similarity (e.g., Carvalho & Goldstone, 2014). While a direct comparison of the efficacy of same- and different-category comparison for learning high-similarity and low-similarity relational categories has not been empirically established, we see this as a promising avenue for further research.

There is another important reason for manipulating category similarity in future research: it may serve to broaden our understanding of the observational comparison advantage, across both attribute and relational categories. Though our study focuses on relational category learning,

the literature suggests that the observational over classification benefit may be general to both attribute and relational categories – though this benefit is not uniform (Carvalho & Goldstone, 2015). Carvalho and Goldstone (2015) crossed learning schedule (blocking vs. interleaving) with learning mode (observational vs. classification) in feature-based category learning. Importantly, they used categories that had high within-category similarity and high between-category similarity. They found an observational advantage at test for the blocking (sequential same-category comparison) group, but not for the interleaving group (in fact, a classification advantage was found). In a follow-up experiment using the same design, Carvalho and Goldstone (2015; Exp 2) maintained the high between-category similarity of the prior experiment – however, they used categories with low within-category similarity. In this case, the observational benefit for the blocking group went away.

While the data suggest that the observational comparison advantage takes place in both relational and attribute categories, it seems that there is a moderating factor in the similarity within and between categories. Based on the available evidence, we put forth a novel *comparison load hypothesis* that may direct future research on learning mode and comparison. The hypothesis asserts that the type of comparison (same- or different-category comparison) and the category structure (dimensionality and the degree of within- and between-category similarity) jointly determine the amount of information that must be encoded. Circumstances in which successful category learning depends on mapping and representing *many* elements during comparison (i.e., high comparison load) will benefit from the observational mode, as the mode does not itself incur additional load or a competing goal. Comparison load is high for same-category comparison when within-category similarity is high as there are a larger number of common elements between members that must be mapped, bound, and represented within the

comparison (relative to low similarity). Correspondingly, load is high for different-category comparison when between-category similarity is low, as a larger number of differences must be mapped and represented (relative to high similarity). The currently available data are consistent with this hypothesis, however future research that manipulates mode, comparison type (same/different-category comparison), within/between-category similarity, and category type (relational or attribute) will be necessary to test the hypothesis more fully.

Across three experiments, and under different kinds of comparison, we showed that the classification mode erodes the quality of comparison-based learning. Experiment 3 was aimed at elucidating why the classification mode plays this role. Specifically, we manipulated both label onset in the observational mode – a potential cue to compare – and the collection of responses in the classification mode – a potential cue to the importance of the guess-and-correct cycle within the learning task. Although we did not find reliable evidence that label onset delay explains the mode difference, we did show a reliable impact of response collection. This effect is consistent with the interpretation that classification, included as a concurrent task to comparison, alters the learner's prioritizations within the learning task – shifting relative priority to the guessing component and away from the comparison component. Finding that learning improves by removing response collection indicates, quite obviously, that what is perceived as central to the learner is shaped by what is (seemingly) important to the experimenter. Learners provide their response data and are given corrective feedback on that data, which establishes the guess-and-correct cycle as the task of chief importance. By prioritizing the guess-and-correct cycle over the available comparison opportunities, learning was stifled. However, when the response collection is removed – i.e., the task is without cues indicating the priority of the guess-and-correct cycle –

the relative importance of the two tasks is not established. As such, learners engaged more deeply in comparison and learning outcomes improved.

This work has notable implications. For basic research interests, our work clearly demonstrates the issues that arise from embedding comparison within a classification learning task. This work thus prescribes that future studies that seek to evaluate the efficacy of comparison relative to other learning conditions should refrain from nesting it within the classification mode. Though our work only speaks to the integration of comparison and classification, we expect similar effects to occur when other resource-demanding tasks are nested within a classification learning task. For a cleaner assessment of task-based manipulations in category learning, we recommend using the observational mode. Further, given the prevalence of relational concepts in education and the importance of comparison-based pedagogical techniques (Goldwater & Schalk, 2016), this work suggests that – irrespective of the type of comparison – educators should isolate comparison-based learning activities from other potentially competing learning tasks. More generally, by identifying a clear path for making comparison and category learning function together to promote concept formation, this broadens the applied potential of each of the techniques and suggests an instructional approach that leverages the power of both to improve upon existing methods. The present findings suggest the possibility that other instructional techniques that rely on comparison operating in conjunction with another task may show better outcomes if the task that embeds comparison is conducted in a passive mode. In terms of even broader consequences for the learning sciences, there is widespread debate on the relative merits of providing good answers to learn from versus withholding answers to allow for active or discovery-based learning; in all such cases there is the possibility that the success of an important, active learning mechanism (alignment or otherwise) depends on the learner having

less to do at the same time – an idea supported by the benefits of ‘passive’, worked-examples relative to active, problem-solving-based learning, especially early in learning (Atkinson & Renkl, 2007; Renkl, Atkinson, Maier, & Staley, 2002; Sweller & Cooper, 1985) .

In sum, side-by-side comparison is known to promote relational learning, however the conditions of learning must support an engaged effort at comparison. We believe that the traditional classification learning paradigm does not provide adequate supports; however, removing the immediate performance emphasis inherent in the guess-and-correct cycle opens the door to comparison-based gains. These results are revealing about the nature of core processes in higher-order cognition and how they interact. In translational terms, our findings point to a promising, but simple, recipe for promoting relational concept acquisition: inductive category learning across pairs of labeled examples.

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.
- Arnold, B. F., Hogan, D. R., Colford, J. M., & Hubbard, A. E. (2011). Simulation methods to estimate design power: an overview for applied research. *BMC Medical Research Methodology*, 11(1), 94.
- Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, 30(5), 666-677.
- Asmuth, J., & Gentner, D. (2017). Relational categories are more mutable than entity categories. *Quarterly Journal of Experimental Psychology*, 70(10), 2007-2025.
- Atkinson, R. K., & Renkl, A. (2007). Interactive example-based learning environments: Using interactive elements to encourage effective processing of worked examples. *Educational Psychology Review*, 19(3), 375-386.
- Bates, D., Maechler, M., Bolker, B., Walker, S., (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

- Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, 42(3), 481-495.
- Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, 22(1), 281-288.
- Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(6), 1147-1156.
- Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, 11(3), 356-373.
- Christie, S., & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive Science*, 38(2), 383-397.
- Corral, D., & Jones, M. (2014). The effects of relational structure on analogical learning. *Cognition*, 132(3), 280-300.
- Corral, D., Kurtz, K. J., & Jones, M. (2018). Learning relational concepts from within- vs. between-category comparisons. *Journal of Experimental Psychology: General*, 147(11), 1571-1596.
- Crossley, M. J., Roeder, J. L., Helie, S., & Ashby, F. G. (2018). Trial-by-trial switching between procedural and declarative categorization systems. *Psychological Research*, 82(2), 371-384.

- Davidson, N. S., & Gelman, S. A. (1990). Inductions from novel categories: The role of language and conceptual structure. *Cognitive Development*, 5(2), 151-176.
- Doumas, L. A., & Hummel, J. E. (2013). Comparison and mapping facilitate relation discovery and predication. *PloS One*, 8(6), e63889.
- Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. *The Quarterly Journal of Experimental Psychology*, 68(6), 1203-1222.
- Erickson, M. A. (2008). Executive attention and task switching in category learning: Evidence for stimulus-dependent representation. *Memory & Cognition*, 36(4), 749-761.
- Erickson, J. E., Chin-Parker, S., & Ross, B. H. (2005). Inference and classification learning of abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1), 86.
- Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.
- Gentner, D. (1981). Some interesting differences between verbs and nouns. *Cognition and Brain Theory*, 4, 161-178.
- Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj (Ed.), *Language development: Vol. 2. language, thought and culture* (pp. 301-334). Hillsdale, NJ: Erlbaum.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155-170.

- Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199-241). London: Cambridge University Press. (Reprinted in *Knowledge acquisition and learning*, 1993, 673-694).
- Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp.195-235). Cambridge, MA: MIT Press.
- Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, 34 (5). 752-775.
- Gentner, D. (2016). Language as cognitive toolkit: How language supports relational thought. *American Psychologist*. 71(8):650-657.
- Gentner, D., Anggoro, F. K., & Klibanoff, R. S. (2011). Structure mapping and relational language support children's learning of relational categories. *Child Development*, 82(4), 1173-1188.
- 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* (pp. 151–175). Washington, DC: American Psychological Association.
- Gentner, D., Levine, S. C., Ping, R., Isaia, A., Dhillon, S., Bradley, C., & Honke, G. (2016). Rapid Learning in a Children's Museum via Analogical Comparison. *Cognitive Science*, 40(1), 224-240.

- Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52(1), 45-56.
- Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, 65(2), 263-297.
- 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.
- Gick, M. L., & Paterson, K. (1992). Do contrasting examples facilitate schema acquisition and analogical transfer?. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 46(4), 539-550.
- Goldwater, M. B., Don, H. J., Krusche, M. J., & Livesey, E. J. (2018). Relational discovery in category learning. *Journal of Experimental Psychology: General*, 147(1), 1.
- Goldwater, M. B., & Markman, A. B. (2011). Categorizing entities by common role. *Psychonomic Bulletin & Review*, 18(2), 406-413.
- Goldwater, M. B., Markman, A. B., & Stilwell, C. H. (2011). The empirical case for role-governed categories. *Cognition*, 118(3), 359–376.
- Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, 142(7), 729-757.

- Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4), 493-498.
- Higgins, E. J. (2017). The Complexities of Learning Categories Through Comparisons. In *Psychology of learning and motivation* (Vol. 66, pp. 43-77). Academic Press.
- Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language*, 66(4), 731-746.
- Kotovskiy, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, 67(6), 2797-2822.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1), 22-44.
- Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, 14, 560-576.
- Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. In *Psychology of learning and motivation* (Vol. 63, pp. 77-114). Academic Press.
- Kurtz, K. J., & Boukrina, O. (2004). Learning relational categories by comparison of paired examples. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*.

- 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. (2013). Detecting anomalous features in complex stimuli: the role of structured comparison. *Journal of Experimental Psychology: Applied*, 19(3), 219-232.
- Kurtz, K. J., & Gentner, D. (1998). Category learning and comparison in the evolution of similarity structure. In *Proceedings of the Twentieth Annual Conference of the 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). Package 'lmerTest'. *R Package Version*, 2(0).
- Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, 43(2), 266-282.
- Loewenstein, J. (2010). How one's hook is baited matters for catching an analogy. In *Psychology of learning and motivation* (Vol. 53, pp. 149-182). Academic Press.
- Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review*, 6(4), 586-597.
- Loewenstein, J., Thompson, L., & Gentner, D. (2003). Analogical learning in negotiation teams: Comparing cases promotes learning and transfer. *Academy of Management Learning &*

Education, 2(2), 119-127.

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

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

Markman, A. B., & Gentner, D. (1993b). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, 32(4), 517-535.

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(4), 329-358.

Ming, N. C. (2009). Analogies vs. contrasts: A comparison of their learning benefits. In B. Kokinov, D. Gentner, & K. Holyoak (Eds.), *New frontiers in analogy research: Proceedings of the second international conference on analogy* (pp. 338-347). Sofia, Bulgaria: New Bulgarian University.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16(5), 519-533.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(3), 289-316.

- Namy, L. L., & Clepper, L. E. (2010). The differing roles of comparison and contrast in children's categorization. *Journal of Experimental Child Psychology*, 107(3), 291-305.
- Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General*, 131(1), 5-15.
- Patterson, J. D., & Kurtz, K. J. (2014). *Engaging the Comparison Engine: Implications for Relational Category Learning and Transfer*. Poster session presented at the Annual Conference of the Cognitive Science Society, Pasadena, CA.
- Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition*, 121(1), 83-100.
- R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(5), 1261-1275.
- Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. *Journal of Experimental Education*, 70, 293-315.
- Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, 99(3), 561-574.

- 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(3), 529-544.
- Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive Science*, 36(5), 919-932.
- Smith, L. A., & Gentner, D. (2014). The role of difference-detection in learning contrastive categories. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the Thirty-sixth Annual Conference of the Cognitive Science Society* (pp. 1473-1478). Austin, TX: Cognitive Science Society.
- 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.
- Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2(1), 59-89.
- Thai, K. P., Krasne, S., & Kellman, P. J. (2015). Adaptive Perceptual Learning in Electrocardiography: The Synergy of Passive and Active Classification. In *Proceedings of the Thirty-seventh Annual Conference of the Cognitive Science Society*.
- Tohill, J. M., & Holyoak, K. J. (2000). The impact of anxiety on analogical reasoning. *Thinking & Reasoning*, 6(1), 27-40.

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory?. *Psychonomic Bulletin & Review*, 19(5), 899-905.

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.

Tables and Figures

Table 1

Adjusted Means and Standard Errors

	Same-Category Comparison		Single-item		Mixed Pairs Comparison	
	Classification	Observational	Classification	Observational	Classification	Observational
Old test	.78(.04)	.92(.02)	.85(.03)	.84(.03)	.82(.03)	.89(.02)
New test	.72(.05)	.86(.03)	.77(.04)	.74(.04)	.78(.04)	.80(.04)
Transfer	.70(.04)	.85(.03)	.74(.04)	.74(.04)	.73(.04)	.78(.03)

Table 2

Adjusted Means and Standard Errors

	Same-Category Comparison		Different-Category Comparison	
	Classification	Observational	Classification	Observational
Old test	.83(.02)	.89(.02)	.82(.03)	.90(.02)
New test	.75(.04)	.84(.03)	.76(.04)	.85(.03)
Transfer	.76(.03)	.85(.02)	.72(.03)	.82(.02)

Table 3

Adjusted Means and Standard Errors

	Observational	Obs Delayed Labels	Covert Classification	Classification
Old test	.87(.03)	.82(.03)	.84(.03)	.78(.04)
New test	.81(.04)	.78(.04)	.79(.04)	.75(.04)
Transfer	.82(.03)	.73(.04)	.80(.03)	.70(.04)

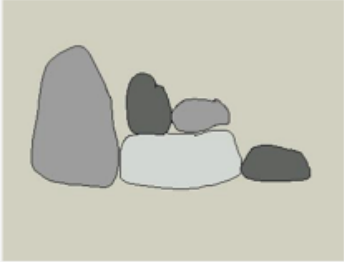
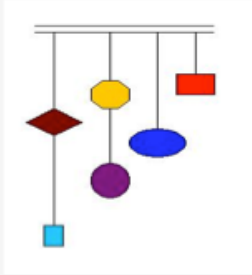
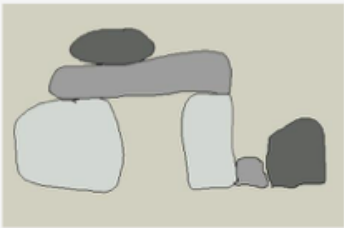
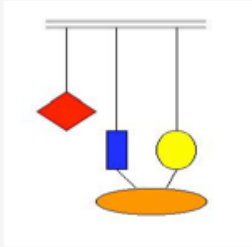
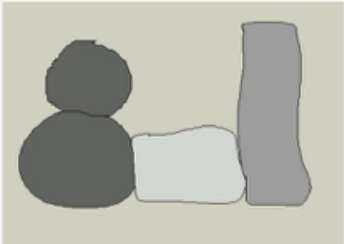
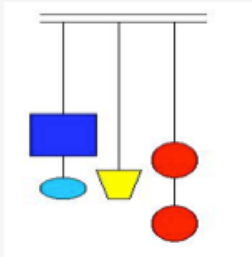
Category	Training/Test	Far Transfer
Monotonic decrease in height from left to right		
One supported by/bridging two		
Similar size, shape, and color in stack of two		

Figure 1. Example stimuli pertaining to all three experiments. Each row represents a different category. Archaeological stimuli in the center column were used for training and the within-domain test. Mobile stimuli in the right column were used as far transfer stimuli.

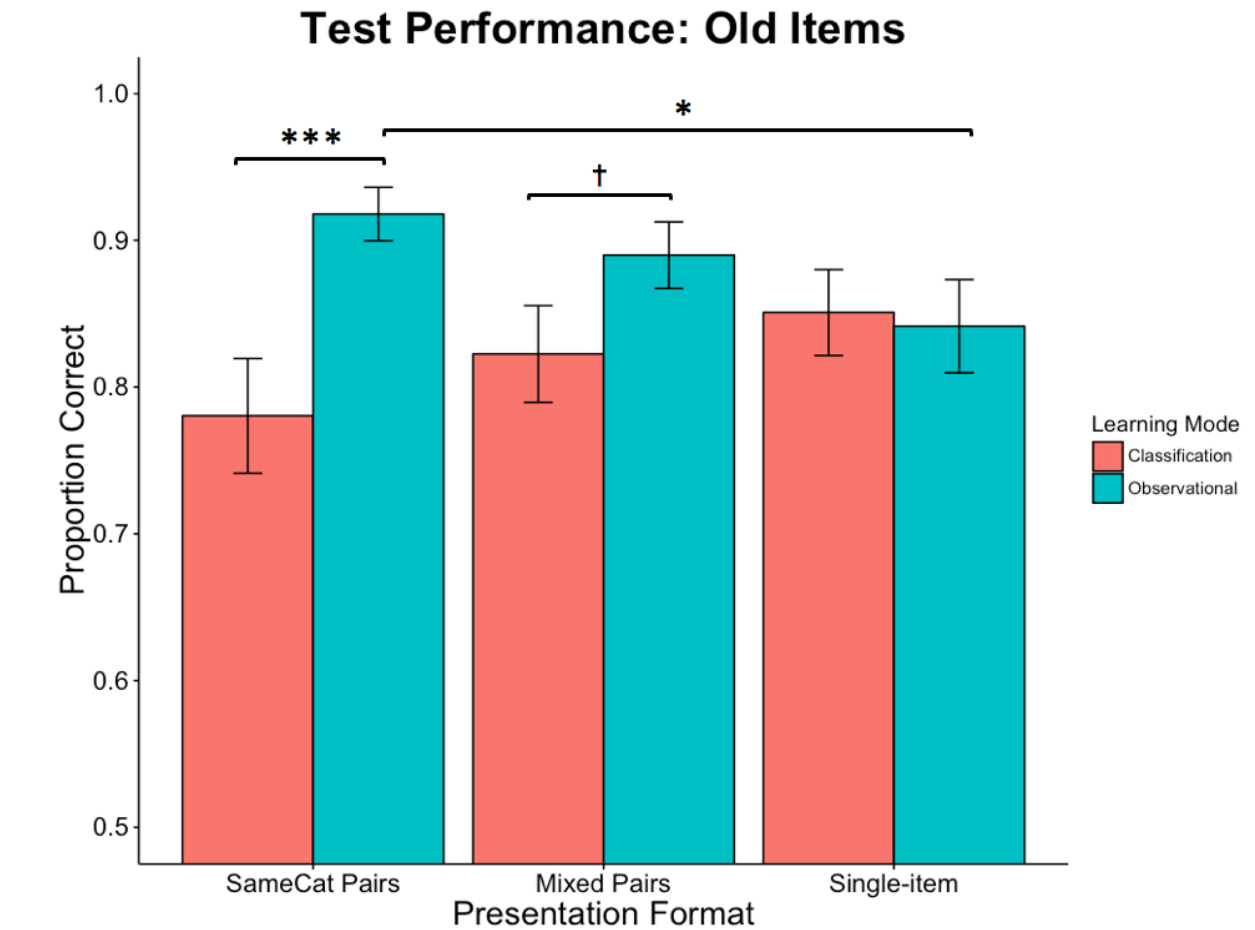


Figure 2. Mean old item test performance by condition – Experiment 1. Values represent adjusted means. Error bars represent +/- 1 SE.

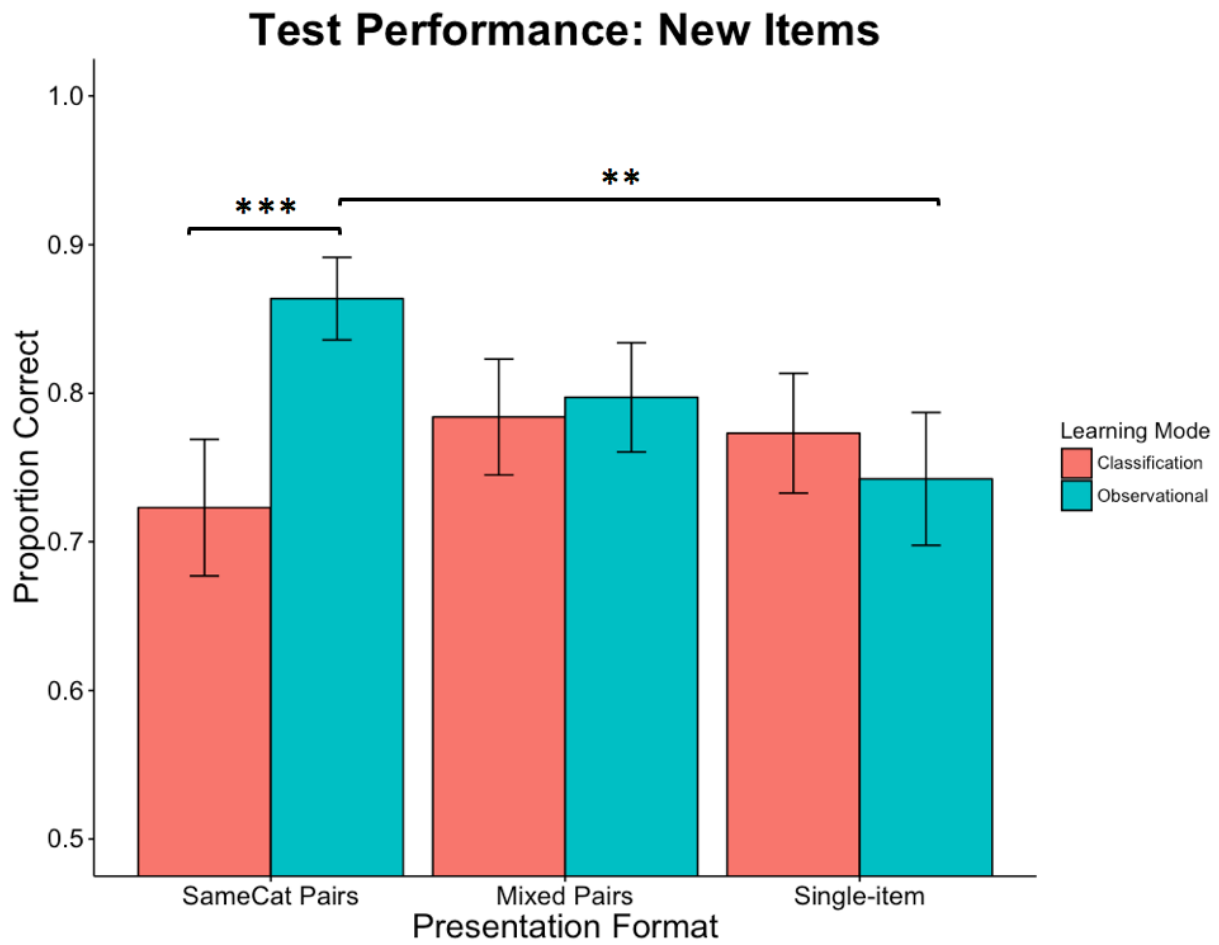


Figure 3. Mean new item test performance by condition – Experiment 1. Values represent adjusted means. Error bars represent ± 1 SE.

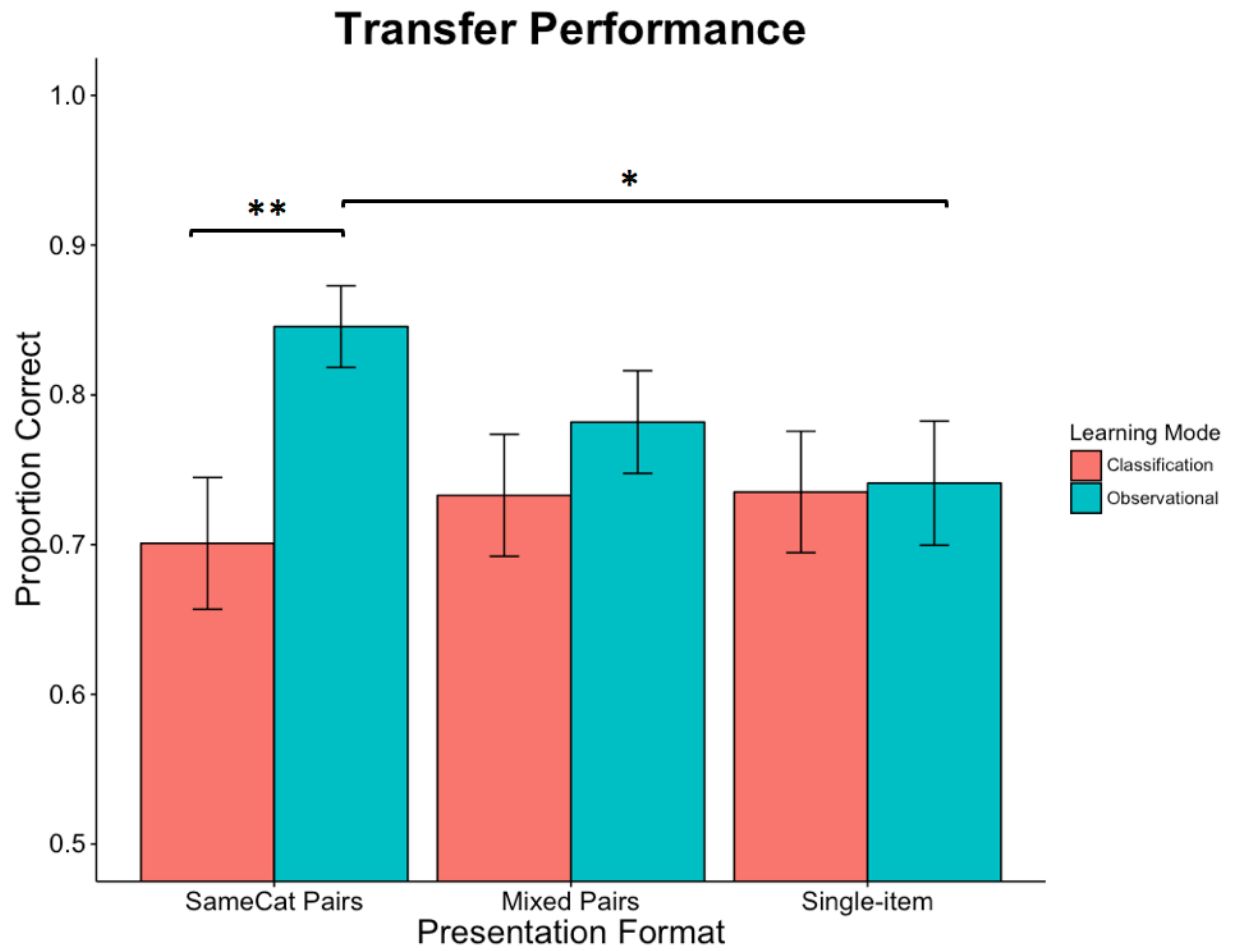


Figure 4. Mean transfer performance by condition – Experiment 1. Values represent adjusted means. Error bars represent +/- 1 SE.

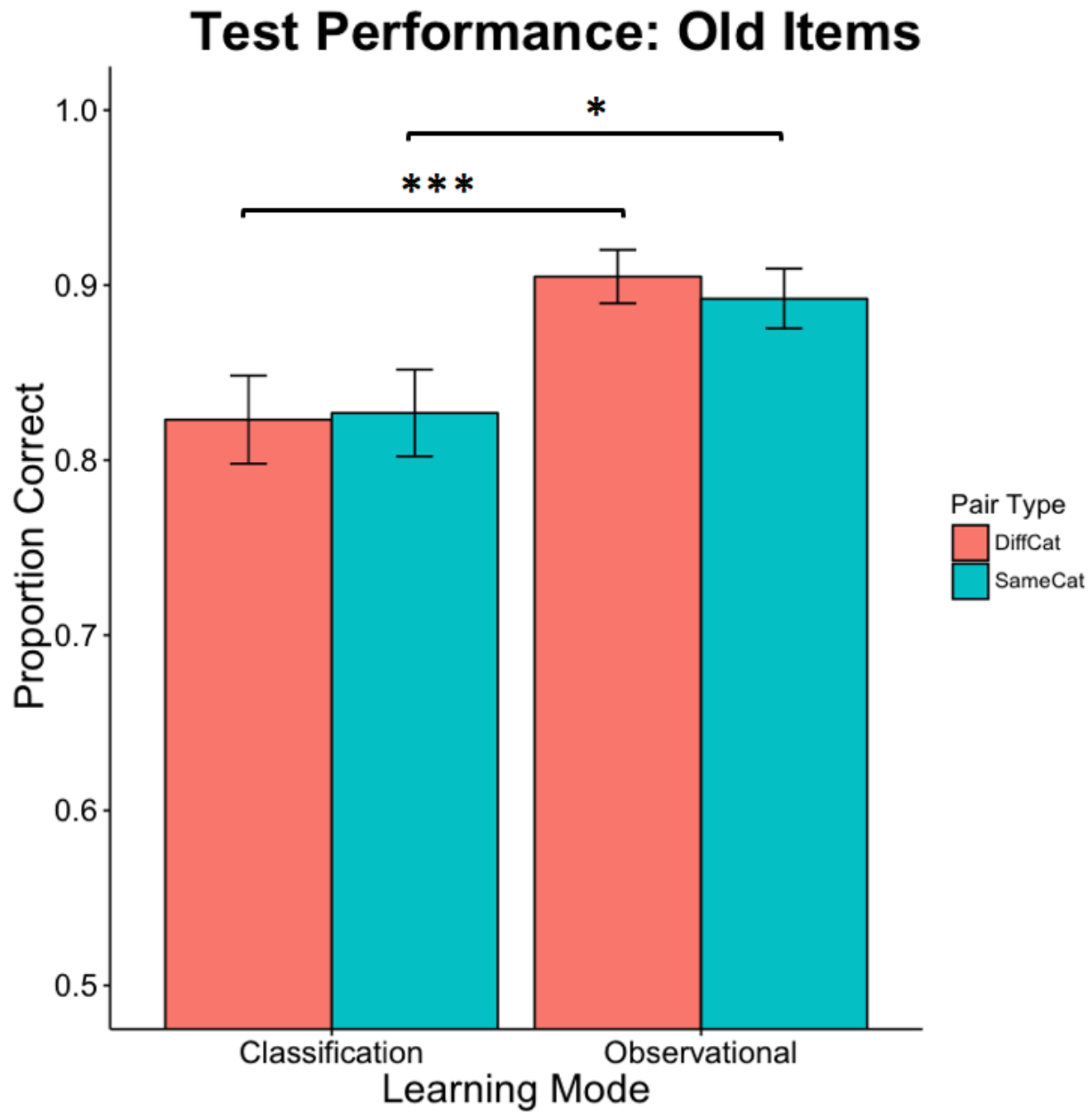


Figure 5. Mean old item test performance by condition – Experiment 2. DiffCat and SameCat refer to different- and same-category pairs respectively. Values represent adjusted means. Error bars represent ± 1 SE.

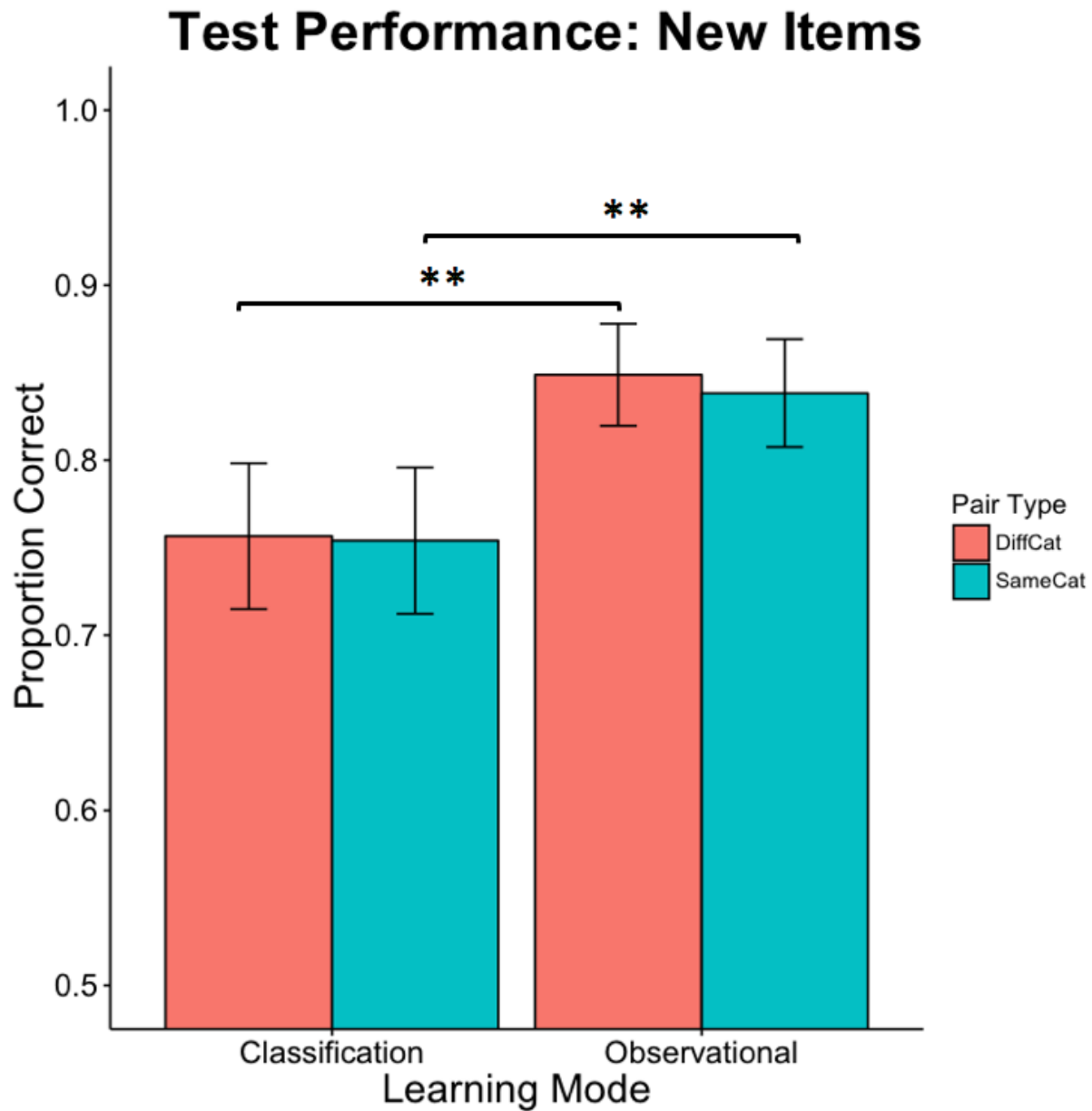


Figure 6. Mean new item test performance by condition – Experiment 2. DiffCat and SameCat refer to different- and same-category pairs respectively. Values represent adjusted means. Error bars represent ± 1 SE.

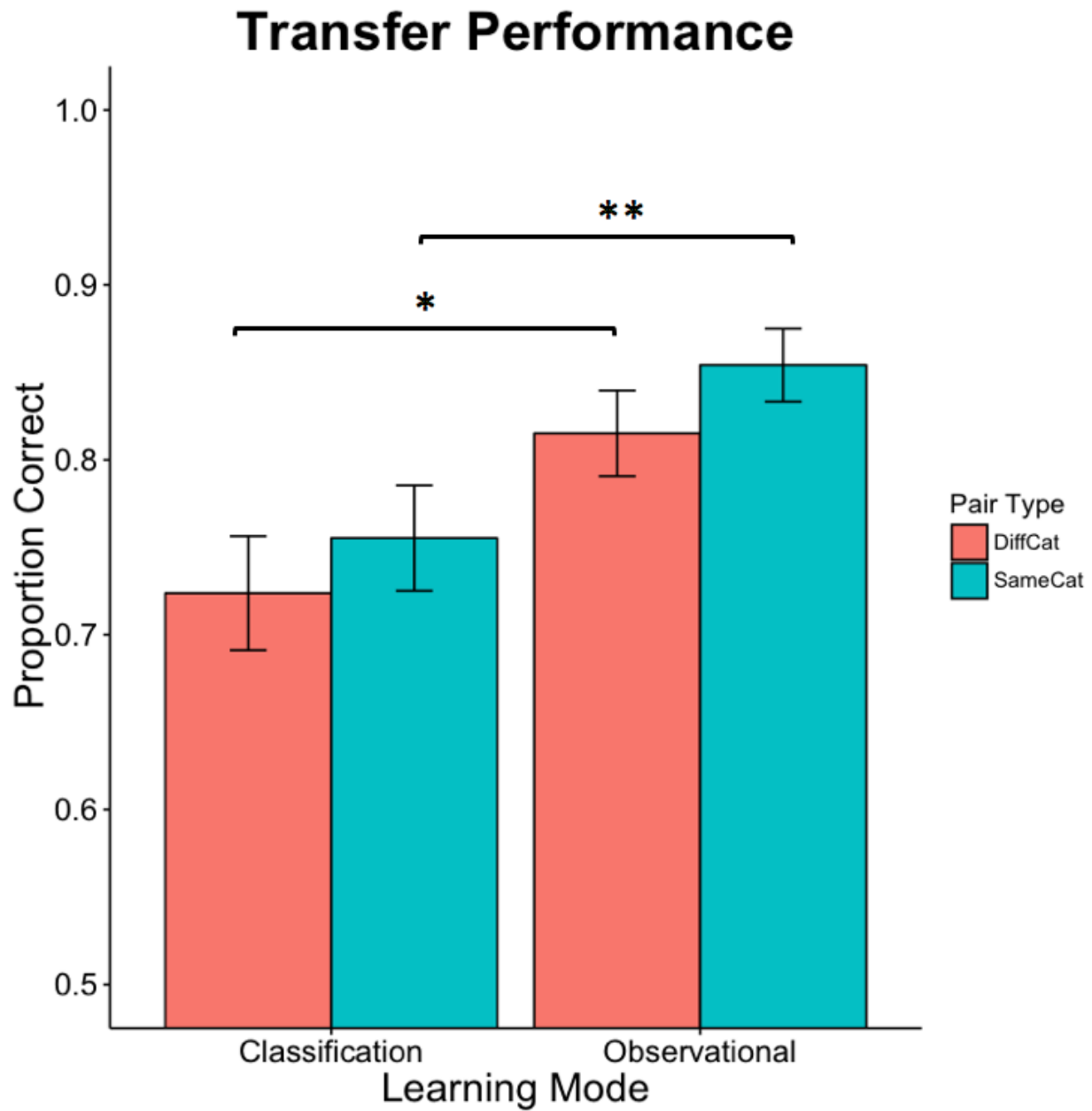


Figure 7. Mean transfer performance by condition – Experiment 2. DiffCat and SameCat refer to different- and same-category pairs respectively. Values represent adjusted means. Error bars represent ± 1 SE.

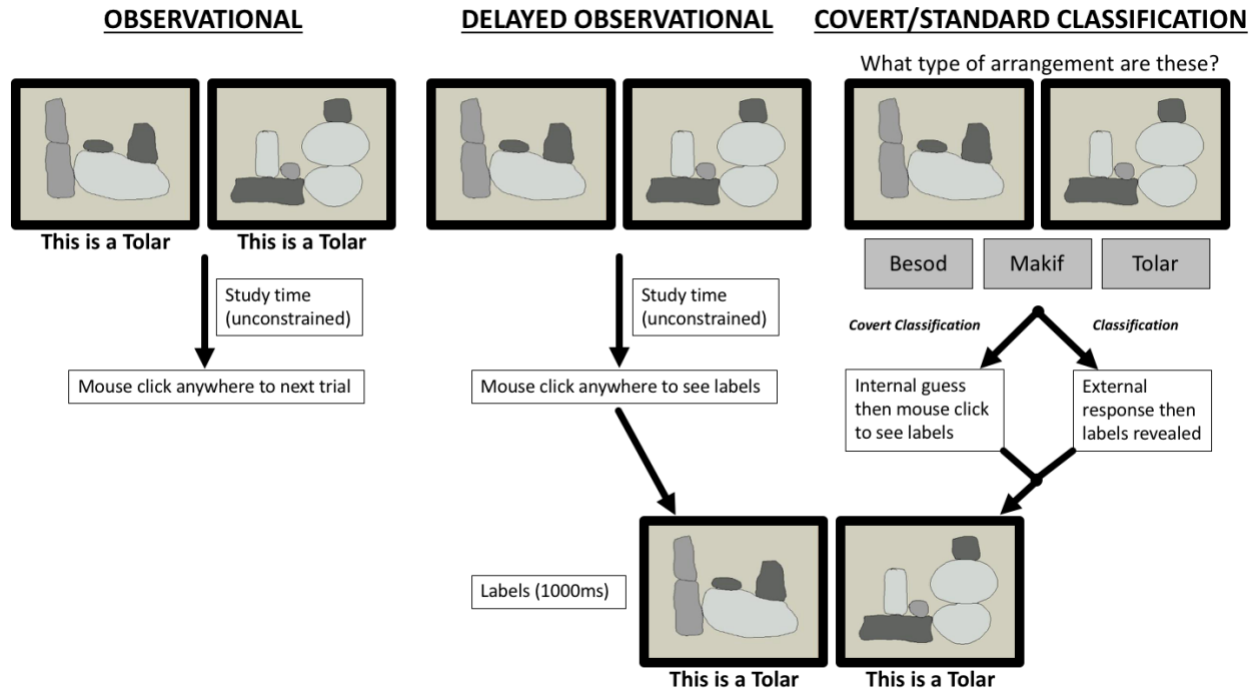


Figure 8. Procedure by condition in Experiment 3.

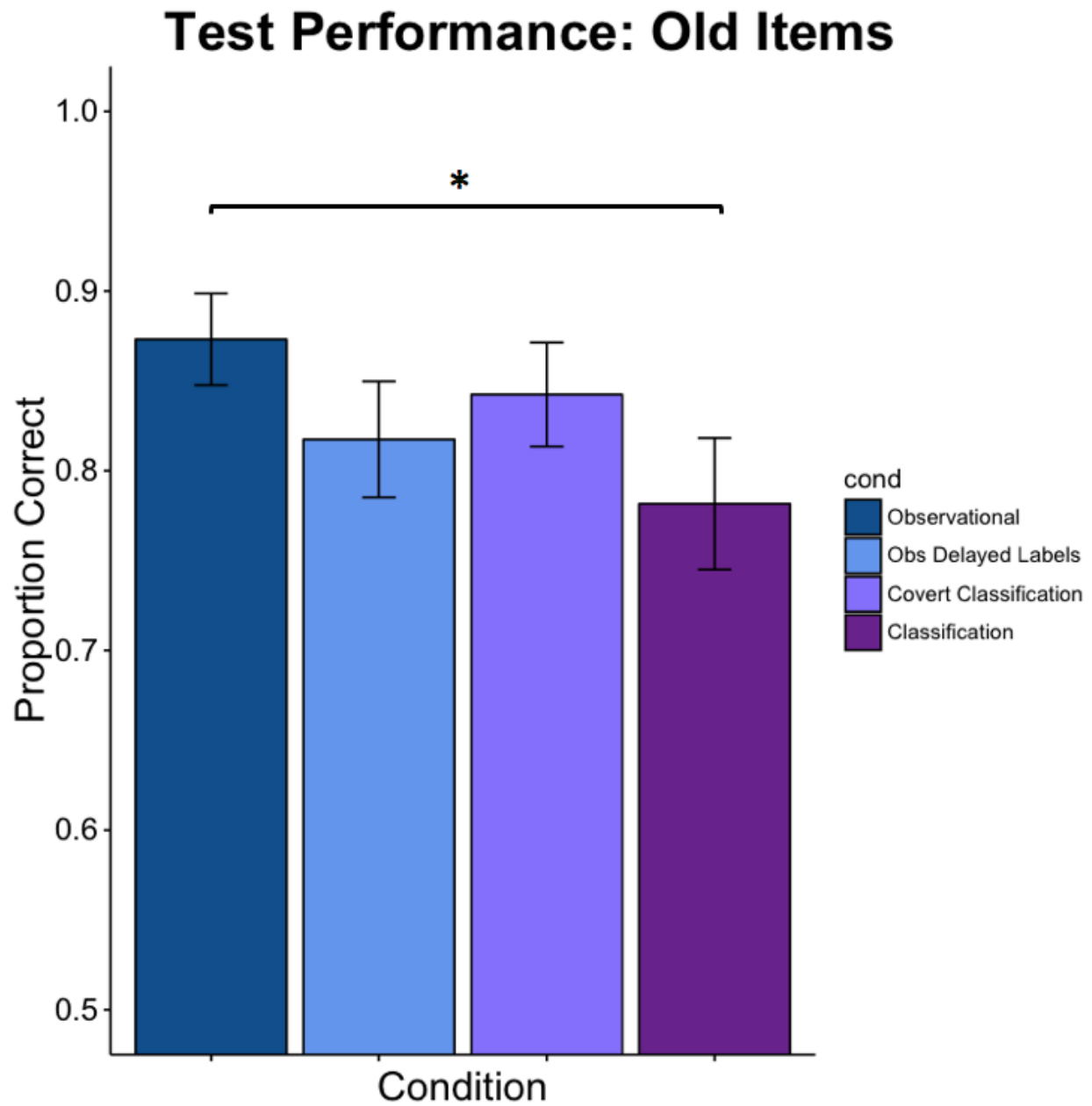


Figure 9. Mean old item test performance by condition – Experiment 3. Values represent adjusted means. Error bars represent ± 1 SE.

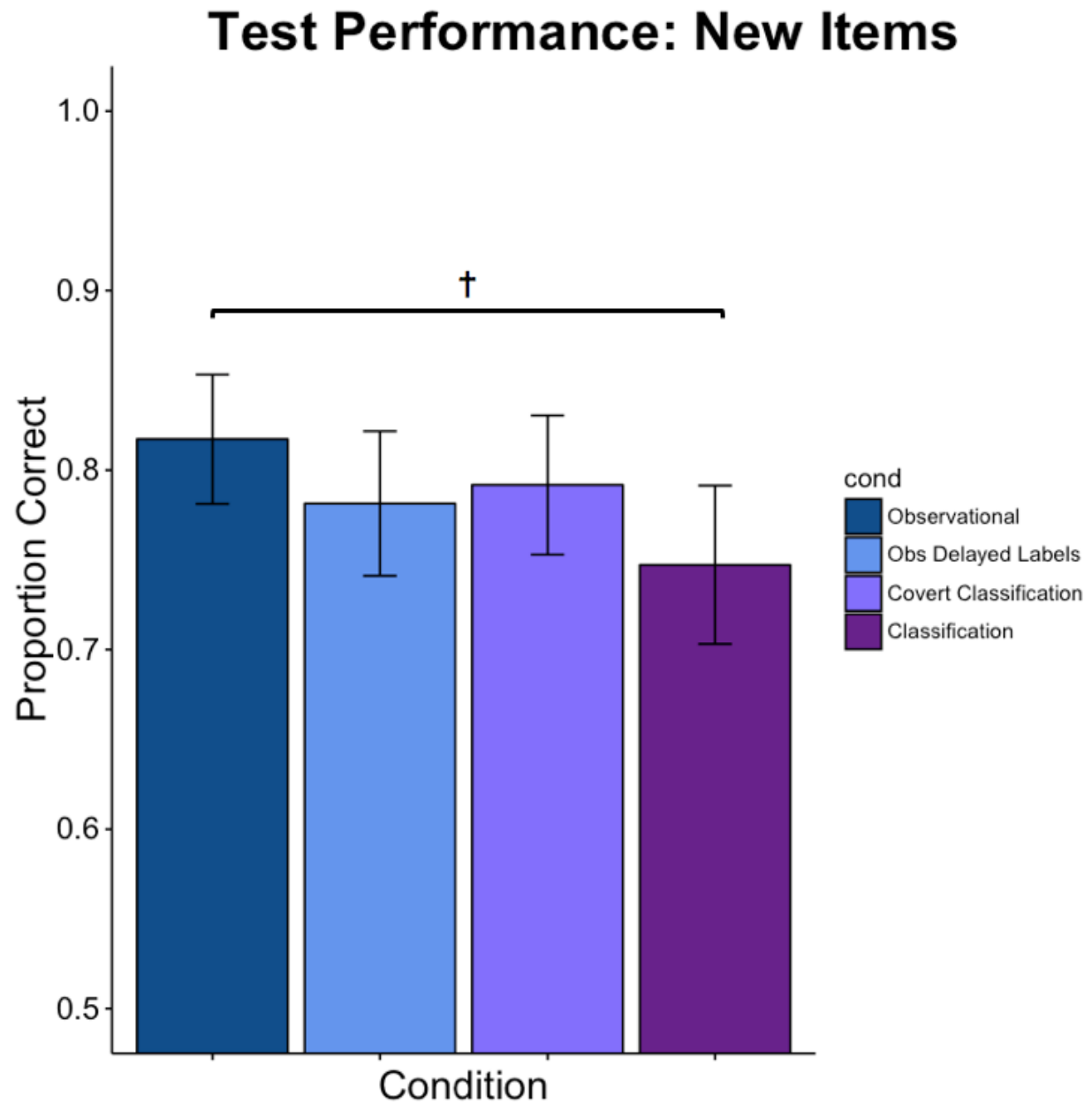


Figure 10. Mean new item test performance by condition – Experiment 3. Values represent adjusted means. Error bars represent ± 1 SE.

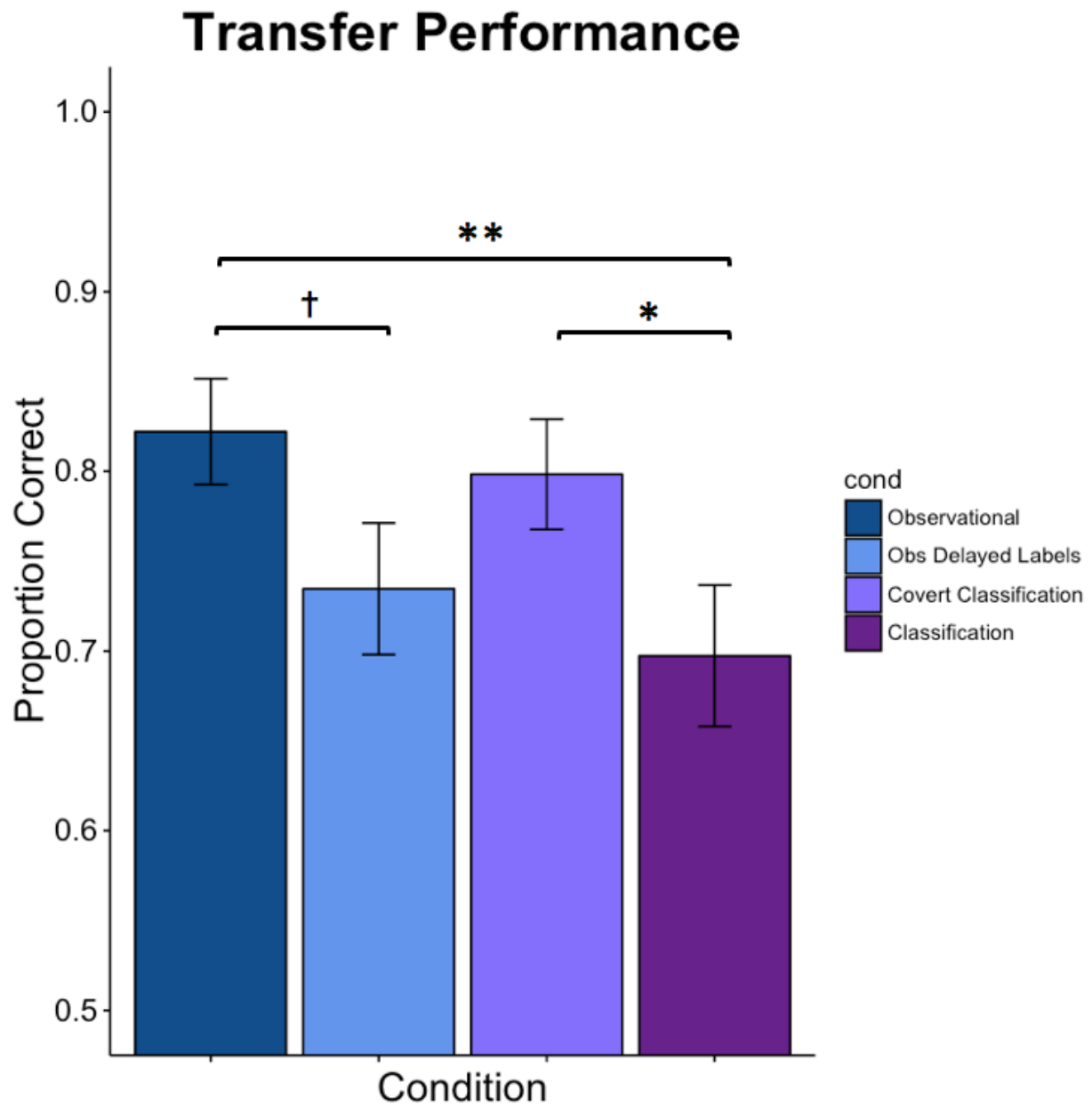


Figure 11. Mean transfer performance by condition – Experiment 3. Values represent adjusted means. Error bars represent ± 1 SE.

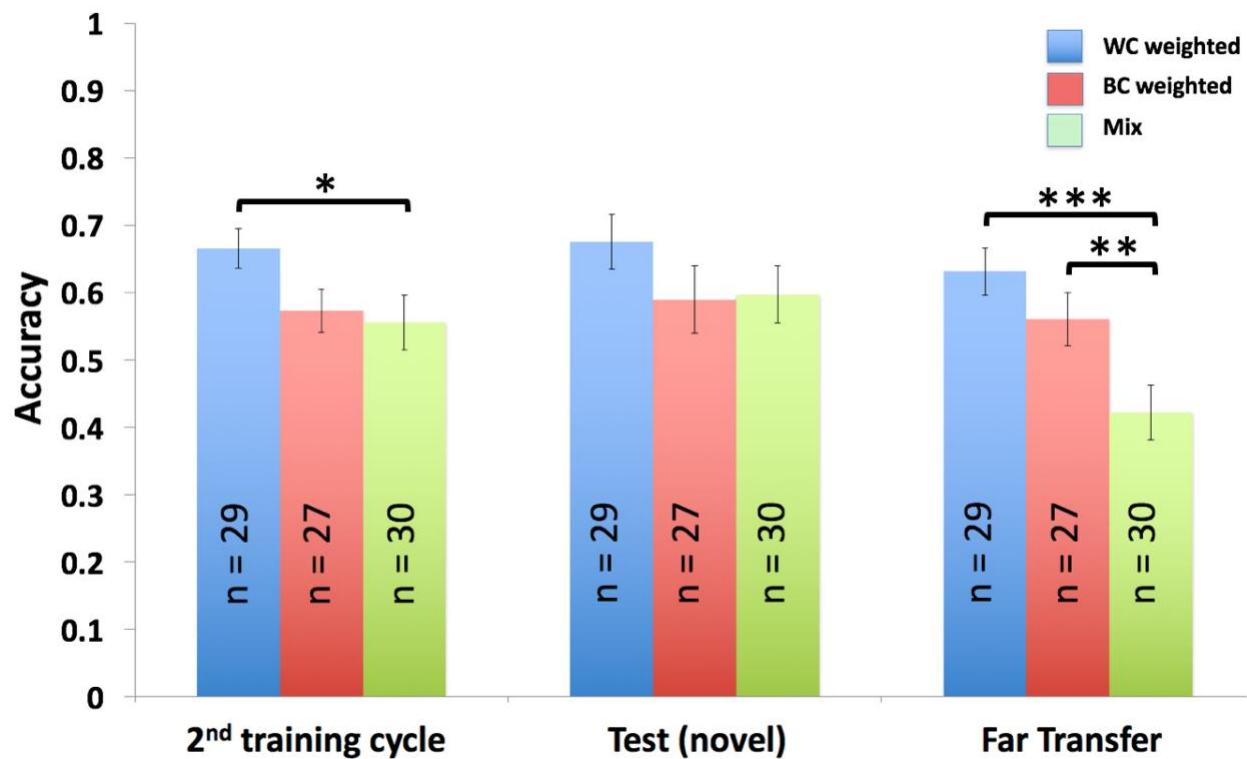


Figure 12. Mean accuracy by condition from Patterson & Kurtz (2014), presented as a poster at the thirty-sixth annual meeting of the Cognitive Science Society. The within-category (WC) weighted condition consisted of a 75/25 ratio of same-category pairs to different-category pairs. The between-category (BC) weighted condition was the same, but with the ratios for same-category and different-category pairs reversed. The Mix condition consisted of a 50/50 mixture of same and different-category pairs, much like the mixed pairs condition in the present investigation. Values represent raw means. Error bars represent +/- 1 SE.