Journal of Experimental Psychology: General
2018, Vol. 147, No. 1, 1–35

0096-3445/18/$12.00 http://dx.doi.org/10.1037/xge0000387

Relational Discovery in Category Learning

Micah B. Goldwater and Hilary J. Don
The University of Sydney

Moritz J. F. Krusche
Maastricht University and University College-London

Evan J. Livesey
The University of Sydney

Learning categories defined by the relations among objects supports the transfer of knowledge from initial learning contexts to novel contexts that share few surface similarities. Often relational categories have correlated (but nonessential) surface features, which can be a distraction from discovering the category-defining relations, preventing knowledge transfer. This is one explanation for "the inert knowledge problem" in education wherein many students fail to spontaneously apply their learning outside the classroom. Here we present a series of experiments using artificial categories that correlate surface features and relational patterns during learning. Our goal was to determine what task parameters and individual differences in learners shift focus to the relational aspect of the category and foster transfer to novel disparate exemplars. We consistently showed that the effectiveness of task structure manipulations (e.g., the sequence of learning exemplars) depended on the learners' strategies (e.g., whether learners are oriented toward discovering rules or focusing on exemplars). Further, we found support that "inference-learning," wherein learners are presented with incomplete exemplars and learn how to infer the missing pieces, is an effective way to promote relational discovery and transfer, even for learners who are not predisposed to make such discoveries.

Keywords: categories and concepts, individual differences, inert knowledge, knowledge transfer, relational categories

Recognizing common patterns in the relations among objects is critical for higher-level cognition, underlying much of causal, logical, and analogical thinking (Gentner, 1983; Halford, Wilson, & Phillips, 2010). Some have even argued that this capacity for relational cognition is what separates humans from our nearest primate cousins, as, for example, infants as young as 9 months of age can readily abstract simple relations across diverse sets of objects that other great apes require thousands of trials to learn (e.g., Ferry, Hespos, & Gentner, 2015; Gentner, 2003; Penn, Holyoak, & Povinelli, 2008). Supporting a mature capacity for relational cognition is our long-term knowledge store of *relational categories*— categories defined by common relational patterns. For instance, *barrier* classifies any object or situation that stands in the way of an agent achieving its goal, such as a giant boulder or childhood poverty (Gentner, & Kurtz, 2005; see also, Markman & Stilwell, 2001; Goldwater, Markman, & Stilwell, 2011). Because relational categories are defined by relational patterns independent of the intrinsic features of objects, they powerfully promote "far transfer," which is the application of prior knowledge to new contexts that share few surface similarities with the initial context of learning (Goldwater & Schalk, 2016). Indeed, novel extensions of relational concepts to disparate domains have been behind many important historical insights, for example the development of Kepler's Laws of Planetary Motion, or how Claude Shannon implemented the structure of Boolean logic in electric circuits (Gentner et al., 1997; Markman & Wood, 2009).

This kind of far knowledge transfer is common in the use of "everyday" relational categories, such as understanding the commonalities between the two exemplars of *barrier*. However, seeing the underlying connection between, and thus solutions to, disparate problems is not always easy. Many examples of the difficulty that this form of far transfer poses for students can be found in educational settings, especially in comparing complex natural phenomena, such as the common positive feedback mechanisms in global warming and economic pricing-bubbles. This ability to see relational similarities across seemingly very different problems is a key difference between novices and experts (Chi, Feltovich, & Glaser, 1981; Rottman, Gentner, & Goldwater, 2012; Stains & Talanquer, 2008).

Synthesizing this work on category representation, expertise, and education, Goldwater and Schalk (2016) argue that (a) The

Micah B. Goldwater and Hilary J. Don, School of Psychology, The University of Sydney; Moritz J. F. Krusche, Faculty of Psychology and Neuroscience, Maastricht University, and Department of Experimental Psychology, University College-London; Evan J. Livesey, School of Psychology, The University of Sydney.

Moritz J. F. Krusche is now at the Warwick Business School, University of Warwick.

This research was funded by Australian Research Council Grant DP150104267 awarded to Micah B. Goldwater and Evan J. Livesey Some of these data were presented at the 2016 International Meeting of the Psychonomic Society, the 2016 Annual Meeting of the Cognitive Science Society, and the 2016 Australasian Experimental Psychology Conference.

Correspondence concerning this article should be addressed to Micah B. Goldwater, School of Psychology, The University of Sydney, Brennan MacCallum Building A18, Sydney, NSW 2006 Australia. E-mail: micah.goldwater@sydney.edu.au

majority of concepts in formal education can be considered relational categories (e.g., *cue competition* in psychology classifies sequences of learning events based on common predictive relationships among simultaneously presented event elements, not the identity of the elements themselves, *catalyst* in chemistry classifies molecules by their role in chemical processes, not by their intrinsic chemical composition, or the *ur-myth* in literary criticism which classifies a narrative by its thematic structure and historical precedent, not its specific content); and (b) The "inert knowledge problem" (Whitehead, 1929), wherein students appear to understand their instruction within the learning context but do not see how that learning applies to new problems and situations, can be conceived of as a failure of relational categorization. That is, the lack of transfer is not just the failure to access the most relevant memory or prior example, as it is more typically conceived (see Trench & Minervino, 2017, for review). Rather, it often involves the failure to classify a novel situation as relevant to a previously instructed relational category. For example, Schwartz and colleagues' work on physics education (e.g., Schwartz, Chase, Oppezzo, & Chin, 2011) showed that when students engage in typical instruction informing them how to solve a class of problems, and then repeatedly practice on a series of examples, they would most likely fail to spontaneously transfer these procedures to novel problems. However, when the students compared and contrasted the exemplar problems to generate their own more general description of how the category of problems needed to be solved before receiving direct instruction, transfer was achieved at a higher rate.

Despite the importance of relational categories to higher-level cognition, and applied issues in education, the majority of research in the category learning literature has focused on categories concerning the commonalities in features intrinsic to individual objects, such as *dog* and *guitar* (see Murphy, 2004 for extensive review). While learning categories of any kind will involve some common cognitive mechanisms, there is ample evidence that relational categories are formally distinct from those that are feature-based, requiring a distinct representational format to capture their meaning, and distinct processes to acquire and reason with them (see Markman, 1999 for thorough discussion; see also Hummel & Holyoak, 2003; Doumas, Hummel, & Sandhofer, 2008, among others). The distinction is especially pertinent to educational settings because, after all, the goal of a modern education is not to focus on objects in isolation or to memorize lists of facts, but to foster the ability to conceptualize the natural and social world in terms of abstract yet coherent schemas that foster flexible reasoning inside and outside the classroom (Resnick, 2010).

Building on recent related efforts (e.g., Corral & Jones, 2014; Doumas & Hummel, 2013; Kurtz, Boukrina, & Gentner, 2013; Tomlinson & Love, 2010) the present series of experiments had two primary aims related to relational discovery in category learning. The first was to further examine the factors that lead to the discovery and transfer of relational knowledge. The second was to serve as a novel experimental model for the abstraction and transfer of relational knowledge in education in an attempt to recreate the inert knowledge problem in the laboratory. That is, previous laboratory research on inert knowledge has focused on the failures of learners to retrieve and apply an individual or small number of relevant prior examples. We hope that by reframing the problem as one rooted in a failure of categorization, we can bring another suite of experimental methods to bear (and see Little & McDaniel, 2015a for a similar perspective).

Feature-Based Categories Versus Relational Categories

The present studies contrast relational discovery with a focus on learning more surface features of the same stimuli, and it is the unique properties of relational categories in particular that are important in this endeavor. Cognitive processes that naturally lend themselves to relational learning (e.g., stimulus comparison, hypothesis testing and rule learning) may well be involved to some degree in the learning of simple categories defined by concrete stimulus features (e.g., see, Livesey & McLaren, 2009; Little & McDaniel, 2015b; Natal, McLaren, & Livesey, 2013 for direct attempts to make these connections, but also Ashby & Maddox, 2005; Goodman, Tenenbaum, Griffiths, & Feldman, 2008; Nosofsky, Palmeri, & McKinley, 1994 among many others for rule-driven hypothesis testing models of category learning). However, there is ample evidence that the learning and representation of relational *categories* (i.e., those defined by abstract relations among stimuli without reference to a particular set of stimulus features) possess characteristics that are qualitatively distinct from the learning and representation of feature-defined categories.

The literature now has many examples of how feature-based and relational categories differ, both examining familiar categories, and the process of learning novel categories. For example, when considering familiar relational categories, (e.g., *friend, drug, home*), people are more likely to list properties extrinsic to individual category members (e.g., *always there for you, changes your mental state, a place you live in,* respectively), while feature-based categories (e.g., *TV, knife, celery*) elicit more properties intrinsic to category members (e.g., *has a screen, sharp, green,* respectively Goldwater et al., 2011). Members of relational categories (two examples of *barrier* are *boulder & poverty*) are rated as less similar to each other than are members of feature-based categories (two examples of *vegetable* are *celery* and *carrot*; Gentner, & Kurtz, 2005); idealized exemplars are more prominent for relational categories, while average exemplars are more prominent for feature-based categories (e.g., the ideal friend seems more important in understanding the concept of *friend* than the statistically average friend, while the statistically average TV is more important in understanding the concept of *TV* than the ideal TV; Goldwater et al., 2011; Kittur, Hummel, & Holyoak, 2004; Rein, Goldwater, & Markman, 2010); and relational categories support inductions about situations (e.g., what is eaten at breakfast vs. dinner parties), rather than about individual category members (e.g., the nutritional value of starches vs. fruits; Ross & Murphy, 1999). In theoretically related work on analogical reasoning, shared surface features between exemplars or situations (e.g., two stories about birds) often drive memory retrieval, but commonalities in deeper relational structure drive reasoning about two domains once both are retrieved (e.g., a story about conflict resolution between birds, and another about conflict resolution between warring nations, see, e.g., Gentner, Rattermann, & Forbus, 1993; Ross, 1987 and Forbus, Gentner, & Law, 1995; Hummel & Holyoak, 1997 for the computational differences between memory retrieval and reasoning).

Relational and feature-based categories are also learned in distinct patterns. For example, words referring to relational categories

are almost universally learned later than words referring to feature-based categories (Gentner, 1982; Gentner & Boroditsky, 2001). Feature-based categories are readily learned with features that are merely probabilistically predictive of category membership, that is without any single feature or set of necessary and sufficient features common to all category members (Rosch & Mervis, 1975). In direct contrast, there is evidence that relational categories that do not contain any category-defining relations cannot be learned directly (i.e., they require restructuring the category space to create defining relations Jung & Hummel, 2015). Last, Davis, Goldwater, and Giron (2017) show that learning feature-based and relational categories differ in their neuorological correlates. However, in contrast to these differences in learning processes, Little and McDaniel (2015b) found that the same explicit strategy can be beneficial for learning both kinds of categories (see further discussion below).

Although relational categories are psychologically distinct from feature-based categories, it is important to still recognize that relational category members often have typical features, and these typical features are important to both the representation of the category and how the category is learned. For example, predators tend to have eyes on the front of their heads, while prey animals tend to have eyes more laterally. Although eye position does not define an animal as predator or prey, understanding how eye-position connects to ecological relations can aid understanding these ecological relations. Here, we are building on the argument from Murphy and Medin (1985) that we use naïve theories of the causal relations among the observable features to explain and cohere category membership. For example, Medin, Wattenmaker, and Hampson (1987), Rehder and Ross (2001), and Spalding and Murphy (1996) discuss how cohering causal and thematic relations help support category construction and learning when features are not shared across exemplars. However, relational categories can also have correlated features that are more arbitrary to allow us to easily recognize category members when the category defining relation is not easily observable on its own, such as different uniform colors signifying different ranks in a hierarchy.

Similar to these relational categories, educators often design instruction wherein deep structure and surface content are correlated. For example consider mathematical word problems. The particular semantics of the story and the objects involved are not critically important; the goal is to be able to see past the content and recognize the applicable solution procedure. Bassok, Chase, and Martin (1998) showed that textbook writers however were quite constrained by correspondences between the semantics of the objects from the stories and the mathematical operations, such that they would write problems where apples would be divided between baskets, but you would never add up apples and baskets into some total number of objects. Mayer (1981) surveyed algebraic word problems from a large sample of textbooks and noted that certain kinds of topics (e.g., about the motion of objects vs. the total output of a production line) were associated with particular categories of solution procedures or "problem templates," even if there is no necessary link between them (see also work by Greeno and colleagues, e.g., Nesher, Greeno, & Riley, 1982).

On the one hand, these connections between more surface semantics and deeper problem structure may ease initial learning, but they can also obfuscate potential deeper connections (e.g., motion and output problems are both essentially about a multiplicative relationship between some rate and how much time something happened at that rate), and lead to misapplications of a solution procedure when learners have to solve an example problem where the correlation between surface content and solution procedure is broken (e.g., Blessing & Ross, 1996). How do students learn how to focus on the deep structure?

Relational Learning in Education

In both children's conceptual development and in formal education, there is a characteristic pattern of learning called the *relational shift* where learners often initially focus on these correlated surface features before discovering the common deep relational structure, (Chi et al., 1981; Doumas et al., 2008; Gentner & Rattermann, 1991; Keil & Batterman, 1984; Rottman et al., 2012; Stains & Talanquer, 2008, but see Bulloch & Opfer, 2009, for a different pattern). Further, recognizing the surface commonalities initially can play an important role in the discovery of the underlying relational structure later on (Braithwaite & Goldstone, 2015; Kotovsky & Gentner, 1996). However, unlike in typical child development learning everyday concepts, in education many learners never truly discover the abstractions, limiting the chances that learners will be able to apply their education outside the classroom (Whitehead, 1929). Unfortunately, this lack of transfer is perhaps the single most common finding in education research.

Although correlated features have an important role in relational category learning, recent empirical examinations have used categories with no diagnostic or even probabilistically predictive perceptual features (e.g., Corral & Jones, 2014; Doumas, & Hummel, 2013; Kurtz et al., 2013; Jung & Hummel, 2015, cf. Livins, Spivey, & Doumas, 2015; Tomlinson & Love, 2010) or have used sets of categories where surface features and relational structure are orthogonal allowing for the adoption of either a feature-based or relational categorization strategy, but not a strategy that uses both sources of information together (e.g., Goldwater & Gentner, 2015). Hence, we ask: when features are correlated with, but not essential to a relational category, how do learners discover the category-defining relations? We hope answering this question explains both basic processes of concept learning, and motivates targets for educational research.

The Current Experiments

Category Learning Task

In the series of experiments presented here, we designed a pair of contrasting artificial categories that differed both in (spatial) relational structure, and in more surface perceptual features. Using typical artificial category learning methods (e.g., stimuli are presented one at a time, subjects guess its category membership, and corrective feedback is presented), our aim was to see what factors would lead individuals to learn the relations, the features, or both.

For all four experiments we used the same artificial categories, however the exact set of stimuli varied for every subject. For each subject, stimuli were randomly generated to fit the following constraints. The stimuli were three vertical lines composed of different colored squares. In one category, the line lengths changed monotonically from left to right (either increasing or decreasing), and in the other category the line lengths changed nonmonotoni-

cally (the middle line spatially was either the longest or shortest). Examples of monotonic and nonmonotonic categories are shown in Figure 1, Panel A. These were the relational rules that defined the categories, and they can be generalized across considerable perceptual variation. The positioning of the lines, the exact line lengths, and the differences in lengths across the three lines varied randomly (within the category defining constraints), but every exemplar fit the relational rule. That is, the relative lengths of the left, middle, and right lines was deterministic of category membership and thus it was impossible for an exemplar from one category and an exemplar from the other category to have identical relative line lengths. Importantly, there is little reason to expect that these visuospatial line length relations can be extracted from macroscene statistics in a short glance (e.g., because the category can be increasing or decreasing from right to left, the relations cannot be encoded by a single low-spatial frequency shape, see Alvarez, 2011).

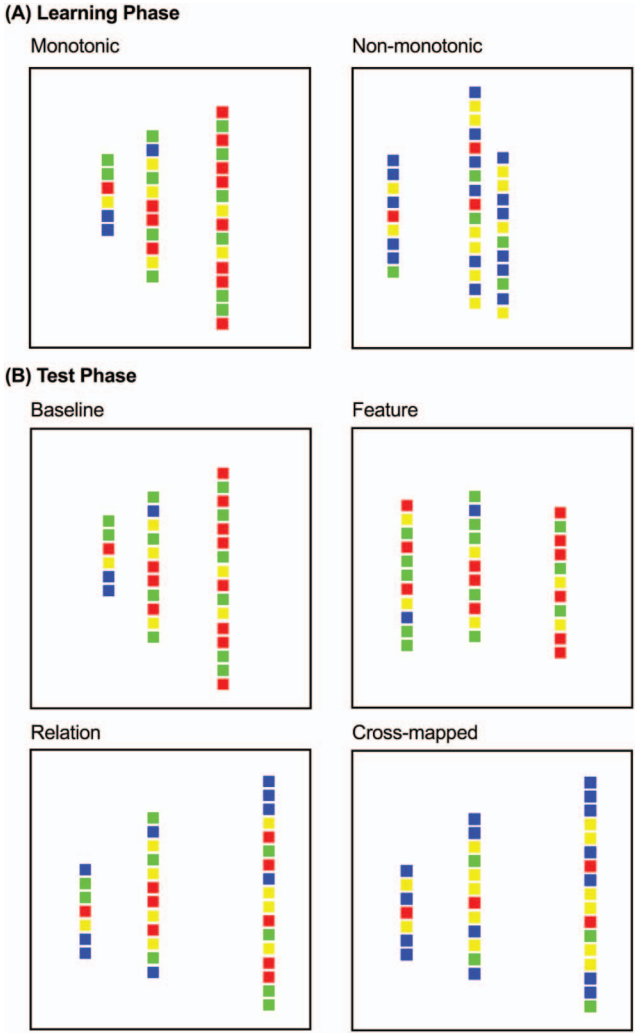

Figure 1. Example stimuli from the categorization task. Panel A shows exemplar arrays from each category. The lines in the left exemplar increase in length monotonically from left to right, and are largely composed of red and green squares, while the lines in the right exemplar do not vary in size monotonically from left to right, and are largely composed of blue and yellow squares. Panel B shows an example of each type of test item. Baseline trials are similar to learned exemplars from each category. Feature trials equate line lengths, such that classification can only be based on the color features. Similarly, relation trials equate the color distributions, such that classification can only be based on the line lengths. Cross-mapped trials have color distributions characteristic of one category and line lengths deterministic of the other category. See the online article for the color version of this figure.

On the other hand, the more surface perceptual features that were part of each category were quite visually salient, based in stark color contrasts, and could be more easily learned via scene statistics. Each square within each line could be one of four colors: red, green, blue, or yellow. The stimuli from one category had, on average, 70% of two of the colors and 30% of the other two colors. The other category had the reverse percentages. Color presentation was stochastic such that the exact percentages varied randomly from trial to trial. If, for instance, the "Blicket" category was associated with the colors red and green then more than a quarter of the squares would be red on about 90% of Blicket trials and more than a quarter of the squares would be green on about 90% of Blicket trials. Further, the color proportion differences were obvious on some trials but relatively subtle on others; in the above example, about a quarter of Blicket exemplars would possess less than 65% red and green squares. Further, it was possible (though unlikely) for exemplars from both categories to have the same color proportions, unlike the completely deterministic line length relations (see below for more detailed stimulus generation description). In summary then, the color features were certainly predictive but were probabilistically related to category labels and were relatively difficult to use on some trials.

During the test phase of each experiment, subjects classified several different test items (See Figure 1, Panel B, for examples of each test trial type). Baseline trials provided another measure of learning accuracy, as they were similar to the stimuli presented during the learning phase, maintaining the association between color distribution and line length. The remaining trials aimed to isolate whether participants had learned about the features or the relations. The feature trials maintained the color distributions from training, with equal line lengths, such that accurate classification required use of the color distribution features. Similarly, relation trials maintained the relative line length differences as training, but with an even color distribution, such that accurate classification required use of the relational rule. Cross-mapped trials were conflicting items that reversed the association between color distribution and line lengths (see Markman & Gentner, 1993). That is, the colors that were more likely to appear in the nonmonotonic line length category were now presented with monotonic line lengths. These test trials assess whether the features or the relations are more prominent in the participant's representation of the categories.

After learning and test phases, participants in all four experiments completed a short 'Far transfer' phase in which the monotonicity/nonmonotonicity relations still defined the same categories, but the rule itself had to be applied to very different stimulus features. That is, the luminance of three arrays of circles appearing in a grid were either increasing or decreasing monotonically from left to right, or the luminance change from left to right was nonmonotonic, for example, the brightest or darkest circles were in the center (See Figure 2 for examples of far transfer stimuli).

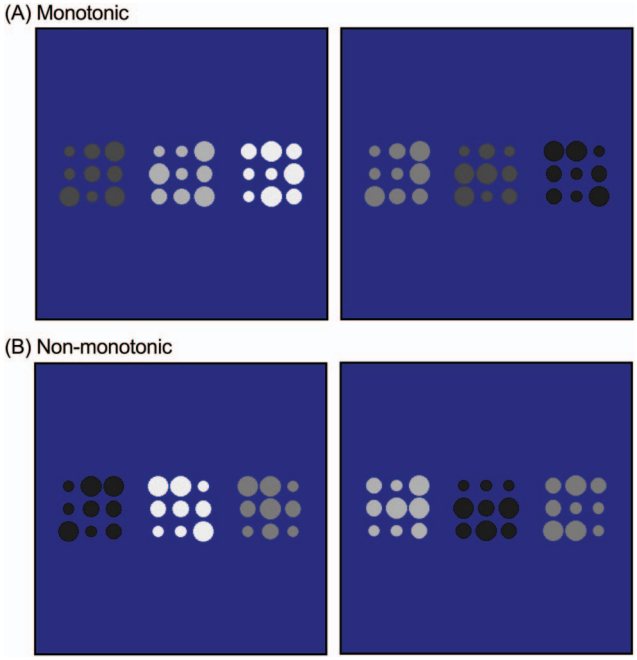

Figure 2. Examples of far transfer phase stimuli. The stimuli in panel A are members of the monotonic category because the luminance of the circles varies monotonically (either increasing or decreasing) from left to right. The stimuli in panel B are members of the nonmonotonic category because the luminance of the circles does not vary monotonically from left to right. See the online article for the color version of this figure.

Unlike the training phase, which possessed predictive color proportions in addition to the line length rule, in the far transfer phase there were no additional featural properties that were correlated with the relational rule, and hence a strategy that exclusively focused on the features during learning would be of no use here.

Corrective feedback was included in this phase for two reasons. First, initial piloting showed that, without feedback, overall subjects performed at chance. Second, it is common in educational settings to only detect differential far transfer performance between learning conditions when additional learning aids are provided along with the transfer materials (e.g., Schwartz & Martin, 2004). That is, in these settings transfer from the learning phase is measured as successful "preparation for future learning" (Bransford & Schwartz, 1999). Indeed, in educational settings sometimes the most relevant form of transfer is not solving a single transfer problem, but bringing appropriate prior knowledge to bear when learning a new topic such that knowledge coherently accumulates over time. Consider for example, how in math education, each new skill prepares you for the next. On the flip side, the degree to which the subjects fail to transfer what they learned in the initial part of the experiment reflects the degree to which their relational knowledge was inert and concrete.[1]

Because of the differences between feature-based and relational categories discussed above, we suspected that different task conditions, and differences between our subjects would systematically lead to differences in what was learned. In theory, however, every participant could learn both the features and the relations because they were positively correlated during learning. To preview the results, across all four experiments it was clear that the features that were probabilistically associated with the categories were easier to learn than the deterministic relations. However, across all experiments, we showed an interesting and systematic pattern of when the featural advantage was modulated and even reversed. Furthermore, not a single subject (of 308) correctly classified more than 80% of both the relation- and feature-isolated test trials. Across all four experiments, performance on the two kinds of test trials were negatively correlated, $r(306) = -.56$, $p < .0001$. See Figure 3 for a plot of each participant's feature and relation test trial accuracy. There were roughly three kinds of subjects: subjects who learned the relations quite well but not the features, subjects who learned the features quite well but not the relations, and subjects who did not learn much of anything. Across the four experiments, our goal was to vary task parameters and attempt to measure the differences across individuals that made learning the relations versus the features more likely.

Task-Structure Manipulations

In Experiment 1, we simply gave verbal hints in the instructions of the task to either focus on the features or the relations. This was to show that generally, subjects could learn either aspect of the categories.

In Experiment 2, we varied the sequence of exemplars into either blocking or interleaving schedules (manipulated between subjects). That is, for the blocking condition, there was a 75% probability that exemplars of the same category repeated across successive trials, while in the interleaving condition there was a 25% chance. Theories of category learning rooted in theories of memory have emphasized the benefits of interleaving when the primary challenge to learning is to discriminate between otherwise confusable categories (e.g., Bjork, Dunlosky, & Kornell, 2013), whereas theories rooted in analogical reasoning predict blocking should support learning the common relations across exemplars (e.g., Gentner, 2010). We predicted interleaving would lead to more featural learning because switching between categories should highlight the distinction in the color distributions as a way to discriminate between categories, while blocking may help discover the relatively less salient relational commonalities (see further discussion below).

In Experiment 3, we directly varied how salient the relations and features were by varying the minimal length difference between the lines, and saturation of the colors. We predicted increasing the minimal line length difference would lead to more relational learning, and decreasing the color saturation would decrease featural learning (because perceptual salience can affect category learning even when predictive validity is not changed, see Kruschke & Johansen, 1999).[2]

In Experiment 4, we varied whether subjects learned via classification (as they had in Experiments 1–3), or whether they learned via inference. In the classification task, participants were

[1] Here, we are not making a distinction between inertness and a lack of abstraction because we believe that the reason knowledge is inaccessible is often because it was encoded in an overly specific or concrete manner.

[2] The data from across our experiments suggest that the features are more salient because on average the cross-mapped data favors feature-based choices. However, as pointed out by an anonymous reviewer, of course one cannot simply generalize this to features versus relations on the whole.

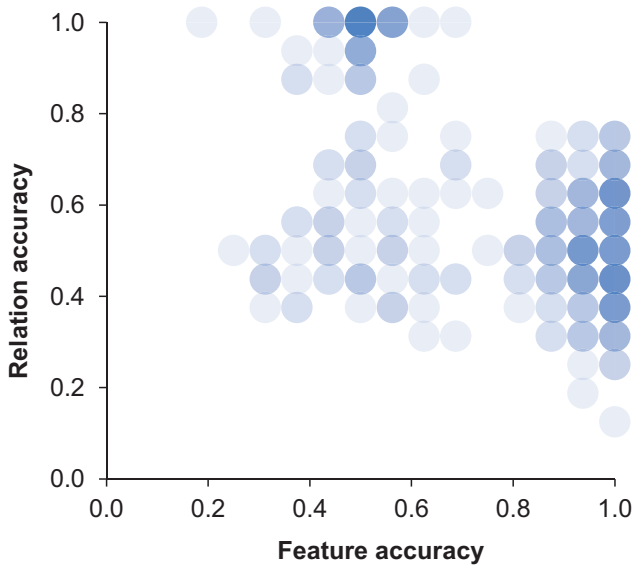

Figure 3. Scatterplot showing the negative relationship for all 308 subjects between Feature and Relation test trial accuracy. Darker circles represent more subjects with equivalent values. See the online article for the color version of this figure.

presented with an exemplar and needed to guess the correct category. In contrast, in the inference-learning task, participants were presented with the category label and an incomplete line array, and had to guess which of two lines completed the pattern (see Markman & Ross, 2003 for review of the work comparing learning via classification vs. inference). In the inference-learning task, the same line that allowed for the successful completion of the relational pattern also possessed the category-appropriate color distribution, while the other line mismatched both in features and relative length. Despite the availability of a perceptual pattern completion strategy, the previous literature on inference-learning shows that the task tends to elicit a strategy wherein subjects focus on the statistical and semantic relations-among the features within a category (e.g., Erickson, Chin-Parker, & Ross, 2005; Markman & Ross, 2003; Sakamoto & Love, 2010). However, the current Experiment 4 was perhaps the strictest test of how inference-learning focuses learners on within-category relations because in this task a much simpler featural solution is also readily available to the subjects.

Individual Differences

In addition to varying these task structure parameters, in Experiments 2–4 we assessed multiple individual differences in learning capacity and learning strategy that could be predictive of learning relational or featural information. Growing evidence suggests that there are large and *stable* individual differences across tasks and content domains in the propensity to transfer structure across examples (e.g., McDaniel, Cahill, Robbins, & Wiener, 2014). However, what is unclear is whether the effectiveness of varying task-structure parameters to foster relational discovery and transfer will depend on these individual differences. One reason this is important is because it is currently unclear if the recommendations that laboratory learning research is making for education only benefit a subset of learners.

We assessed learning capacity using Ravens Progressive Matrices (RPM) as a measure of fluid intelligence (an assessment of how well people can discover novel visuospatial relational rules), operation-span (O-span; a complex working memory span task), which has shown to be predictive of category learning ability independently of strategy (Craig & Lewandowsky, 2012), and the Cognitive Reflection Test (CRT; Frederick, 2005), which measures the tendency of individuals to inhibit immediately available responses and engage in deliberative processes, and has been shown to be predictive of rule-based transfer in causal learning (Don, Goldwater, Otto, & Livesey, 2016).

To assess learning strategy, we focused on related distinctions developed for cognitive laboratory experiments (Experiments 2 and 3), and to characterize "real-world" study habits (Experiment 4). To measure learning strategy in Experiment 2 and 3, we used a laboratory-based task from Little and McDaniel (2015b) designed to distinguish whether individuals enter a learning task in the mindset to memorize which sets of exemplars belong to which category, or whether they intend to search for rules or principles that underlie all the exemplars in a category (see Yang & Lewandowsky, 2004; Little & Lewandowsky, 2009 for a similar distinction). This distinction between rule versus exemplar learning is reminiscent of a distinction between "deep" and "surface" learning strategies that has been a focus in the educational psychology literature for the past 40 years (Marton, & Säljö, 1976; see Alexander, Peterson, Dumas, & Hattan, 2016 for review). Deep learning strategies seek to elaborate upon the learning materials, integrating multiple ideas, and finding more general principles. Surface learning strategies focus more on just learning what is directly presented. In Experiment 4, we assessed differences in deep versus surface learning strategy using the Motivational Strategies for Learning Questionnaire [MSLQ] developed by Pintrich and Degroot (1990), which is perhaps the most common assessment that educational researchers have used to measure study habits. See the General Discussion for comparisons between these constructs from cognitive and educational psychology.

Across these tasks, we sought to answer multiple novel questions with regard to these individual differences in capacity and strategy. First, we asked whether or not cognitive capacity and learning strategy are correlated, and then whether they predict similar or differing aspects of categorization performance. Across McDaniel et al. (2014) and Little and McDaniel (2015b), there was an inconsistent pattern of whether strategy and capacity are related, so we aimed to add some clarity to this issue. Second, we asked whether capacity or strategy interact with task structure. As mentioned above, the premise for doing so is that task manipulation may facilitate relational learning for some individuals more than others, depending on their abilities and strategies. For example, if in Experiment 2, blocking exemplars led to more relational learning overall, this effect may be confined to those who are already rule-oriented or those who demonstrate relatively high cognitive ability, or it may benefit those who are not rule-oriented or who demonstrate relatively poor cognitive ability. These interactions are important for how we interpret the effects of manipulating task parameters. To make applied recommendations, it would be ideal

to identify task manipulations that promote relational discovery for learners with suboptimal strategies or capacities, while not costing more capable learners (or even benefitting them as well).

Experiment 1

Experiment 1 served as a simple proof of concept, establishing the presence of feature-based and relational learning in the basic classification task used in all experiments. Three groups were each given the classification task with differing initial instructions. The 'feature hint' group was explicitly told to pay attention to the colors of the squares, the 'relation hint' group was explicitly told to compare the lengths of the lines, and the 'no hint' group was given neither, matching the instructions used in the later experiments. These hints were designed to strongly bias learning toward either the featural or relational property of the task (see Kalish, Lewandowsky, & Davies, 2005 for how hints can elicit changes in categorization strategy; and Spalding & Ross, 1994 for how explicit hints and directions to compare examples focus attention to aspects of the stimuli in a similar manner). Instructions can have a strong influence over the direction of attention to specific features and the strength of subsequent learning in feature-based learning tasks, though their effect does not entirely overwhelm other biases carried over from previous learning history (e.g., Don & Livesey, 2015; Jung & Hummel, 2015; Shone, Harris, & Livesey, 2015). We expected that these explicit hints would be sufficient to guide attention toward (and away from) relevant featural and relational properties of the task.

Participants then completed training on the basic classification task with corrective feedback. They were then tested in the absence of feedback using a range of test items that (a) tested baseline performance using more training exemplars, (b) isolated the predictive features by presenting the color proportions used in training on three lines of equal length, (c) isolated the diagnostic relational rule by presenting lines of different lengths with equal proportions of the four colors, and (d) pitted the features and relation against each either by presenting the color proportions predictive of one category on lines arranged to follow the rule that defined the other category. Finally, they completed a 20-trial Far Transfer phase wherein corrective feedback was reintroduced.

Method

This experiment, and the following three were approved by the University of Sydney Human Research Ethics Committee protocol #2013/1077, titled "Learning from Examples."

Subjects. Fifty-four psychology undergraduate students (30 female; mean age = 20.3 years) from the University of Sydney participated in this experiment in return for course-credit. There were 18 subjects each in the No Hint, Feature Hint, and Relation Hint conditions (randomly allocated).

Materials. The experiment was programmed and run via PsychToolbox version 3 for Matlab (Kleiner et al., 2007). Stimuli during the learning phase were generated with the following constraints. Each exemplar comprised a visual display of small squares that formed larger vertical lines (see Figure 1 for an example). Every display had three lines that varied in length from six to 18 squares, with the constraint that each line was different in line length to the other two by at least one square. The three lines were placed probabilistically in a 20 × 20 square grid display. The categories, labeled as "Snarg" and "Blicket" were defined in two ways. First, they were defined by the relations among the line lengths. In one category, the line lengths would vary in size monotonically from left to right (note that this included equal numbers of exemplars varying from shortest to longest and longest to shortest moving left to right). In the other category, the lines varied nonmonotonically such that the shortest or longest line was always in the middle. The length of each line was randomized with the constraints that the shortest line had at least 6 squares, the longest line had no more than 18 squares, and the three lines were of different lengths that retained the predetermined relational properties. Any given Snarg exemplar could have three lines of the same length as those in a Blicket exemplar, but these lines could not be spatially arranged in the same order moving left to right. The relative lengths of the left, middle, and right lines was deterministic of category membership and thus it was impossible for a Snarg exemplar and a Blicket exemplar to have identical relative line lengths.

In addition, each square was one of four colors (red, green, blue, and yellow) sampled stochastically using proportions that differed for the two categories. One category had a prototypical color distribution of 35% red, 35% green, 15% blue, and 15% yellow, and the other category had a prototypical color distribution of 15% red, 15% green, 35% blue, and 35% yellow. The color of each square was determined by sampling without replacement from a large pool of color values. Specifically, for each trial, an array of 100 color values was created, 35 of each of the two more prevalent colors for that category and 15 of each of the two less prevalent colors for that category. The color values were then randomly shuffled and each successive value in the array was used to color one square until all squares of all three lines were complete. Because the three lines together used 20–50% of the total pool, this means that color sampling was probabilistic. Sampled in this way, it was possible (though in practice unlikely) that an exemplar from the Snarg category and an exemplar from the Blicket category could have identical color proportions.

In the learning phase, there were 184 trials with corrective feedback. Trial order was pseudorandomized such that over the course of training there was an equal probability of repeating the same category versus alternating categories from one trial to the next.[3] In the test phase, there were four kinds of trials, with 16 exemplars of each kind (8 for each of the two categories), making 64 trials in total. Baseline trials presented stimuli of the same kind as the learning phase. Relation test trials had flat color distributions, that is, 25% of squares were from each of the four colors. Thus, accurate classification required using the relative line lengths. Feature test trials presented lines of equal lengths, but with the color distributions of the learning phase. Thus accurate classification required use of the color distributions. "Cross-mapped" test trials reversed the associations between color distributions and relative line length. That is, the color distribution of the monotonic

[3] We used this form of trial randomization in anticipation of running Experiment 2, where trial sequencing became a key focus. The method was developed by Nicks (1959), and is widely used in human conditioning and reaction time (RT) experiments where unpredictable sequencing is desirable, for instance in the Perruchet effect (e.g., see Lee Cheong Lem, Harris, & Livesey, 2015; Livesey & Costa, 2014; Perruchet, 2015 for a review).

line length category were now displayed with the nonmonotonic line length category and vice versa.

In the Far Transfer phase, stimuli used the same relational principle (monotonic vs. nonmonotonic), but were designed to be surfacely vastly different. A rectangular array of circles of different shades of gray was shown against a blue background. The circles were arranged into 3 sets of 3 × 3, with the three sets positioned adjacent to one another horizontally (see Figure 2 example stimuli). The circles within each set varied in size in an uninformative manner (3 sizes, 3 of each size within each set, organized randomly). All circles within a set were the same shade of gray (drawn from 5 possible luminance values) but the three sets always varied in shade. Correct classification could only be achieved through the order of gray shades, which was monotonically lightening or darkening for one category, and changing nonmontonically for the other category such that the middle set was either the lightest or the darkest. The relational rule in training and far transfer always matched (i.e., the monotonic category for the lines was always the monotonic category for the gray shades).

Procedure. Each participant was seated at an individual workstation and completed the experiment using a desktop PC. Participants first received instructions on screen informing them that they would see displays of colored patterns that belonged to two different categories, "blickets" and "snargs." They were instructed that there were multiple things to learn from the displays, and that their task was to discover which patterns belonged to which category, based on the feedback provided.

The Feature Hint additionally received instructions that read "Note: here's a hint for how to solve the task. Pay attention to the colors that appear in each pattern. The Relation Hint received instructions that read "Note: here's a hint for how to solve the task. Compare the lengths of the lines that appear in each pattern." The No Hint group was not given any additional instructions.

Participants completed 184 trials of the training stage in random order.[4] On each trial, the stimulus appeared centrally, and the category names 'Snarg' and 'Blicket' appeared immediately below. Participants selected their response by clicking on the appropriate category label and then pressing the spacebar. Responses were self-paced, and followed by corrective feedback. Prior to the test phase, participants were instructed that they would see some new examples that may look different from the displays they had already seen, and to classify each example into the categories, based on what they had already learned. Similarly, prior to the far transfer phase, participants were told that they would see new displays that look very different, and to try to classify each one into the categories they had learned. The procedure for the test and far transfer trials was the same except that for the test trials, feedback was omitted. The task took approximately 25 min in total.

Results and Discussion

There were several measures for this task, given it had three distinct phases (learning, test, and far transfer), and that there were four kinds of test trials. Throughout the four experiments, we analyzed the learning phase as a whole (to compare across conditions), then examined condition effects for each kind of test trial individually.

Group means and statistical analysis of group effects for the learning, test, and far transfer phases of Experiment 1 are summarized in Table 1. The hint had no significant effect on accuracy during the learning phase, or performance on baseline trials. However, there were significant group effects on the test trials that isolated what was learned. The feature hint and no hint groups showed better accuracy on feature trials than the relation hint group, $t(51) = 4.05$, $p < .001$. The relation hint group showed greater accuracy on relation trials compared to the other two conditions, $t(51) = -3.86$, $p < .001$. Similarly, the relation hint groups showed greater relationally consistent responses on cross-mapped trials than the other conditions, $t(51) = -4.95$, $p < .001$. There were no significant differences between the no hint and feature hint groups on any of these trial types, highest $t(51) = 1.58$, $p = .12$. Importantly, although the relation hint improved relational performance on the test trials, it did not result in improved performance in the far transfer phase, where there was no significant difference between conditions.

In summary, this first experiment clearly showed a few important findings. (a) Without any sort of hint, overall subjects in this task learn the distributions of colored squares (i.e., the features) associated these categories more than they learn the rule pertaining to the relative line length. (b) When instructed to focus on the relative line lengths, subjects were generally capable of learning the relational rule. However, the subjects who were given the relation hint did not carry any sort of advantage into the far transfer phase of the experiment. This pattern is reminiscent of the inert knowledge problem where students do not seem to apply what they have learned to novel stimuli. We will refrain from discussing the nature of these initial far transfer results in any further detail until the General Discussion, where we consider all four experiments together.

Experiment 2

In the next experiment, we focused on how the sequence of exemplars might shift focus toward the relations without any explicit hint in the instructions. We ran two between-subjects conditions, manipulated during the learning phase: blocking, which presents longer sequences of exemplars from the same category, and interleaving, which presents sequences of exemplars that instead tend to alternate between categories exemplars. In a recent review, Bjork et al. (2013) argued that interleaving leads to better learning because it supports between-category discrimination via repeatedly highlighting differences between exemplars from different categories. The interleaving advantage was consistent across many learning tasks, such as discriminating between

[4] We used a unique randomized trial order and stimulus randomization for every participant in the category learning task in these experiments (stimuli and trial order were identical for participants in the tasks measuring the individual differences of interest, e.g., the word-learning task, o-span, ravens, CRT, and later the MSLQ). In this study, we had to reach a compromise between, on the one hand, the aim of comparing experimental variables manipulated at a group level and, on the other hand, differences explained at the level of individuals. The first of these aims is better served by unique randomization. However, this is not ideal for the second aim of studying individual differences because it introduces a source of variance that is confounded with real individual differences in, say, learning orientation and cognitive ability. Future research looking primarily at individual differences may be better served by using identical stimulus and trial sequencing across the cohort.

Table 1
Results From the Learning, Test, and Far Transfer Phase of Experiment 1

Phase	Means			Group effects
	No hint	Feature hint	Relation hint	
Learning	.80 (±.04)	.88 (±.03)	.80 (±.04)	$F(2, 51) = 1.53$, $p = .226$, $\eta_p^2 = .06$
Test				
Baseline	.86 (±.05)	.92 (±.05)	.90 (±.03)	$F(2, 51) = .36$, $p = .7$, $\eta_p^2 = .014$
Feature	**.79 (±.06)**	**.90 (±.04)**	**.60 (±.05)**	$F(2, 51) = 9.33$, $p < .001$, $\eta_p^2 = .27$
Relation	**.58 (±.05)**	**.57 (±.04)**	**.81 (±.05)**	$F(2, 51) = 7.46$, $p = .001$, $\eta_p^2 = .23$
Cross-mapped	**.29 (±.09)**	**.11 (±.06)**	**.69 (±.09)**	$F(2, 51) = 13.52$, $p < .001$, $\eta_p^2 = .35$
Far transfer	.66 (±.04)	.59 (±.05)	.63 (±.04)	$F(2, 51) = .60$, $p = .55$, $\eta_p^2 = .02$

Note. Standard error of the mean (*SEM*) is shown in parentheses. Cross-mapped trial means represent the proportion of relationally consistent responses. All other means represent accuracy. Significant results are highlighted in bold.

different painters' works (Kornell, & Bjork, 2008), or between different classes of math problems (Rohrer & Pashler, 2010). In each of these cases, the primary obstacle to learning was discrimination, as for example many incorrect answers to math problems were because students misapplied the solution procedure from a different category of problem that on the surface looked quite similar.

However, more recent data have complicated the issue. Carvalho and Goldstone (2014, 2015a) showed that interleaving is beneficial when exemplars between categories are easily confused, such that between-category discrimination is the key barrier to learning, but this interleaving advantage is mitigated when within-category exemplar similarity is low (see Higgins & Ross, 2011; Higgins, 2017 for a similar pattern). In this low-similarity case, discovering what exemplars from the same category have in common is the main barrier to learning, so blocking becomes advantageous presumably because it affords within-category comparisons on successive trials. Supporting this interpretation, Carvalho and Goldstone (2015b) and Rawson, Thomas, and Jacoby (2015) present additional evidence that when task properties promote within-category comparisons (such as by providing category labels and definitions, respectively), the interleaving advantage is eliminated (See Carvalho & Goldstone, 2015b).

This complex pattern across several papers motivated the hypothesis tested in Experiment 2. We predicted interleaving would lead to more featural learning because the featural distinction was more salient, while blocking may help discover the less salient relational commonalities. The color proportions that form the basis of feature-based transfer are relatively similar across exemplars within each category and differ between categories. In contrast, the line lengths that one needs to compare to discover and transfer the relational rule vary within each category as much as they do between categories, even though the rule derived from a comparison of those features is deterministic. Therefore we expected that blocking learning trials during training may facilitate relational discovery in our task.

In addition, we assessed subjects on several measures of their cognitive capacity and learning strategy to measure what individual differences predicted relational versus featural learning, and whether any of these differences interacted with the blocking versus interleaving manipulation. Measures of cognitive capacity included RPM, O-span and CRT tasks.

To assess learning strategy, we used the word-learning task from Little and McDaniel (2015b), wherein subjects learn to categorize 12 nonce words as names for either animals or plants. The final letter of the nonce word deterministically predicted category membership. Subjects could successfully categorize the 12 training items by either discovering and applying the last-letter rule, or by memorizing which six exemplars belonged to each category. After training, there was a test phase with additional nonce words. Some had one of the rule-following final letters, but were otherwise novel items ("rule-trials," analogous to our relation test trials). Some were identical to previous words, except now had a final letter that was different from either of the rule-following letters ("exemplar trials," analogous to our feature test trials), and some had the same letters as a word from training except now had the final letter of the other category ("ambiguous trials," analogous to our cross-mapped trials). After the test phase, Little and McDaniel simply asked subjects to indicate their strategy during learning, on a continuous scale from entirely relying on memorization of exemplars to entirely focused on searching for a rule. The learner's self report predicted their performance on the test phase items. Rule-learners performed better on rule trials and indicated rule-consistent responses on the ambiguous trials. Exemplar-learners performed better on exemplar-trials and indicated exemplar-consistent responding on ambiguous trials.

In the current experiment we used this same task, and rely on this same self-report learning strategy item, to test whether self-reported strategy from the word-learning task predicts performance on the primary categorization task. In the same paper, Little and McDaniel ran an additional categorization task wherein the exemplars were pairs of shapes, and the categorization rule concerned whether they were of identical shape and color (i.e., it was a relational rule). One of the points of their paper was that this individual difference in the propensity to search for rules is independent of whether the rule is a relational rule, as in the shapes task, or based on a single discriminatory feature, as in the word-learning task. Building on their finding, we used the word-learning task because we wanted to assess this rule-searching strategy more generally, and not just measure a more specific orientation toward relational rules. Some have argued that rule learning generally relies on the relational alignment of exemplars (Gentner &

Medina, 1998), suggesting we should see connections across these tasks.

Method

Subjects. One hundred psychology undergraduate students from the University of Sydney participated in this experiment as part of their course work. Seven were eliminated from analyses for having substantial data missing, and seven more were eliminated for scoring below 25% accuracy on RPM.[5] This left 86 participants (56 female; mean age = 21.8 years), 43 subjects in the blocking condition, and 43 in the interleaving condition. Because of a technological error, data from the O-span assessment were missing for an additional 26 subjects, so for those analyses only, 60 subjects' data were included.

Materials.

Categorization task. The primary category-learning task was identical to the No hint condition of Experiment 1, except for the sequencing of trials during the learning phase, which was manipulated between-subjects to form blocking and interleaving conditions. In the blocking condition successive trials presented exemplars from the same category 75% of the time (i.e., a high probability of category repetition), whereas in the interleaving condition, exemplars from the same category only followed each other 25% of the time (i.e., a high probability of category alternation). The test and far transfer phases of the experiment were identical to Experiment 1 in both conditions.

Individual differences measures. The following tasks were included as measures of individual differences in learning strategy and capacity.

Word-learning study. The second task in the experiment was the "word-learning study" from Little and McDaniel (2015b, Experiment 2), used to distinguish learners oriented to discover underlying rules from learners attempting to memorize exemplars. In this task, participants categorized 12 articulable nonword letter strings as either plants or animals. Each letter string had a distinct first letter and stem, but ended with either a "k" or a "t." The final letter determined the category of the letter string, for example, all plants ended in "k." As noted above, this is not a relational rule; it is a rule that focuses on a single feature of each stimulus. Our motivation for using this task was to examine whether being oriented toward deterministic single-feature rules would generalize to the relational rules in our task.

During training, participants categorized each letter string six times, while receiving corrective feedback. In a transfer phase, participants categorized four trained items and 12 new items, consisting of four ambiguous, four rule-favored and four exemplar-favored transfer items. The ambiguous transfer items took the stems of trained items and switched the final letter, for example, the final letter became "t" if it was previously a "k." Rule-favored transfer items had novel stems, but the final letter was either a "t" or a "k," and exemplar-favored items had trained stems with novel final letters (these trials are similar to the current cross-mapped trials). The rationale for this design was that exemplar-strategy subjects would base their classifications on word-stems, while rule-strategy subjects would base their classifications on the last letter only. Stimuli were identical for each participant. Training and transfer phase items were presented in a random order that was the same for all participants.

This task was followed by a questionnaire querying the subjects' strategies and whether they were aware of a rule that defined the categories of letter-strings. Five questions were included based on those used in Little and McDaniel (2015b; see Appendix for all questions), however, as in Little and McDaniel, analyses focused on the question "While you were learning the categories, were you more focused on trying to learn the individual items, or trying to develop a rule for why items were members of each category?" Responses were made on a linear analog scale ranging from "Relied solely on memorization" to "Relied solely on developing a rule," which was translated to a numeric score from 0 to 100.

Raven's progressive matrices. Subjects then completed an abbreviated form of the RPM (Raven, 2000). Subjects chose which of eight possible choices would best fit a missing cell in a 3×3 matrix. One choice was most appropriate to complete both the vertical and horizontal patterns within the matrix. After two practice trials with feedback, subjects had unlimited time to complete 20 individually presented trials, without feedback. This took on average 10 min.

O-span. Next, subjects completed the O-span test of working memory (Turner & Engle, 1989; Lewandowsky, Oberauer, Yang, & Ecker, 2010). The task was run in an identical manner to Lewandowsky et al. (2010), and involved serial recall of a list of letters presented in conjunction with a secondary task. On each trial, participants were presented with a fixation cross for 1.5 seconds. An arithmetic equation would then appear on the screen, for example, (2 + 4 = 6), and subjects were required to judge the accuracy of the equation by the pressing the "/ " and "Z" keys to make "Yes, this is correct" and "No, this is not correct" responses, respectively. After a response was made, or the maximum response time of 3 seconds had passed, the equation disappeared and a letter appeared on the center of the screen for 1 second. Subjects were required to encode this letter for later recall. The next equation appeared after a 100-ms blank interval. This sequence continued until all equations and letters in the list had been presented. Subjects were then asked to recall each letter in the order of their presentation. A question mark and blinking underscore appeared, and participants typed all letters that were presented during the trial, in the correct order of their presentation. There was an intertrial interval of 500 ms, and participants could take a self-paced break after every 3 trials.

Letters were all consonants (excluding Y and Q), and lists contained no letter repetitions. The first operand in each equation varied between 1 and 10, and the second varied between −9 and 10, excluding 0. Results were all positive integers. The number of equations and letters in each list ranged from four to eight. Each list length was presented 3 times each, resulting in a total of 15 trials. There were three practice trials prior to the start of the experimental trials. All participants received the same order of equations, letters and trials. A score was generated for each participant as the proportion of accurately recalled letters.

[5] Because we used a highly capable and generally intelligent population, scores near chance on Ravens are unrealistically low. We observed a small number of scores in the range of 5% to 25% that were accompanied by very fast response times, strongly suggestive of a nonserious attempt at RPM (and potentially other components of the experiment) rather than low fluid ability, thus they were removed completely from the analysis.

Cognitive reflection task. Finally, the subjects completed a computerized version of the Cognitive Reflection Task (CRT). The task consisted of three questions taken directly from Frederick (2005; see Appendix). The items were developed so that each quickly elicits an intuitive answer, which is incorrect, and participants must think more reflectively to arrive at the correct answer. Responding was self-paced. Participants typed their responses using the keyboard.

Procedure. Each subject sat individually at a desktop PC. One program took them through the entire series of experiments in the order previously described: the relational categorization task, the word-learning task, Raven's matrices, the O-span, and then the CRT. The series of tasks took the majority of a 2-hr class period.

Results

First we analyzed the primary category-learning task in a manner similar to Experiment 1. Then, we related the individual differences assessments to performance on this primary task (see the Appendix for a complete table of correlations).

Categorization task. Group means and statistical analysis of group effects are shown in Table 2. In the primary categorization task, trial sequencing did not affect categorization accuracy during learning, or accuracy on baseline, feature, or relation test trials. However, blocking resulted in greater relationally consistent responses than interleaving on the cross-mapped trials (though relational responding was still relatively low across the board). There was no effect of trial sequencing on performance in the far transfer phase. However, the effects of trial sequencing become clearer once taking into account the individual differences measures.

Individual differences. We analyzed the connection between learning tasks and the individual difference measures. Because there are many measures across several tasks, we will present in the main body of the text what is theoretically driven and informative. For the primary category learning task, we will limit these analyses to predicting accuracy in the learning phase, performance on the relation test trials specifically, and far transfer, but see the Appendix for the complete matrices of correlations.

Learning capacity. To assess relationships with cognitive capacity, we used the subject's accuracy on each of the RPM, O-span, and CRT. RPM significantly predicted accuracy during the learning phase ($r(84) = .26$, $p = .015$), but no measure of cognitive capacity predicted performance on relation trials, or learning strategy.[6]

Learning strategy. To assess relationships with learning strategy, we focus on participants' rating on the learning strategy question during the word-learning task, because it most directly assessed how they approached the task initially, and it was the basis for how Little and McDaniel (2015b) classified their subjects. Similar to Little and McDaniel's (2015b) results, the learning orientation question was highly correlated with the other questions (i.e., strategy during transfer, and whether or not they noticed the rule).

Overall, there was no relationship between learning strategy and performance on relation test trials, $r(84) = .096$, $p = .38$, but when split between blocking and interleaving conditions, an interesting pattern was revealed. There was a significant positive relationship between learning strategy and relation test trial performance for the blocking condition, $r(41) = .35$, $p = .02$, and a nonsignificant negative relationship for the interleaving condition, $r(41) = -.18$, $p = .249$. Critically, there was a significant difference between the correlations, $z = 2.45$, $p = .014$. This pattern is essentially an interaction between strategy and task structure. None of our three capacity measures showed this pattern, that is, they showed no relationship with relation learning overall, and no relationship for either the blocking or interleaving conditions specifically. Strategy uniquely interacts with relational learning.

To further investigate this interaction, we split subjects into self-reported rule- and exemplar-oriented groups based on the 1–100 score on the learning strategy question. We classified the top 40% of subjects within each between-subjects condition as rule-learners, and the bottom 40% of subjects as exemplar-learners, as this allowed for a clear separation between the groups on this measure of strategy, while still keeping the majority of our subjects for further analysis. For this experiment, that led to four groups of 17 subjects each (exemplar-blocking, rule-blocking, exemplar-interleaving, rule-interleaving). Figure 4 plots the results from each phase of the categorization task for each of these groups.

There was a significant interaction between learning strategy and trial sequence for relation trials, $F(1, 64) = 7.05$, $p = .01$, $\eta_p^2 = .10$. Further analysis of simple effects showed that rule-learners were more accurate in the blocking condition than in the interleaving condition, $t(32) = 2.88$, $p = .007$, $d = 0.99$, although there was no effect of trial sequencing on exemplar-learners, $t(32) = -0.78$, $p = .44$, $d = -0.27$. Similarly, this interaction was also significant for the cross-mapped trials, $F(1, 64) = 4.05$, $p = .048$, $\eta_p^2 = .06$. Analysis of simple effects showed that rule-learners again showed greater relation-consistent responding in the blocking than interleaving condition, $t(32) = 3.05$, $p = .005$, $d = 1.05$, whereas there was no difference for the exemplar-learners, $t(32) = 0.22$, $p = .826$, $d = 0.08$. No other interactions were significant, all Fs < 1, including (most notably) for far transfer accuracy. Thus, even rule learners in the blocking condition (who showed the most relational responding at test overall) were not able to transfer this relational knowledge to the far transfer phase any better than the other participants.

Discussion

In Experiment 2, the manipulation of trial sequencing elicited modest overall changes in relational learning, as seen most clearly in the cross-mapped trials. However, these changes interacted in an interesting way with learning strategy. On the whole, exemplar-learners learned more about the surface features than the relations, and trial sequencing seemed to do little to change that. On the other hand, trial sequencing had a large effect on rule-learners, and little effect on exemplar-learners, shown by the interaction between learning strategy and trial sequencing on relation and cross-mapped test trials. For rule-learners, the blocking condition eliciting more relational learning than the interleaving condition. However, even the rule-learners in the blocking condition failed to show a relation

[6] Because of the missing O-Span and CRT data, we do not interpret these null results too strongly.

Table 2
Results From the Learning, Test, and Far Transfer Phase of Experiment 2

Phase	Means		Group effects
	Blocking	Interleaving	
Learning	.82 (±.02)	.82 (±.02)	$t(84) = -.29$, $p = .773$, $d = -.06$
Test			
Baseline	.85 (±.03)	.88 (±.03)	$t(84) = -.78$, $p = .436$, $d = -.17$
Feature	.79 (±.04)	.87 (±.03)	$t(84) = 1.78$, $p = .079$, $d = -.38$
Relation	.59 (±.03)	.52 (±.03)	$t(84) = 1.91$, $p = .06$, $d = .41$
Cross-mapped	**.34 (±.06)**	**.18 (±.04)**	$t(84) = 2.19$, $p = .031$, $d = .47$
Far transfer	.68 (±.03)	.68 (±.03)	$t(84) = .03$, $p = .979$, $d = .01$

Note. Standard error of the mean (*SEM*) shown in parentheses. Cross-mapped trial means represent the proportion of relationally consistent responses. All other means represent accuracy. Significant results are highlighted in bold.

preference overall, that is, they were evenly split between relational and feature-based responding for the cross-mapped trials. Further, they did not seem able to apply their knowledge of the relational rule in the far transfer phase, given they had similar far transfer scores to the rule-learners from the interleaving conditions.

In contrast to measures of learning strategy, cognitive capacity did not predict relational learning specifically. RPM did predict overall rates of learning, however. It is perhaps particularly surprising that RPM did not predict a focus on relational learning, given that RPM is a relational reasoning task. However, Little and McDaniel (2015b) also found that RPM did not predict a rule-learning strategy generally or relational rule learning specifically (but unlike that result, RPM did not predict how well rule-learners discovered the relational rule in the current task). This suggests that learners' approaches to classification tasks are less constrained than tasks such as RPM, and that learners vary in how they selectively *apply* their relational thinking abilities, in addition to varying in their relational thinking abilities.

Experiment 3

Blocking promotes comparison of exemplars from the same category across consecutive trials, which highlights within-category similarity. For individuals searching for rules, blocking in Experiment 2 made the relational structure more salient, suggesting that relational discovery specifically relies upon recognition of within-category similarity. However, even for these individuals, they still showed no relational preference in the cross-mapped trials on the whole. In Experiment 3, we manipulated relational salience more directly by changing the minimum line length difference between the three lines in each exemplar (in terms of number of squares), rendering this relational quality of the exemplars easier to process and perhaps making comparisons of multiple exemplars less necessary. In Experiments 1 and 2, every line was a different length, with a minimum difference of only a single square. This could make the relative lengths of the lines a seemingly unlikely characteristic to build a hypothesis around, or simply render testing such a hypothesis too difficult. In Experiment 3, the minimum line length difference was either one square (as before; hereafter "1-square") or four squares (hereafter "4-square"), manipulated between subjects. We predicted that the 4-square condition would lead to substantially more relational learning. However, perhaps more interestingly, this also allowed us to further examine the existence of interactions between strategy and task structure. That is, if a learner's orientation were not toward the discovery of relational rules, then perhaps making the relevant relation much more visually salient would have no effect. In addition to changing the salience of the relation, we manipulated the salience of the features by changing the saturation of the colored squares. We hypothesized that reducing feature salience could potentially reduce the degree that the colors captured the learner's attention, which may shift attention toward other characteristics of the exemplars, including the relative line lengths.

Method

Subjects. Ninety-two psychology undergraduate students from the University of Sydney participated in this experiment in exchange for course credit. Five were excluded from analyses for having substantial data missing or for scoring below 25% accuracy on RPM, leaving 87 participants (54 female, mean age = 21.7 years).

Materials. The primary categorization task was altered from the previous experiments in the following ways: First, as in Experiment 1, there was a 50/50 chance of repeating or alternating categories on successive trials. Second, there were two variables manipulated between subjects, minimum line length difference (1-square vs. 4-square), and color saturation (full-saturation vs. half-saturation), creating four between-subjects conditions. Third, the far transfer phase was expanded to 40 trials to increase the ability for the effects of learning conditions and learning strategies to be detected. RPM and the word-learning task from Little and McDonald (2015b) were used, but CRT and O-span were not repeated as they added no additional information in Experiment 2.

Procedure. The procedure was the same as before, with the primary category task first, followed by the word-learning task, and then RPM. As noted, to increase the possibility of showing evidence of far transfer, we doubled the number of far transfer trials.

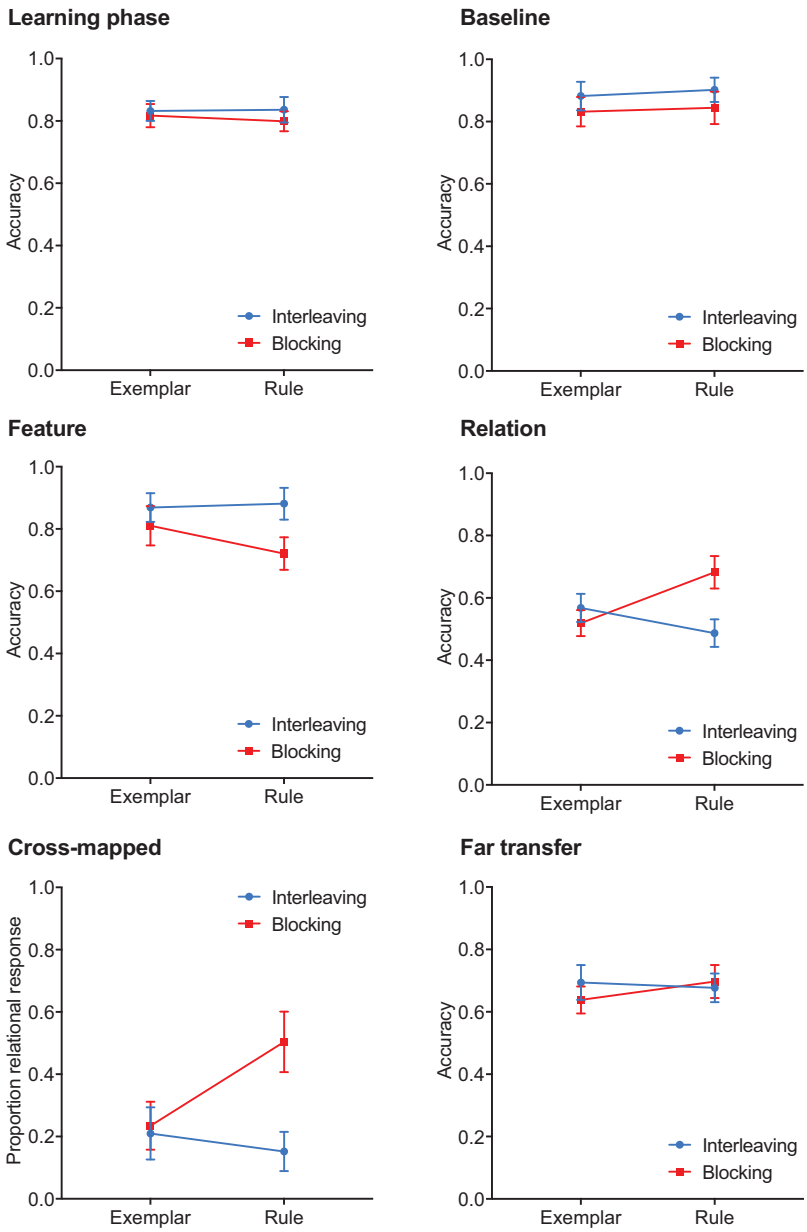

Figure 4. Experiment 2 results for exemplar- and rule-learners. Cross-mapped results are plotted as the proportion of relationally consistent responding. Baseline, Feature, and Relation trials report accuracy. Error bars show standard error of the mean. See the online article for the color version of this figure.

Results

Categorization task. Group means and statistical analyses of group effects are shown in Table 3. Unlike in the previous experiments, there was a strong effect of task manipulation on accuracy during the learning phase. Specifically, although there were no main effects of line length, or color saturation, there was a significant interaction indicating that reducing the color saturation reduced accuracy in the 1-square condition, although it increased accuracy in the 4-square condition.

On baseline trials, there was no significant effect of line length. However, participants in the full-saturation condition showed greater accuracy than in the half-saturation condition, and this was moderated by an interaction with line length, in which the benefit of full-saturation only occurred in the 1-square condition, $F(1, 41) = 9.93$, $p = .003$, $\eta_p^2 = .195$, but not the 4-square condition, $F < 1$.

There were no further interactions between line length and color saturation on the remaining test trials. Increasing the line length resulted in greater accuracy on relation trials, a greater proportion of relational responses on cross-mapped trials, and reduced accuracy on the feature trials. Reducing the color saturation reduced accuracy on feature trials but had no effect on relation, or cross-mapped trials.

Table 3
Results From the Learning, Test, and Far Transfer Phase of Experiment 3

Phase	Means				Group Effects		
	1-square		4-square				
	Full-sat	Half-sat	Full-sat	Half-sat	Main effect color	Main effect line length	Interaction color × line length
Learning	.85 (±.03)	.73 (±.04)	.79 (±.04)	.85 (±.03)	$F(1, 83) = .67$, $p = .415$, $\eta_p^2 = .01$	$F(1, 83) = .61$, $p = .436$, $\eta_p^2 = .01$	$F(1, 83) = 6.77$, $p = .011$, $\eta_p^2 = .08$
Test							
Baseline	.95 (±.02)	.79 (±.05)	.89 (±.04)	.89 (±.04)	$F(1, 83) = 4.69$, $p = .03$, $\eta_p^2 = .05$	$F(1, 83) = .50$, $p = .481$, $\eta_p^2 = .01$	$F(1, 83) = 4.69$, $p = .03$, $\eta_p^2 = .05$
Feature	.86 (±.04)	.73 (±.05)	.74 (±.05)	.65 (±.05)	$F(1, 83) = 4.85$, $p = .03$, $\eta_p^2 = .06$	$F(1, 83) = 4.04$, $p = .048$, $\eta_p^2 = .05$	$F(1, 83) = .15$, $p = .698$, $\eta_p^2 < .01$
Relation	.55 (±.05)	.58 (±.04)	.64 (±.06)	.74 (±.06)	$F(1, 83) = 1.47$, $p = .229$, $\eta_p^2 = .02$	$F(1, 83) = 5.38$, $p = .02$, $\eta_p^2 = .06$	$F(1, 83) = .37$, $p = .546$, $\eta_p^2 = .02$
Cross-mapped	.27 (±.08)	.32 (±.08)	.45 (±.09)	.61 (±.09)	$F(1, 83) = 1.47$, $p = .229$, $\eta_p^2 = .02$	$F(1, 83) = 7.65$, $p = .007$, $\eta_p^2 = .08$	$F(1, 83) = .374$, $p = .543$, $\eta_p^2 < .01$
Far transfer	.66 (±.05)	.63 (±.05)	.68 (±.04)	.69 (±.04)	$F(1, 83) = .01$, $p = .94$, $\eta_p^2 < .01$	$F(1, 83) = .86$, $p = .355$, $\eta_p^2 = .01$	$F(1, 83) = .25$, $p = .617$, $\eta_p^2 < .01$

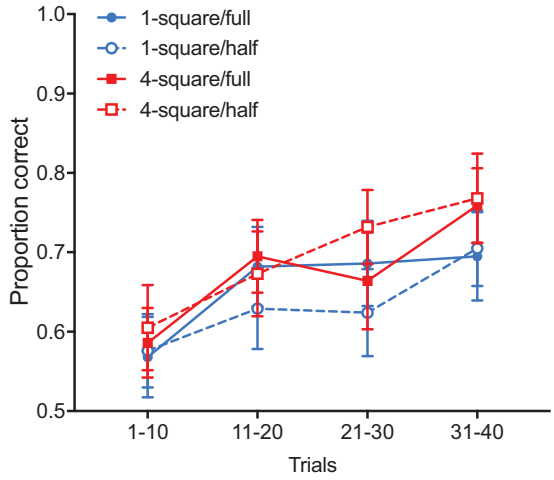

Figure 5. Far Transfer results for Experiment 3 across successive blocks of 10 trials. Error bars show standard error of the mean. See the online article for the color version of this figure.

Accuracy in the far transfer phase for each group is shown in Figure 5. Accuracy increased over the 40 trials, $F(3, 249) = 13.71$, $p < .001$, $\eta_p^2 = .142$, but there were no effects of the task manipulations, highest $F < 1$.

Individual differences.

Learning capacity. Replicating Experiment 2, RPM significantly correlated with accuracy during learning $r(85) = .26$, $p = .01$, but did not significantly correlate with relation learning $r(85) = .16$, $p = .136$, or learning strategy, $r(85) = .09$, $p = .39$.

Learning strategy. Overall, there was no significant correlation between learning strategy and relation learning $r(85) = .17$, $p = .125$. Following the analyses from Experiment 2, we examined whether this correlation differed between learning conditions. There was a significant correlation between learning strategy and relation learning in the 4-square condition, $r(42) = .41$, $p = .006$, but not in the 1-square condition, $r(41) = -.16$, $p = .298$, and these r values were significantly different from each other $z = 2.69$, $p = .004$. There were no significant correlations between learning strategy and relation learning in either the full-saturation condition, $r(42) = .04$, $p = .78$, or the half-saturation condition, $r(41) = .26$, $p = .087$, and these r-values were not different from each other, $z = 1.02$, $p = .308$. As such, the following analyses focused only on the difference between the line length conditions.

As in Experiment 2, we split subjects into the top 40% (rule-learners) and bottom 40% (exemplar-learners) for each condition, based on their reported learning strategy score, creating four groups of 14 subjects for analysis. Figure 6 shows the results for each of these groups.

There were significant interactions between line length and learning strategy for the feature, $F(1, 64) = 15.122$, $p < .001$, $\eta_p^2 = .191$, relation, $F(1, 64) = 12.80$, $p = .001$, $\eta_p^2 = .17$, and cross-mapped, $F(1, 64) = 10.61$, $p = .002$, $\eta_p^2 = .14$ trials. In each of these cases, there was an effect of line length for the rule-learners, but not for the exemplar-learners, highest $t(26) = -0.9$, $p = .376$, $d = -0.34$. On feature trials, rule-learners had better accuracy in the 1-square condition than the 4-square condition, $t(26) = 4.75$, $p < .001$, $d = 1.80$. On

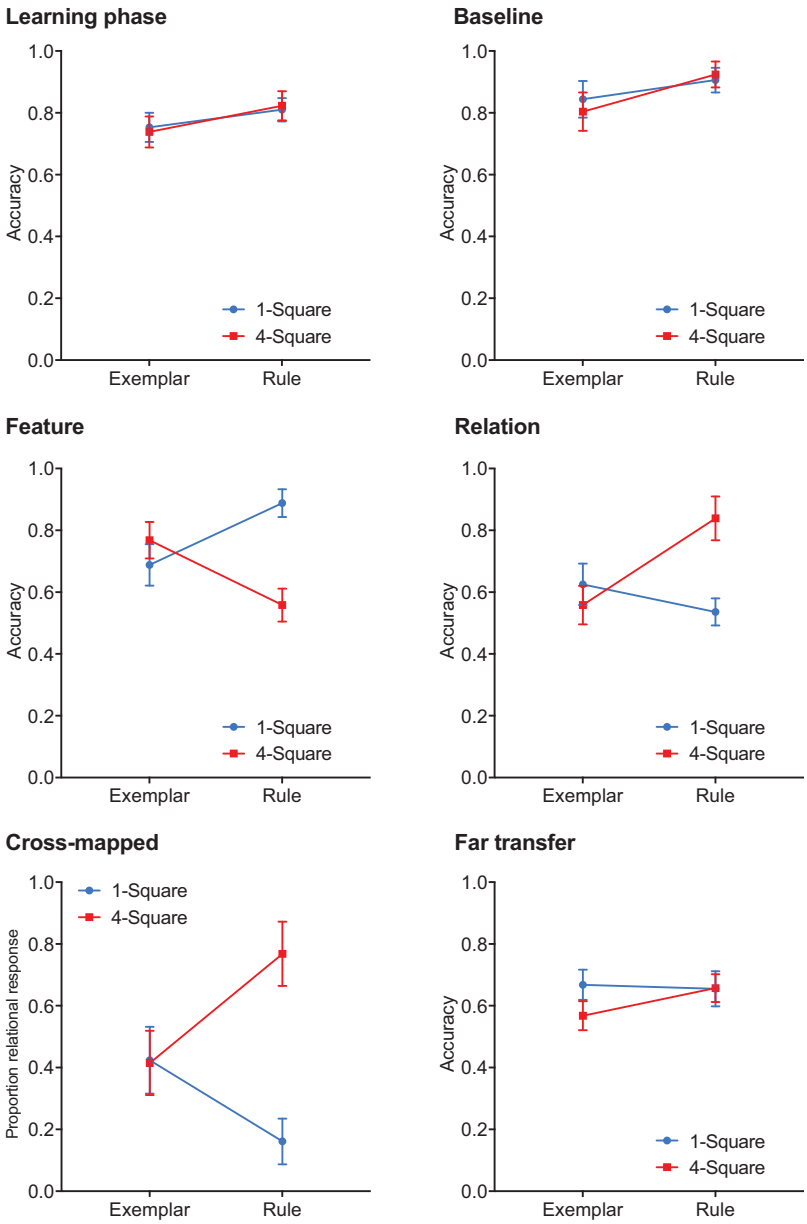

Figure 6. Results of Experiment 3 for exemplar- and rule-learners based on line length condition. See the online article for the color version of this figure.

relation trials, rule-learners had better accuracy in the 4-square condition than in the 1-square condition, $t(26) = -3.65$, $p = .001$, $d = -1.38$. Similarly, on the cross-mapped test-trials, rule-learners responded more relationally in the 4-square condition than in the 1-square condition, $t(32) = -4.76$, $p < .001$, $d = -1.80$. Despite the 4-square condition facilitating relational responding for rule-learners in the test phase, this did not occur in the far transfer phase, $F(1, 52) = 1.05$, $p = .31$, $\eta_p^2 = .02$.

Discussion

Unlike in Experiment 2, the independent variables significantly affected the overall amount of learning, as shown in the learning phase and for the baseline test trials. Specifically, the line length and color saturation interacted. On the whole, reducing the saturation of the color reduced feature learning, and this had a detrimental effect on learning overall for the 1-square condition, but not on the 4-square condition, as the relational rule was much easier to discover in the latter case to compensate.

The test phase showed an interesting asymmetry in the tradeoff between focusing on features versus relations. Increasing relational salience significantly increased relation learning, and correspondingly significantly decreased feature learning. On the other hand, manipulating the salience of the features only had a significant effect on feature learning, with no corresponding significant effect on relation learning.

Interesting asymmetrical patterns continued when considering the learning strategy of the subjects. The data suggested that the kind of categories that rule-learners develop greatly depends on what kind of distinction is salient in the stimuli, which is consistent with Little and McDaniel (2015b). When the relational rule was not salient, rule-learners were highly feature-oriented. When the relational rule was salient, rule-learners were highly relation-oriented. On the other hand, exemplar-learners did not seem to learn the relations very well regardless of either experimental manipulation. This pattern, along with Experiment 2, suggested that discovering relational rules requires both being rule oriented and that the task and/or stimuli highlight the relation.

Although both Experiments 2 and 3 showed the interactions between learning strategy and task structure that are an empirical goal of this paper, neither were suggestive of a learning intervention that can foster relational learning for exemplar-learners. Thus Experiment 4 sought a means by which even learners not predisposed to rule learning could learn relational rules more effectively. This would be a more powerful finding, and one that would, in many ways, be more practically important.

Experiment 4

In an attempt to orient all learners toward relational information, we introduced an additional category learning procedure: inference-learning (see Markman & Ross, 2003 for review) wherein subjects, given the category label and an incomplete exemplar, learn how to infer the missing parts to the exemplar. Much of the previous work on inference-learning (e.g., Sakamoto & Love, 2010) has concerned how inference-learning benefits learning the *statistical* relations among an exemplar's features, that is, how features are correlated across exemplars. Statistical relations are distinct from those typically used in relational categorization (e.g., the current work, Doumas, & Hummel, 2013) because learning feature-correlations does not require explicitly binding distinct representational elements into a relational structure. The distinction can be captured, for instance, by comparing a model of semantic memory that can use distributed representations to capture feature correlations (McRae, Cree, Westmacott, & De Sa's, 1999), and one that binds symbolic representations on top of a distributed semantic memory (Hummel & Holyoak, 1997, 2003). Closer to the current work, Erickson et al. (2005; and see Higgins, 2017) showed that compared with typical classification learning, inference-learning improved learning of abstract coherent categories. In these categories, exemplars share no features in common but the feature-sets contained within each exemplar cohere by the same underlying relation.

Here, we looked to extend that improvement in relational learning to our paradigm, while also examining whether that improvement is specific to participants with a particular learning strategy or whether it applies to all kinds of learners. In addition, we wanted to increase the external validity of our learning strategy analyses by assessing learning strategy via the MSLQ, which is a survey frequently used in education research (see, e.g., Pintrich & Degroot, 1990) wherein students report on the learning strategies they use in their formal education generally. We included items from all constructs, however we were interested specifically in the items that concern an *elaboration* learning strategy. Pintrich, Smith, Garcia, and McKeachie (1991, p. 20) note, "Elaboration strategies help students store information into long-term memory by building internal connections between items to be learned." This construct more than the others, such as effort-regulation, test anxiety, and learning from peers, seems specifically related to relational learning.

Method

Subjects. Eighty-four subjects took part in the experiment. Sixty were first-year psychology students from the University of Sydney rewarded with course credit for their participation, and a further 24 participants were recruited on campus and compensated with 15AUD for their participation. Three participants were excluded based on scores of below 25% accurate on the RPM. This left 81 participants (58 females; mean age = 20.4 years), randomly allocated to the classification condition (n = 41) and the inference-learning condition (n = 40).

Materials. In the primary category-learning phase, the *classification* task presented exemplars in similar fashion to Experiments 1 and 3, with a 50/50 probability of repeating the same category or alternating to an exemplar of the other category. The colors of the line stimuli were full saturation and the minimum line length difference was two squares. The *inference* task presented incomplete arrays in which one line of squares was missing (the position of the missing line was marked by a thin gray rectangular outline). The display was labeled as either a "Snarg" or a "Blicket," and subjects were presented with two options of lines that they could select to complete the pattern (see Figure 7). After they selected the line, the selected line appeared in the array, filling in the blank space. As with the classification conditions throughout the four experiments, the subjects had an opportunity to change their answer before moving on to the next trial. This allowed the subjects to see what the completed array looked like with both options. The correct third line option matched both the color distribution and relational rule of the presented category, whereas the incorrect option mismatched both the color distribution and the relational rule, and so in no way intrinsically biased the focus on one or the other. The classification and inference conditions had identical constraints for trial sequencing and generating individual exemplars. For instance, every pattern

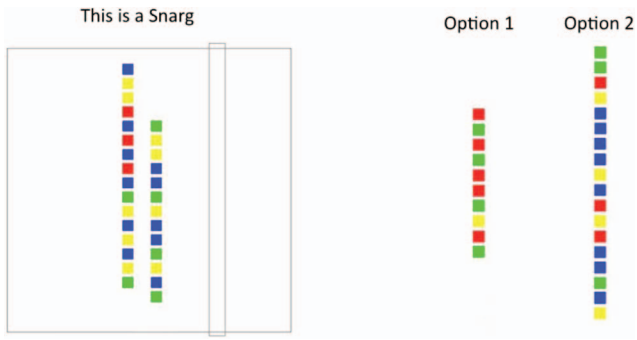

Figure 7. Example of a typical display for the Inference-learning condition from Experiment 4. See the online article for the color version of this figure.

was generated with lawful correct and incorrect choices for the third line, even though the incorrect choice was only shown for the inference condition. After the learning phase, the task continued in an identical manner for both conditions, with all subjects completing classification trials during the test and far transfer phases.

Subjects then completed an abbreviated version of the RPM. To assess learning strategy, subjects completed 40 items from the MSLQ (Pintrich et al., 1991; see the Appendix). For all items, subjects indicated to what degree the statement applied to them on a 1–7 Likert scale. The elaboration items were: (item numbers from the full MSLQ):

62. I try to relate ideas in one subject to those in other subjects whenever possible.

64. When reading for a class, I try to relate the material to what I already know.

67. When I study for a course, I write brief summaries of the main ideas from the readings and the concepts from the lectures.

69. I try to understand the material in a class by making connections between the readings and the concepts from the lectures.

Procedure. The primary task and RPM were run on Mac mini desktop computers at individual workstations, and then the MSLQ was filled out by pen and paper.

Results

Categorization task. Group means and statistical analyses of group effects are shown in Table 4. Accuracy in the learning phase was generally lower for the inference condition compared to the classification condition. This disadvantage for inference-learning persisted into the test phase, as seen on the baseline test trials. The inference condition also had poorer accuracy on the feature trials.

Although there was no effect of condition on the relation test trials, the inference condition responded significantly more relationally than the classification condition on the cross-mapped trials. Moreover, relational responding was well above chance levels in this condition, $t(39) = 3.83$, $p < .001$, $d = 1.24$, indicating the relational advantage was more than a lack of feature learning. Critically, there was a significant benefit in far transfer performance for the inference condition compared to the classification condition (see Figure 8). Accuracy increased over the far transfer phase overall, $F(3, 237) = 5.68$, $p = .001$, $\eta_p^2 = .067$, but did so at a greater rate for the inference condition, $F(3, 237) = 2.69$, $p = .047$, $\eta_p^2 = .033$. Within the inference condition, far transfer performance was significantly correlated with accuracy on cross-mapped test trials, $r(38) = .49$, $p = .001$, further suggesting that the effect of condition on far transfer performance was based on the inference group's superior learning of the relation.

Individual differences.

Learning capacity. In contrast to the previous experiments, RPM did not predict accuracy throughout the learning phase, $r(79) = -.03$, $p = .802$, and this lack of significant relationship held for both the classification, $r(39) = .16$, $p = .314$, and inference, $r(38) = -.25$, $p = .116$, conditions considered alone. In addition, RPM did not predict relation test trial performance $r(79) = .13$, $p = .237$, nor the degree to which subjects reported elaboration strategies, $r(79) = -.06$, $p = .602$.

Learning strategy. Across both conditions, self-reported elaboration strategies did not predict relation test trial performance, $r(79) = .08$, $p = .505$, nor did it significantly predict relation test trial performance for the classification, $r(39) = .25$, $p = .112$, or inference, $r(38) = .01$, $p = .963$ conditions alone.

To be consistent with the previous experiments, we further examined the role of learning strategy on performance. Similar to E2 and E3, we divided subjects into the top 40% and bottom 40% of self-reported elaborators in each condition creating four groups of 16 subjects each. To get an overall sense of the pattern, see Figure 9, which shows that globally for low-elaborators, there was a large effect of task, wherein the inference condition elicited an increased relational focus compared to the classification condition, but for high-elaborators there was little effect of task-condition as they showed a relational focus in both the classification and inference conditions.

There was an interaction between learning strategy and task-condition for the feature test trials, $F(1, 60) = 8.18$, $p = .006$, $\eta_p^2 = .12$, because classification led to greater accuracy for the low-elaborators, $t(30) = 5.45$, $p < .001$, $d = 1.93$, although there was no effect of condition for the high-elaborators, $t(30) = 1.33$, $p = .192$, $d = 0.47$. There was no interaction on the relation test trials, $F < 1$, but there was an interaction on the cross-mapped trials, $F(1,$

Table 4
Results From the Learning, Test, and Far Transfer Phase of Experiment 4

Phase	Means		Group effects
	Classification	Inference	
Learning	**.83 (±.02)**	**.72 (±.03)**	$t(79) = 3.35$, $p = .001$, $d = .74$
Test			
Baseline	**.88 (±.03)**	**.77 (±.04)**	$t(79) = 2.30$, $p = .024$, $d = .51$
Feature	**.74 (±.04)**	**.53 (±.03)**	$t(79) = 4.56$, $p < .001$, $d = 1.02$
Relation	.64 (±.03)	.72 (±.04)	$t(79) = -1.43$, $p = .157$, $d = .32$
Cross	**.39 (±.06)**	**.69 (±.05)**	$t(79) = -3.66$, $p < .001$, $d = .82$
Far transfer	**.64 (±.03)**	**.74 (±.03)**	$t(79) = -2.23$, $p = .028$, $d = .50$

Note. Standard error of the mean (SEM) shown in parentheses. Cross-mapped trial means represent the proportion of relationally consistent responses. All other means represent accuracy. Significant results are highlighted in bold.

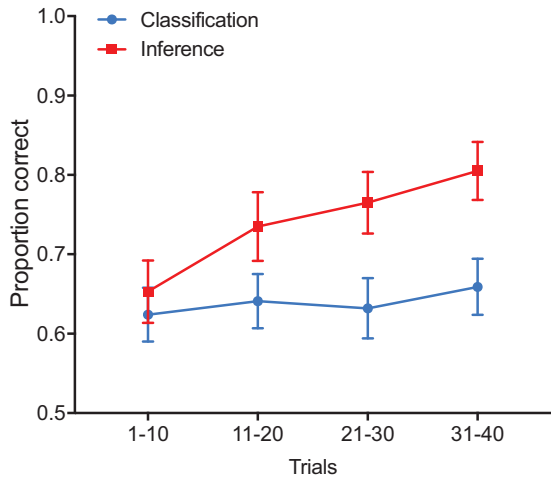

Figure 8. Far Transfer results for Experiment 4 across successive blocks of 10 trials. Error bars show standard error of the mean. See the online article for the color version of this figure.

60) = 7.41, $p = .008$, $\eta_p^2 = .11$, as low-elaborators in the inference condition performed more relationally than low-elaborators in the classification condition, $t(30) = -4.51$, $p < .001$, $d = -1.59$, while there was no effect of condition for high-elaborators, $t(30) = -0.37$, $p = .711$, $d = -0.13$. Although numerically this pattern was similar in the far transfer phase, the interaction did not reach significance, $F(1, 60) = 2.146$, $p = .15$, $\eta_p^2 = .03$.

Discussion

With the results of Experiment 4, all of our primary empirical aims were achieved. We identified learning conditions (specifically, the inference-learning task) that boost learning the relations as shown through a strong relational preference in the cross-mapped test trials and, for the first time in this study, clear evidence of facilitation on far transfer trials. Interestingly, this benefit on cross-mapped and far transfer does not reflect overall general advantages for the inference condition. This is a distinct pattern from Sakamoto and Love (2010) which saw advantages for inference-learning both specifically in learning the statistical relations among the exemplars' features, and increased accuracy overall. Indeed, learners in the inference condition faired worse during training, and on baseline test trials. This suggests that a task that induces a relational focus can come at a cost to some measures of overall learning. We will return to this in the General Discussion.

Further, the inference-learning task benefitted learners with self-reported suboptimal learning strategies. Looking at Figure 9, the low- and high-elaborators have distinct patterns in the classification condition (with the low-elaborators having a greater featural focus), while both high- and low-elaborators show a relational focus in the inference condition. These results build on Erickson et al. (2005) by showing that inference-learning aids relational discovery even for learners who otherwise do not engage in activities that promote relational learning more generally.

General Discussion

This paper has introduced a novel category-learning paradigm that could be solved by focusing either on perceptual features or structural relations. The four experiments then systematically examined variations in task parameters that led to a focus on either the features or the relations. Experiments 2–4 also examined differences in learning strategy and cognitive ability and how these interacted with task structure. There were several key findings. The first is one that we have not yet emphasized in reporting the results of the individual experiments. Across all four experiments, despite the logical possibility, no learner excelled at learning both the featural and relational aspects of the categories; not a single participant of the 308 who were analyzed scored greater than 80% accuracy on both feature and relation test trials (whereas 243 scored greater than 80% on at least one of these test trials). Second, Experiments 2–4 all showed interactions between learning strategy and the task structure manipulations in producing superior featural or relational learning. In classification learning, Experiments 2 and 3 showed that both a rule-learning strategy, and a task structure that highlighted the relations were necessary to discover the relational rule. In Experiment 4, we demonstrated that inference-learning improves relational discovery and transfer even for those with suboptimal learning strategies.

Additionally, in Experiments 1–3, conditions favoring relational learning during the earlier parts of the experiment showed little continued advantage for the far transfer phase. However, In Experiment 4, inference-learning produced superior far transfer performance. It seems that inference-learning better prepared the subjects to learn from the feedback in this later phase (see Bransford & Schwartz, 1999). It is an open empirical question whether inference-learners would be able to apply their relational knowledge to far transfer trials without any feedback.

Limitations Attributable to Stimulus Design

In addition to the open question about the role of feedback in transfer, additional potential limitations of the current study are rooted in the nature of the stimuli. Our goals in stimulus design were to create entirely novel stimuli that reflected the real-world relationship between structural relations and surface features in natural relational categories. In this design, however, the actual relations of line-length, and the color features of those lines do not closely connect to any specific natural relational category in the way for example, the categories of Yamauchi and Markman (1998) can reasonably be described as novel cartoon bugs. In this sense, the current stimuli are actually more similar to fluid intelligence items such as in RPM than any natural category. Yet, our individual differences analyses, wherein RPM scores did not predict relational learning, and yet self-reported real-world learning strategies (from the MSLQ) did predict relational learning suggests that our stimuli did capture something beyond the specific of these stimuli. Of course, it will be important for future research to extend these findings to natural, or at least naturalistic relational categories.

An additional potential limitation of the design is rooted in the relationship between the two categories' relational rules. That is, the two relational rules are defined in contrast to one another (see Goldstone, 1996). Conceptual interrelatedness is quite common in relational categories, and Goldwater and Schalk (2016) discuss how the interconnections among relational concepts can make them a more useful set of tools for reasoning. However, it is unclear how much our results will generalize to sets of relational categories that do not have these clear interrelations. Explicitly varying the level of interrelated-

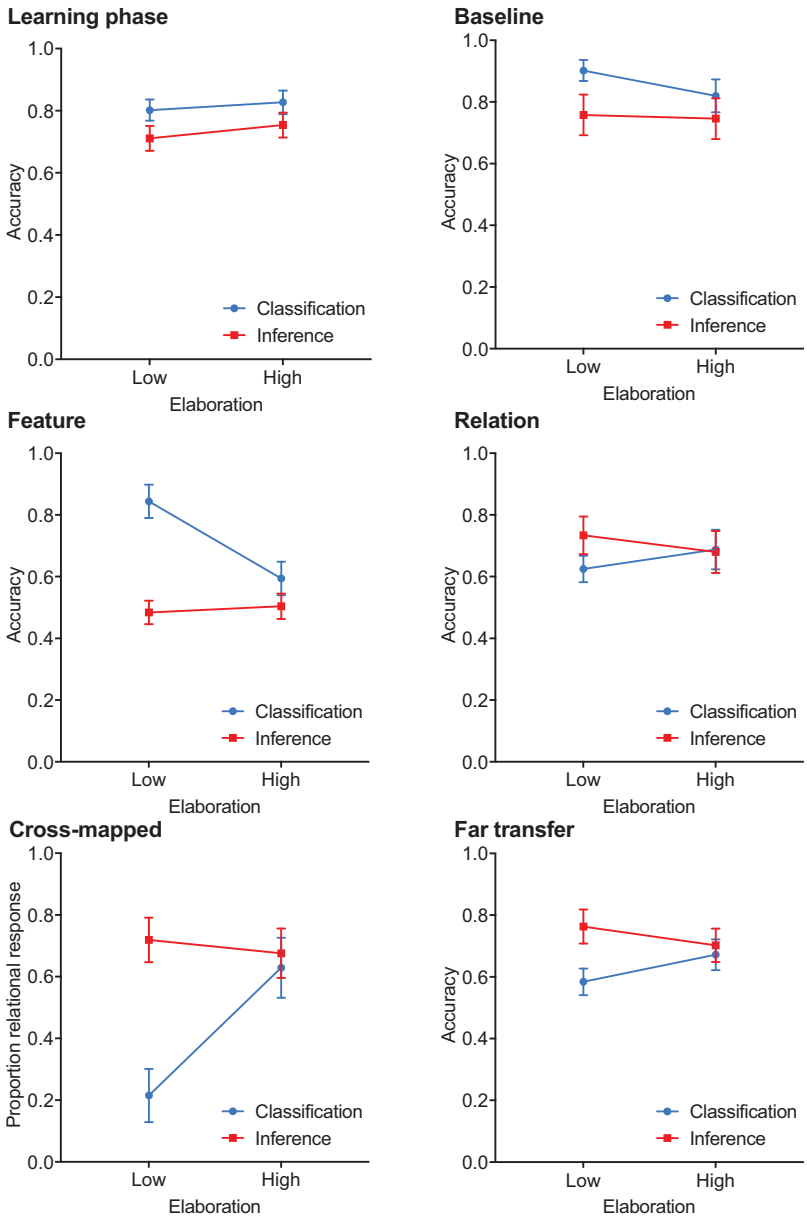

Figure 9. Experiment 4 results for low- and high- elaborators. Cross-mapped results are plotted as the proportion of relationally consistent responding. Baseline, Feature, and Relation trials report accuracy. Error bars show standard error of the mean. See the online article for the color version of this figure.

ness among relational categories being learned in parallel is the topic of ongoing research (and see Jung & Hummel, 2015 for evidence that such interrelations are critical for learning).

Implications for Category Learning Research

For several decades, the primary goal of the category learning literature has been to uncover category representations (e.g., whether categories were represented by prototypes or exemplars, Smith & Minda, 1998) and to characterize the learning processes (e.g., whether there are one or two kinds of learning mechanisms, see Ashby & Maddox, 2005), and so manipulations of stimulus or task structure were targeted toward these theoretical discoveries (e.g., to contrast linearly or nonlinearly separable categories, Medin & Schaffer, 1978). The current research joins a growing literature on categorization that experimentally manipulates task structure with the clear goal of identifying ways to improve learning outcomes, while uncovering learning processes along the way (see Bjork et al., 2013; Goldwater & Schalk, 2016 for recent reviews). This recent literature has identified two clear recommendations we focus on here: (1) Task structures that foster comparison of exemplars are critical for generalization (e.g., Kurtz et al., 2013; Rohrer & Pashler, 2010); and (2) often greater difficulties at learning can produce greater benefits in the long-term,

called "desirable difficulties" (e.g., Bjork & Bjork, 2011, and see Kapur, 2008 for a related perspective from educational research).

The current experiments were consistent with both of these recommendations, and add to our understanding of them. With regard to how best to promote comparisons, selecting the optimal trial sequence may be more complex than some have argued (e.g., Bjork et al., 2013). Promoting between-category comparisons via interleaving seems optimal when the primary challenge is discriminating among categories (Rohrer & Pashler, 2010), but promoting within-category comparisons via blocking (or other means, see Carvalho & Goldstone, 2015b; Rawson et al., 2015) seems critical when learners need to discover within-category structure (Carvalho & Goldstone, 2014; Higgins & Ross, 2011; but also see Rohrer, Dedrick, & Burgess, 2014). Experiment 2 suggested that blocking can improve relational discovery, particularly for people searching for rules. However, the stimuli used in Experiment 2 may have made the task of discovering the relation relatively difficult. In contrast, Experiment 3 directly varied the relational salience and found strong effects on relation and feature learning. Further, there is evidence from young children's word-learning that an optimal learning sequence first supports within-category comparisons to discover the structure, and then between-category comparisons to refine that understanding (Namy & Clepper, 2010).

Expanding on Bjork et al. (2013), Experiment 4 suggested that learning via inference is an additional desirable difficulty: Inference-learning elicited lower accuracy throughout the learning phase, but greater far transfer. An interesting question for further research is whether the effectiveness of inference-learning would change with the trial sequence. Each trial on its own elicits focus on within-category relations, so perhaps an interleaving structure would maximize both within and between-category learning. However, it is also possible that interleaving interferes with local attempts to discover the structure of any given category and, if so, blocking would maximize the potential of the inference-learning procedure. Further, because inference-learning requires using the category label, it would intrinsically support another benefit of interleaved problem-solving practice: the required need to classify a problem before solving it (Rohrer et al., 2014).[7]

In addition to contributing to our understanding of how to improve learning outcomes, the current experiments challenge current category learning models. First, the majority of existing category learning models do not explicitly represent relational structure. That is, they do not represent objects bound by how they relate, with objects and relations as independent representational elements fostering composition into a variety of complex structures (as in Falkenhainer, Forbus, & Gentner, 1989; Hummel & Holyoak, 2003). Typically categorization models represent categories as collections of features (e.g., Tversky, 1977), or as points or vectors in a high-dimensional space (e.g., Nosofsky, 1984). Spatial models are typically applied to feature-based categories, although, as Davis, Goldwater, and Giron (2017) have recently shown, they can be extended to relational categorization as well. That is, just like different features, different relations can be represented as different points or vectors in a space. However, although spatial models can *accommodate* relational categorization data, they cannot truly *account* for relational discovery and learning. For example, if the two kinds of information (features and relations) are represented using the same spatial format, what a priori basis would there be to predict the many differences (e.g., between inference and classification training, blocked and interleaved sequences)? Models of analogy make it quite clear that without the explicit binding of representation elements into a structured representation (i.e., the part of the process that spatial models take for granted), much about relational cognition cannot be captured (see Doumas et al., 2008; Falkenhainer et al., 1989; Forbus, Gentner, & Law, 1995; Hummel & Holyoak, 1997, 2003 among others).

As one reviewer noted, the learning of the color proportions associated with the two categories could be based on relations rather than features. After all, the *relative* frequency of red and green versus yellow and blue serves as a predictive relation of the categories (for instance Blickets might be "redder" than Snargs, and Blickets might be more red than they are yellow). Indeed virtually any feature-based property can be described in relational terms. Importantly, although these properties of the categories can be expressed in relational terms, they are quite readily learned by explicitly nonrelational, feature-based models. For instance, associative networks can learn categorization problems based on the relative frequency of different features even if the numerosity of those features is not explicitly coded into the inputs of the network (e.g., Livesey & McLaren, in press). In comparison with models of relational cognition and analogy, models focusing on feature learning and comparison are algorithmically much simpler (e.g., contrast Tversky, 1977 with Falkenhainer et al., 1989). In the current work, a simple feature-based solution is analogous to what one would expect from applying logistic regression equation where the four predictors are the proportion of each of the four colors. Learning could then be characterized as learning the beta weights of blue and green of one sign and on red and yellow of the opposite sign. In practice, error-correction learning algorithms operate by the same principles and are typically applied to features rather than relations. That is, even learning based the relative values of features can be accounted for without true relational structures and structured-comparisons.

Second, although most categorization models do not explicitly represent relations, it is unclear how category learning models that do (e.g.,Goodman et al., 2008; Mclure, Friedman, & Forbus, 2010; Doumas et al., 2008; Corral & Jones, 2014; Tomlinson & Love, 2006) would account for effects of trial-sequencing, and inference-learning versus classification. The model developed by Love, Medin, and Gureckis (2004) is able to capture some key distinctions between inference and classification because its error-based learning mechanism fosters flexible representation formation fitting to the particular constraints of the learning task. We suspect that integrating these flexible learning mechanisms with the machinery to represent structured relations will be critical to capture the pattern demonstrated in this paper and beyond.

This modeling work will further help to unify theories of analogical reasoning and category learning, which we view as critical for a full account of relational transfer (and lack thereof). We framed this paper at the outset as building on work on the inert knowledge problem that

[7] Rohrer's and colleagues' work has shown that one of the primary hurdles to math problem solving is the proper classification of an exemplar problem as a member of a problem category. Blocking problems allows for the same procedure to be used consecutively without consideration of the problem category for each solution, while interleaving requires problem classification in order for the right procedure to be selected, strengthening exemplar-category ties. Likewise, inference-learning also intrinsically supports exemplar-category ties as the category label is used to predict missing aspects of exemplars.

focuses on the retrieval of individual cases by instead focusing on the generalization of relational category knowledge. However, analogical knowledge transfer from a single case and from a category most likely lie on a continuum, especially given how crucial retrieval of specific exemplars can be in initial category formation (Ross, Perkins, & Tenpenny, 1990). Future modeling and empirical work should focus on how relational knowledge transfer changes as learners accrue experiences with an increasing number relational category exemplars.

The Role of Selective Attention

Central to many models and theories of categorization is the notion of selective attention: that over the course of learning, attention is shifted toward the most relevant dimensions of the stimuli to perform a task (e.g., attending to the dimension most diagnostic of category membership fostering improved classification accuracy). Designs that spatially separate exemplar features have allowed eye-tracking analysis as a direct measure of attention, which have empirically confirmed the importance of attention as posited in formal models such as the GCM (Hoffman & Rehder, 2010; Rehder, Colner, & Hoffman, 2009, and see Blair, Watson, Walshe, & Maj, 2009). What was the role of selective attention in the current findings? The fact that no learner approached ceiling for both the features and the relations suggests that learning either one required selectively attending to that kind of information at the cost of the other. Indeed, Tomlinson, and Love (2006) simulate relational learning as a shift in attention from featural to relational aspects of a stimulus.

In Experiment 1, we explicitly hinted at what aspects of the stimuli to attend to. The relational hint did not mention the relation itself, but the dimension of the stimulus (line length) that the relation operated on. In Experiment 2, blocking elicited attention to the relation because comparing stimuli that share relational structure shifts attention toward those common relations (Gentner, 2010; Markman & Gentner, 1993). In Experiment 3, increasing the line length differences also seemed to elicit attention to that relevant dimension. In Experiment 4, (and in Erickson et al., 2005; Higgins, 2017) inference-learning supported relational learning because predicting missing pieces of exemplars is helped by focusing on how the pieces relate to one another.

Perhaps selective attention explains why feedback was so useful in the Far Transfer phase when similar research (e.g., Kurtz et al., 2013) has not required it. That is, if successful relational learning entailed shifting attention away from the featural dimension of color, then this may have carried into the far transfer phase and made applying the relational rule to the luminance of the gray dots quite counterintuitive. This opens the door for future research examining the potential interaction between attention-shifting during learning and the role of feedback during transfer.

In addition to explaining learning and transfer effects, differences in learning strategy can potentially be reframed as differences in what is attended. That is, "being rule-oriented" could actually be a propensity for attending to relations among objects and events, thus comparing exemplars to find the key within-category commonalities and between-category differences. This is effectively the result of explicitly directing category learners to compare exemplars (Higgins, 2017; Spalding & Ross, 1994). As suggested by an anonymous reviewer, perhaps the exemplar strategy, despite the wording of the self-report measure, does not really reflect an attempt to memorize, but reflects a more diffuse attentional style than rule-learners who are actively trying to discover what to pay the most attention to.

Although selective attention is clearly important to relational learning, it is not sufficient in all cases. Goldwater and Gentner (2015) analyzed self-reported sorting strategies of descriptions of causal phenomena. Even subjects who said they were attempting to sort by their underlying causes (i.e., they were already attending to that dimension of the stimuli) benefitted from an analogical comparison exercise to gain a sharper understanding of what the causal structures were. Likewise, while the hint of Experiment 1 pointed learners toward the relevant dimension, further alignment of exemplars was important to abstract the common relations, allowing for accurate classification. Understanding the role of selective attention in relational learning, and in characterizing differences in learning strategy is an important area of future research.

Implications for Individual Differences Research

The current experiments advanced recent trends on individual differences in category learning. There are mixed results in the literature in regards to the effects of cognitive capacity and learning strategy. There are some situations where capacity is predictive regardless of strategy (Craig & Lewandowsky, 2012), others where strategy is critical, but strategy itself is predicted by capacity (McDaniel et al., 2014), and others where strategy is critical, but strategy is not predicted by capacity (Little & McDaniel, 2015b). In our experiments, cognitive capacity (as measured by RPM) was generally predictive of overall rates of learning, but did not discriminate learning strategy, and thus was not predictive of whether the learner focused on relational or featural information.[8] In contrast, learning strategy predicted the propensity to discover and transfer the relational properties of the task. Thus, our results are most consistent with (Little & McDaniel, 2015b) in that learning strategy was highly predictive of what was learned and not related to learning capacity. However, we note that our sample were undergraduates, and thus are the high-end of cognitive capacity for the population as a whole. We suspect that across the range of cognitive abilities present in the entire population, there will be more clear relationships between capacity and strategy.

The current results build on (Little & McDaniel, 2015b) in two crucial ways. The first is that learning strategy consistently interacted with the task structure in producing learning outcomes. That is, the experimental manipulations had large effects for some subjects, but little effect for other subjects. For classification learning, stimuli and trial sequences mattered little to exemplar learners, while they drastically affected performance in rule learners. Inference-learning however, elicited a relational focus equally for all learners. One implication is that classification is a more open-ended task, one wherein a learner can impose their preferred strategy. Inference-learning on the other hand seems more constrained, eliciting relational learning more broadly, despite the features still being available as the basis for accurate responding.

The second key finding is that a laboratory based learning task, and a survey to elicit self-report of study strategies more generally, predicted learning outcomes in our primary task in similar manner. This helps to both ground the self-report educational measures in highly constrained cognitive tasks, and give evidence to the external-validity

[8] Given that O-span showed no predictive relationship with any measure, our results depart from those of Craig and Lewandowsky (2012), but again, with limited data, this pattern is hard to interpret.

of these highly constrained tasks. Likewise, McDaniel et al. (2014) refer to findings that their laboratory measures of learning strategy predict students' performance in an undergraduate chemistry course. Perhaps it is unsurprising that students approach learning in their course in a similar manner to how they engage in laboratory-based cognitive tasks. In these experiments, as is quite common in psychological research, most participated for course-credit. Why would they not treat these learning tasks similarly to their course-content? Further, many students probably see some of the content of their courses as similarly arbitrary (. . . or hopefully, similarly engaging).

However, despite these connections between task-based and self-report measures, an important future direction is to investigate to whether learners flexibly deploy different strategies in different contexts or are rather consistent in their strategies. McDaniel et al.'s (2014) findings were novel in that it showed an underlying consistency in learners' strategies across superficially disparate tasks, but educational researchers investigating deep versus surface strategies have assumed learners vary in their strategies based on their interest in a topic or motivation in any given moment (e.g., Pintrich, Marx, & Boyle, 1993). On the other hand, research on students' folk epistemologies, and naïve theories of how learning works predict the way they approach their education and learning tasks, perhaps suggesting more consistency across contexts (e.g., Schommer, 1990; Wegner & Nuckles, 2016). Again, more research is needed.

Implications for Educational Research

One of the primary goals of the current research was to serve as a laboratory model for one of the most challenging problems in education: the abstraction and transfer of relational knowledge. To this end, our research makes two clear recommendations for avenues of research with real-world educational materials and environments. The first is to explore these interactions between learning strategy and task structure. For a long time, there have been general intuitions that task structure would interact with individual differences, and yet evidence for any such interactions was scant (Pashler, McDaniel, Rohrer, & Bjork, 2008; though see, e.g., Belenky & Nokes-Malach, 2012; Fuchs et al., 2014; Kalyuga, 2007 for recent demonstrations). However, perhaps this was largely attributable to assessments of the wrong kind of individual difference, as many have searched for differences in learners based on preferred modality of receiving instructional content (i.e., the construct that some students are "verbal learners" whereas others are "visual learners."). Although we find consistent interactions between learning strategy and task structure interactions, our results lead to the same recommendation as Pashler et al. (2008), which is to identify learning tasks that benefit all learners. That is, we did not find any manipulation of the basic form of the task that benefits one group of learners, but hurts the other. In Experiment 2, blocking helped rule-learners, but had no negative effect on exemplar-learners. In Experiment 4, inference-learning helped low-elaborators, but had no negative effect on high-elaborators. Taking the series of experiments as a whole, the results clearly suggest inference-learning as a method worth further exploring for its ability to foster relational learning and far transfer for learners with either strategy as a predisposition.

Of course real-world learning is far more complex than the tasks in our experiments. However, given the success of inference-learning in promoting biological category learning in primary schoolchildren (Sakamoto & Love, 2010), and the growing evidence that laboratory-based category learning methods generalize to educational settings (e.g., Bjork et al., 2013; Goldwater & Schalk, 2016), there are reasons to be confident. Still, beyond the complexity of real-world relational categories and learning environments, a perhaps bigger challenge to learning is the conflicting conceptual understanding students bring to bear to many educational contexts, especially in science (e.g., Chi, Roscoe, Slotta, Roy, & Chase, 2012; Jacobson, Kapur, So, & Lee, 2010). Goldwater and Schalk (2016) argue that a crucially underexplored research area is building laboratory-models of conceptual change by implementing multiphase category learning experiments where there are conflicts between earlier and later phases (see Ramsburg & Ohlsson, 2016; Sewell & Lewandowsky, 2011 for work along these lines), and then applying such models to the classroom. Real-world concepts are interrelated, forming conceptual systems. For example, concepts in physics such as "force," "momentum," "acceleration," and "work" are in part defined in relation to one another, and with common subcomponents (such as distance, mass, and time). An additional important challenge for lab-based learning experiments is to build on these methods to examine the learning of conceptual systems.

Conclusion

Our aim in this paper was twofold: to understand basic processes in relational category learning, and to identify avenues for research into educational interventions to improve relational transfer of knowledge. We are confident that the two goals go hand-in-hand because of the commonalities in the learning patterns inside and outside of the classroom. However, as we have mentioned, there is a clear difference between learning the relational concepts that native speakers of any language use readily, and classroom-based learning wherein many students fail to ever achieve an understanding supporting transfer. Ultimately, as children we all needed to abstract everyday relational concepts to properly communicate with and function among our peers because of their daily relevance. It is unclear whether students see the same need for their formally taught concepts. Further, there are only so many exemplars we can possibly ask students to process. We hope that our research helps the design of learning tasks that are maximally efficient: that is, to support the most abstract understanding given the limited time-on-task.

References

Alexander, P., Peterson, E. G., Dumas, D., & Hattan, C. (2016). A retrospective and prospective examination of cognitive strategies and academic development: Where have we come in twenty-five years? In A. O'Donnell (Ed.), *Handbook of educational psychology*. New York, NY: Oxford University Press.

Alvarez, G. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15,* 122–131.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149–178. http://dx.doi.org/10.1146/annurev.psych.56.091103.070217

Bassok, M., Chase, V. M., & Martin, S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology, 35,* 99–134. http://dx.doi.org/10.1006/cogp.1998.0675

Belenky, D. M., & Nokes-Malach, T. J. (2012). Motivation and transfer: The role of mastery-approach goals in preparation for future learning. *Journal of the Learning Sciences, 21,* 399–432. http://dx.doi.org/10.1080/10508406.2011.651232

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In

M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York, NY: Worth.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64,* 417–444. http://dx.doi.org/10.1146/annurev-psych-113011-143823

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1196–1206. http://dx.doi.org/10.1037/a0016272

Blessing, S. B., & Ross, B. H. (1996). Content effects in problem categorization and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 792–810. http://dx.doi.org/10.1037/0278-7393.22.3.792

Braithwaite, D. W., & Goldstone, R. L. (2015). Effects of variation and prior knowledge on abstract concept learning. *Cognition and Instruction, 33,* 226–256. http://dx.doi.org/10.1080/07370008.2015.1067215

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24,* 61–100.

Bulloch, M. J., & Opfer, J. E. (2009). What makes relational reasoning smart? Revisiting the perceptual-to-relational shift in the development of generalization. *Developmental Science, 12,* 114–122. http://dx.doi.org/10.1111/j.1467-7687.2008.00738.x

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,* 481–495. http://dx.doi.org/10.3758/s13421-013-0371-0

Carvalho, P. F., & Goldstone, R. L. (2015a). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology, 6,* 505. http://dx.doi.org/10.3389/fpsyg.2015.00505

Carvalho, P. F., & Goldstone, R. L. (2015b). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review, 22,* 281–288. http://dx.doi.org/10.3758/s13423-014-0676-4

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152. http://dx.doi.org/10.1207/s15516709cog0502_2

Chi, M. T., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive Science, 36*(1), 1–61.

Corral, D., & Jones, M. (2014). The effects of relational structure on analogical learning. *Cognition, 132,* 280–300. http://dx.doi.org/10.1016/j.cognition.2014.04.007

Craig, S., & Lewandowsky, S. (2012). Whichever way you choose to categorize, working memory helps you learn. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 65,* 439–464. http://dx.doi.org/10.1080/17470218.2011.608854

Davis, T., Goldwater, M., & Giron, J. (2017). From concrete examples to abstract relations: The rostrolateral prefrontal cortex integrates novel examples into relational categories. *Cerebral Cortex, 27,* 2652–2670. http://dx.doi.org/10.1093/cercor/bhw099

Don, H. J., Goldwater, M. B., Otto, A. R., & Livesey, E. J. (2016). Rule abstraction, model-based choice, and cognitive reflection. *Psychonomic Bulletin & Review, 23,* 1615–1623. http://dx.doi.org/10.3758/s13423-016-1012-y

Don, H. J., & Livesey, E. J. (2015). Resistance to instructed reversal of the learned predictiveness effect. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 68,* 1327–1347. http://dx.doi.org/10.1080/17470218.2014.979212

Doumas, L. A., & Hummel, J. E. (2013). Comparison and mapping facilitate relation discovery and predication. *PLoS ONE, 8,* e63889. http://dx.doi.org/10.1371/journal.pone.0063889

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review, 115,* 1–43. http://dx.doi.org/10.1037/0033-295X.115.1.1

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,* 86–99. http://dx.doi.org/10.1037/0278-7393.31.1.86

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

Ferry, A. L., Hespos, S. J., & Gentner, D. (2015). Prelinguistic relational concepts: Investigating analogical processing in infants. *Child Development, 86,* 1386–1405. http://dx.doi.org/10.1111/cdev.12381

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science, 19,* 141–205. http://dx.doi.org/10.1207/s15516709cog1902_1

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives, 19,* 25–42. http://dx.doi.org/10.1257/089533005775196732

Fuchs, L. S., Schumacher, R. F., Sterba, S. K., Long, J., Namkung, J., Malone, A., . . . Changas, P. (2014). Does working memory moderate the effects of fraction intervention? An aptitude–treatment interaction. *Journal of Educational Psychology, 106,* 499–514. http://dx.doi.org/10.1037/a0034341

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj (Ed.), *Language development: Language, thought and culture* (Vol. 2, pp. 301–334). Hillsdale, NJ: Erlbaum.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170. http://dx.doi.org/10.1207/s15516709cog0702_3

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

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34,* 752–775. http://dx.doi.org/10.1111/j.1551-6709.2010.01114.x

Gentner, D., & Boroditsky, L. (2001). Individuation, relational relativity and early word learning. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 215–256). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511620669.010

Gentner, D., Brem, S., Ferguson, R. W., Markman, A. B., Levidow, B. B., Wolff, P., & Forbus, K. D. (1997). Analogical reasoning and conceptual change: A case study of Johannes Kepler. *Journal of the Learning Sciences, 6,* 3–40. http://dx.doi.org/10.1207/s15327809jls0601_2

Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W.-K. Ahn (Ed.), *Categorization inside and outside the lab* (pp. 151–175). Washington, DC: American Psychological Association.

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

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development* (pp. 225–277). London, England: Cambridge University Press.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology, 25,* 524–575. http://dx.doi.org/10.1006/cogp.1993.1013

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition, 24,* 608–628. http://dx.doi.org/10.3758/BF03201087

Goldwater, M. B., & Gentner, D. (2015). On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. *Cognition, 137,* 137–153. http://dx.doi.org/10.1016/j.cognition.2014.12.001

Goldwater, M. B., Markman, A. B., & Stilwell, C. H. (2011). The empirical case for role-governed categories. *Cognition, 118,* 359–376. http://dx.doi.org/10.1016/j.cognition.2010.10.009

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin, 142,* 729–757. http://dx.doi.org/10.1037/bul0000043

Goodman, N. D., Tenenbaum, J. B., Griffiths, T. L., & Feldman, J. (2008). Compositionality in rational analysis: Grammar-based induction for concept learning. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for rational models of cognition*. Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199216093.003.0017

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences, 14,* 497–505. http://dx.doi.org/10.1016/j.tics.2010.08.005

Higgins, E. J. (2017). The complexities of learning categories through comparisons. *Psychology of Learning and Motivation, 66,* 43–77. http://dx.doi.org/10.1016/bs.plm.2016.11.002

Higgins, E. J., & Ross, B. H. (2011). Comparisons in category learning: How best to compare for what. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 1388–1393). Austin, TX: Cognitive Science Society, Austin.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General, 139,* 319–340.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104,* 427–466. http://dx.doi.org/10.1037/0033-295X.104.3.427

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220–264. http://dx.doi.org/10.1037/0033-295X.110.2.220

Jacobson, M. J., Kapur, M., So, H. J., & Lee, J. (2011). The ontologies of complexity and learning about complex systems. *Instructional Science, 39,* 763–783.

Jung, W., & Hummel, J. E. (2015). Making probabilistic relational categories learnable. *Cognitive Science, 39,* 1259–1291. http://dx.doi.org/10.1111/cogs.12199

Kalish, M. L., Lewandowsky, S., & Davies, M. (2005). Error-driven knowledge restructuring in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 846–861. http://dx.doi.org/10.1037/0278-7393.31.5.846

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19,* 509–539. http://dx.doi.org/10.1007/s10648-007-9054-3

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26,* 379–424. http://dx.doi.org/10.1080/07370000802212669

Keil, F. C., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning & Verbal Behavior, 23,* 221–236. http://dx.doi.org/10.1016/S0022-5371(84)90148-8

Kittur, A., Hummel, J. E., & Holyoak, K. J. (2004). Feature- vs. relation-defined categories: Probab(alistic)ly not the same. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the twenty-sixth annual conference of the cognitive science society* (pp. 696–701). Mahwah, NJ: Erlbaum.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36,* 1–16.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19,* 585–592. http://dx.doi.org/10.1111/j.1467-9280.2008.02127.x

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67,* 2797–2822. http://dx.doi.org/10.2307/1131753

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1083–1119. http://dx.doi.org/10.1037/0278-7393.25.5.1083

Kurtz, K. J., Boukrina, O., & Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1303–1310. http://dx.doi.org/10.1037/a0031847

Lee Cheong Lem, V. A., Harris, J. A., & Livesey, E. J. (2015). Testing the limits of the Perruchet effect in choice response time tasks. *Journal of Experimental Psychology: Animal Learning and Cognition, 41,* 385–394. http://dx.doi.org/10.1037/xan0000079

Lewandowsky, S., Oberauer, K., Yang, L. X., & Ecker, U. K. (2010). A working memory test battery for MATLAB. *Behavior Research Methods, 42,* 571–585. http://dx.doi.org/10.3758/BRM.42.2.571

Little, D. R., & Lewandowsky, S. (2009). Beyond nonutilization: Irrelevant cues can gate learning in probabilistic categorization. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 530–550. http://dx.doi.org/10.1037/0096-1523.35.2.530

Little, J. L., & McDaniel, M. A. (2015a). Some learners abstract, others memorize examples: Implications for education. *Translational Issues in Psychological Science, 1,* 158–169. http://dx.doi.org/10.1037/tps0000031

Little, J. L., & McDaniel, M. A. (2015b). Individual differences in category learning: Memorization versus rule abstraction. *Memory & Cognition, 43,* 283–297. http://dx.doi.org/10.3758/s13421-014-0475-1

Livesey, E. J., & Costa, D. S. J. (2014). Automaticity and conscious control in single and choice response time versions of the Perruchet effect. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 67,* 646–664. http://dx.doi.org/10.1080/17470218.2013.824014

Livesey, E. J., & McLaren, I. P. L. (2009). Discrimination and generalization along a simple dimension: Peak shift and rule-governed responding. *Journal of Experimental Psychology: Animal Behavior Processes, 35,* 554–565. http://dx.doi.org/10.1037/a0015524

Livesey, E. J. & McLaren, I. P. L. (in press). Revisiting peak shift on an artificial dimension: Effects of stimulus variability on generalization. *Quarterly Journal of Experimental Psychology*.

Livins, K. A., Spivey, M. J., & Doumas, L. A. (2015). Varying variation: The effects of within- versus across-feature differences on relational category learning. *Frontiers in Psychology, 6,* 129. http://dx.doi.org/10.3389/fpsyg.2015.00129

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111,* 309–332. http://dx.doi.org/10.1037/0033-295X.111.2.309

Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Erlbaum.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25,* 431–467. http://dx.doi.org/10.1006/cogp.1993.1011

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129,* 592–613. http://dx.doi.org/10.1037/0033-2909.129.4.592

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence, 13,* 329–358. http://dx.doi.org/10.1080/09528130110100252

Markman, A. B., & Wood, K. L. (Eds.). (2009). *Tools for innovation: The science behind the practical methods that drive new ideas*. New York,

NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195381634.001.0001

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. *British Journal of Educational Psychology, 46,* 4–11. http://dx.doi.org/10.1111/j.2044-8279.1976.tb02980.x

Mayer, R. E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. *Instructional Science, 10,* 135–175. http://dx.doi.org/10.1007/BF00132515

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General, 143,* 668–693. http://dx.doi.org/10.1037/a0032963

McLure, M. D., Friedman, S. E., & Forbus, K. D. (2010). Learning concepts from sketches via analogical generalization and near-misses. In *Proceedings of the 32nd annual meeting of the cognitive science society* (pp. 465–470). Austin, TX: Cognitive Science Society.

McRae, K., Cree, G. S., Westmacott, R., & De Sa, V. R. (1999). Further evidence for feature correlations in semantic memory. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 53,* 360–373.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238. http://dx.doi.org/10.1037/0033-295X.85.3.207

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19,* 242–279. http://dx.doi.org/10.1016/0010-0285(87)90012-0

Murphy, G. (2004). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316. http://dx.doi.org/10.1037/0033-295X.92.3.289

Namy, L. L., & Clepper, L. E. (2010). The differing roles of comparison and contrast in children's categorization. *Journal of Experimental Child Psychology, 107,* 291–305. http://dx.doi.org/10.1016/j.jecp.2010.05.013

Natal, S. D. C., McLaren, I. P. L., & Livesey, E. J. (2013). Generalization of feature- and rule-based learning in the categorization of dimensional stimuli: Evidence for dual processes under cognitive control. *Journal of Experimental Psychology: Animal Behavior Processes, 39,* 140–151. http://dx.doi.org/10.1037/a0031352

Nesher, P., Greeno, J. G., & Riley, M. S. (1982). The development of semantic categories for addition and subtraction. *Educational Studies in Mathematics, 13,* 373–394. http://dx.doi.org/10.1007/BF00366618

Nicks, D. C. (1959). Prediction of sequential two-choice decisions from event runs. *Journal of Experimental Psychology, 57,* 105–114. http://dx.doi.org/10.1037/h0045193

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114. http://dx.doi.org/10.1037/0278-7393.10.1.104

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101,* 53–79. http://dx.doi.org/10.1037/0033-295X.101.1.53

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest, 9,* 105–119. http://dx.doi.org/10.1111/j.1539-6053.2009.01038.x

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31,* 109–130. http://dx.doi.org/10.1017/S0140525X08003543

Perruchet, P. (2015). Dissociating conscious expectancies from automatic link formation in associative learning: A review on the so-called Perruchet effect. *Journal of Experimental Psychology: Animal Learning and Cognition, 41,* 105–127. http://dx.doi.org/10.1037/xan0000060

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82,* 33–40. http://dx.doi.org/10.1037/0022-0663.82.1.33

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research, 63,* 167–199. http://dx.doi.org/10.3102/00346543063002167

Pintrich, P. R., Smith, D., Garcia, T., & McKeachie, W. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. Ann Arbor, MI: The University of Michigan.

Ramsburg, J. T., & Ohlsson, S. (2016). Category change in the absence of cognitive conflict. *Journal of Educational Psychology, 108,* 98–113. http://dx.doi.org/10.1037/edu0000050

Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology, 41,* 1–48. http://dx.doi.org/10.1006/cogp.1999.0735

Rawson, K. A., Thomas, R. C., & Jacoby, L. L. (2015). The power of examples: Illustrative examples enhance conceptual learning of declarative concepts. *Educational Psychology Review, 27,* 483–504.

Rehder, B., Colner, R. M., & Hoffman, A. B. (2009). Feature inference learning and eyetracking. *Journal of Memory and Language, 60,* 393–419.

Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1261–1275. http://dx.doi.org/10.1037/0278-7393.27.5.1261

Rein, J. R., Goldwater, M. B., & Markman, A. B. (2010). What is typical about the typicality effect in category-based induction? *Memory & Cognition, 38,* 377–388. http://dx.doi.org/10.3758/MC.38.3.377

Resnick, L. B. (2010). Nested learning systems for the thinking curriculum. *Educational Researcher, 39,* 183–197. http://dx.doi.org/10.3102/0013189X10364671

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review, 21,* 1323–1330. http://dx.doi.org/10.3758/s13423-014-0588-3

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher, 39,* 406–412. http://dx.doi.org/10.3102/0013189X10374770

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605. http://dx.doi.org/10.1016/0010-0285(75)90024-9

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 629–639. http://dx.doi.org/10.1037/0278-7393.13.4.629

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology, 38,* 495–553. http://dx.doi.org/10.1006/cogp.1998.0712

Ross, B. H., Perkins, S. J., & Tenpenny, P. L. (1990). Reminding-based category learning. *Cognitive Psychology, 22,* 460–492. http://dx.doi.org/10.1016/0010-0285(90)90010-2

Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive Science, 36,* 919–932. http://dx.doi.org/10.1111/j.1551-6709.2012.01253.x

Sakamoto, Y., & Love, B. C. (2010). Learning and retention through predictive inference and classification. *Journal of Experimental Psychology: Applied, 16,* 361–377. http://dx.doi.org/10.1037/a0021610

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology, 82,* 498–504. http://dx.doi.org/10.1037/0022-0663.82.3.498

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology, 103,* 759–775. http://dx.doi.org/10.1037/a0025140

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22,* 129–184.

Sewell, D. K., & Lewandowsky, S. (2011). Restructuring partitioned knowledge: The role of recoordination in category learning. *Cognitive Psychology, 62,* 81–122. http://dx.doi.org/10.1016/j.cogpsych.2010.09.003

Shone, L. T., Harris, I. M., & Livesey, E. J. (2015). Automaticity and cognitive control in the learned predictiveness effect. *Journal of Experimental Psychology: Animal Learning and Cognition, 41,* 18.

Stains, M., & Talanquer, V. (2008). Classification of chemical reactions: Stages of expertise. *Journal of Research in Science Teaching, 45,* 771–793.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1411–1436. http://dx.doi.org/10.1037/0278-7393.24.6.1411

Spalding, T. L., & Murphy, G. L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 525–538. http://dx.doi.org/10.1037/0278-7393.22.2.525

Spalding, T. L., & Ross, B. H. (1994). Comparison-based learning: Effects of comparing instances during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1251–1263.

Tomlinson, M. T., & Love, B. C. (2006, July). From pigeons to humans: Grounding relational learning in concrete examples. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, No. 1, p. 199). Cambridge, MA: MIT Press.

Tomlinson, M. T., & Love, B. C. (2010). When learning to classify by relations is easier than by features. *Thinking & Reasoning, 16,* 372–401. http://dx.doi.org/10.1080/13546783.2010.530464

Trench, M., & Minervino, R. (2017). Cracking the problem of inert-knowledge: Portable strategies to access distant analogs from memory. *Psychology of Learning and Motivation, 66,* 1–41. http://dx.doi.org/10.1016/bs.plm.2016.11.001

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28,* 127–154. http://dx.doi.org/10.1016/0749-596X(89)90040-5

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

Wegner, E., & Nückles, M. (2016). Training the brain or tending a garden? Students' metaphors of learning predict their self-reported learning patterns. *Frontline Learning Research, 3,* 95–109.

Whitehead, A. N. (1929). *The aims of education and other essays*. New York, NY: Free Press.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and language, 39,* 124–148.

Yang, L. X., & Lewandowsky, S. (2004). Knowledge partitioning in categorization: Constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1045–1064. http://dx.doi.org/10.1037/0278-7393.30.5.1045

Appendix

Questionnaires and Correlation Matrices

Learning Strategy Questionnaire (Experiments 2 & 3)

Learning strategy. While you were learning the categories, were you more focused on trying to learn the individual items, or trying to develop a rule for why items were members of each category? The linear analog response scale ranged from "Relied solely on memorization" to "Relied solely on developing a rule."

Transfer strategy. While you were categorizing new words, were you relying on similarity to the old words, or relying on a rule? The linear analog response scale ranged from "Relied completely on similarity" to "Relied completely on a rule."

Rule awareness. Were you aware that the last letter of each word determined which category it belonged to? The linear analog response scale ranged from "Not at all aware" to "Very aware."

Rule knowledge. If the word ended with 'k,' which category did it belong to? The linear analog scale ranged from "Always animal" to "Always plant."

Rule knowledge. If the word ended with 't,' which category did it belong to? The linear analog scale ranged from "Always animal" to "Always plant."

Cognitive Reflection Task Questions (Frederick, 2005)

(a) A bat and a ball cost $1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost? ____ cents

(b) If it takes 5 machines 5 min to make 5 widgets, how long would it take 100 machines to make 100 widgets? ____ min

(c) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? ____ days

Table A1
Correlation Matrix for all Participants in Experiment 2

Variable	Learning accuracy	Baseline trials	Feature trials	Relation trials	Cross-mapped trials	Far transfer	WLT strategy	RPM	O-span	CRT
Learning Accuracy										
r	—									
p										
Baseline trials										
r	.781**	—								
p	<.001									
Feature trials										
r	.611**	.504**	—							
p	<.001	<.001								
Relation trials										
r	−.027	.113	−.503**	—						
p	.807	.302	<.001							
Cross-mapped trials										
r	−.373**	−.292**	−.877**	.684**	—					
p	<.001	.006	<.001	<.001						
Far transfer										
r	.284**	.289**	−.062	.315**	.216*	—				
p	.008	.007	.573	.003	.046					
WLT strategy										
r	.009	.058	−.032	.096	.09	.02	—			
p	.936	.598	.773	.38	.411	.857				
RPM										
r	.262*	.214*	.264*	−.086	−.238*	.023	0	—		
p	.015	.047	.014	.43	.027	.835	.997			
O-span										
r	.105	.162	−.014	−.01	−.062	.009	.18	.414**	—	
p	.423	.216	.915	.942	.638	.947	.17	.001		
CRT										
r	.179	.14	.221*	−.159	−.17	.156	.027	.551**	.246	—
p	.099	.198	.041	.144	.117	.151	.804	<.001	.059	

Note. Correlations with O-span based on 60 participants, all other correlations based on 86 participants.
* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

(*Appendix continues*)

Table A2
Correlation Matrix for Participants in the Blocked Group From Experiment 2

Variable	Learning accuracy	Baseline trials	Feature trials	Relation trials	Cross-mapped trials	Far transfer	WLT strategy	RPM	O-span	CRT
Learning accuracy										
r	—									
p										
Baseline trials										
r	.789**	—								
p	<.001									
Feature trials										
r	.573**	.475**	—							
p	<.001	.001								
Relation trials										
r	.101	.237	−.456**	—						
p	.518	.127	.002							
Cross-mapped trials										
r	−.307*	−.185	−.874**	.694**	—					
p	.045	.235	<.001	<.001						
Far transfer										
r	.315*	.308*	−.222	.411**	.382*	—				
p	.04	.045	.153	.006	.012					
WLT strategy										
r	.013	.084	−.081	.346*	.22	.152	—			
p	.932	.594	.606	.023	.157	.331				
RPM										
r	.278	.225	.275	−.07	−.201	−.053	−.099	—		
p	.071	.147	.074	.658	.196	.736	.53			
O-span										
r	−.002	.132	.048	.136	−.019	−.078	−.146	.523**	—	
p	.99	.479	.797	.465	.92	.678	.432	.003		
CRT										
r	.11	.094	.14	−.115	−.133	−.004	−.307*	.525**	.376*	—
p	.482	.547	.371	.464	.395	.981	.045	<.001	.037	

Note. Correlations with O-span based on 31 participants, all other correlations based on 43 participants.
* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

(*Appendix continues*)

Table A3
Correlation Matrix for Participants in the Interleaved Group From Experiment 2

Variable	Learning accuracy	Baseline trials	Feature trials	Relation trials	Cross-mapped trials	Far transfer	WLT strategy	RPM	O-span	CRT
Learning accuracy										
r	—									
p										
Baseline trials										
r	.780**	—								
p	<.001									
Feature trials										
r	.681**	.533**	—							
p	<.001	<.001								
Relation trials										
r	−.144	.009	−.522**	—						
p	.357	.952	<.001							
Cross-mapped trials										
r	−.473**	−.417**	−.871**	.633**	—					
p	.001	.005	<.001	<.001						
Far transfer										
r	.258	.272	.14	.215	.005	—				
p	.094	.078	.369	.166	.972					
WLT strategy										
r	.001	.018	.004	−.18	−.049	−.126	—			
p	.996	.908	.982	.249	.755	.422				
RPM										
r	.259	.17	.147	.041	−.141	.105	.056	—		
p	.094	.277	.348	.792	.368	.501	.72			
O-span										
r	.24	.263	−.025	−.11	−.127	.094	.464*	.468*	—	
p	.211	.169	.899	.572	.51	.629	.011	.01		
CRT										
r	.224	.152	.242	−.128	−.124	.302*	.294	.518**	.227	—
p	.149	.331	.117	.415	.428	.049	.056	<.001	.236	

Note. Correlations with O-span based on 29 participants, all other correlations based on 43 participants.
* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

(*Appendix continues*)

Table A4
Correlation Matrix for all Participants in Experiment 3

Variable	Learning accuracy	Baseline trials	Feature trials	Relation trials	Cross-mapped trials	Far transfer	WLT strategy	RPM
Learning accuracy								
r	—							
p								
Baseline trials								
r	.831**	—						
p	<.001							
Feature trials								
r	.19	.280**	—					
p	.077	.009						
Relation trials								
r	.309**	.267*	−.711**	—				
p	.004	.013	<.001					
Cross-mapped trials								
r	.068	.017	−.882**	.831**	—			
p	.53	.876	<.001	<.001				
Far transfer								
r	.300**	.304**	−.067	.242*	.125	—		
p	.005	.004	.535	.024	.25			
WLT Strategy								
r	.197	.166	−.039	.166	.077	.108	—	
p	.067	.125	.717	.125	.48	.321		
RPM								
r	.261*	.269*	−.064	.161	.105	.166	.093	—
p	.015	.012	.554	.136	.331	.125	.392	

* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

Table A5
Correlation Matrix for Participants in the 1-Square Condition of Experiment 3

Variable	Learning accuracy	Baseline trials	Feature trials	Relation trials	Cross-mapped trials	Far transfer	WLT strategy	RPM
Learning accuracy								
r	—							
p								
Baseline trials								
r	.815**	—						
p	<.001							
Feature trials								
r	.494**	.536**	—					
p	.001	<.001						
Relation trials								
r	.14	.096	−.569**	—				
p	.371	.539	<.001					
Cross-mapped trials								
r	−.181	−.157	−.826**	.757**	—			
p	.246	.314	<.001	<.001				
Far transfer								
r	.105	.191	.134	.05	−.059	—		
p	.501	.219	.392	.752	.706			
WLT Strategy								
r	.18	.089	.321*	−.162	−.246	−.041	—	
p	.247	.572	.036	.298	.112	.794		
RPM								
r	.215	.353*	−.041	.226	.183	.004	−.059	—
p	.167	.02	.795	.145	.241	.979	.706	

* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

(*Appendix continues*)

Table A6
Correlation Matrix for Participants in the 4-Square Condition of Experiment 3

Variable	Learning accuracy	Baseline trials	Feature trials	Relation trials	Cross-mapped trials	Far transfer	WLT strategy	RPM
Learning accuracy								
r	—							
p								
Baseline trials								
r	.846**	—						
p	<.001							
Feature trials								
r	−.04	.073	—					
p	.795	.638						
Relation trials								
r	.417**	.390**	−.795**	—				
p	.005	.009	<.001					
Cross-mapped trials								
r	.222	.125	−.922**	.858**	—			
p	.147	.417	<.001	<.001				
Far transfer								
r	.477**	.413**	−.223	.366*	.233	—		
p	.001	.005	.145	.015	.128			
WLT strategy								
r	.204	.241	−.375*	.409**	.315*	.255	—	
p	.185	.115	.012	.006	.037	.094		
RPM								
r	.333*	.228	−.157	.204	.151	.348*	.275	—
p	.027	.137	.308	.183	.327	.021	.071	

* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

Table A7
Correlation Matrix for Participants in the Full Color Saturation Condition of Experiment 3

Variable	Learning accuracy	Baseline trials	Feature trials	Relation trials	Cross-mapped trials	Far transfer	WLT strategy	RPM
Learning accuracy								
r	—							
p								
Baseline trials								
r	.764**	—						
p	<.001							
Feature trials								
r	.158	.285	—					
p	.305	.061						
Relation trials								
r	.188	.176	−.800**	—				
p	.221	.253	<.001					
Cross-mapped trials								
r	−.062	−.159	−.959**	.852**	—			
p	.69	.301	<.001	<.001				
Far transfer								
r	.198	.206	−.151	.184	.155	—		
p	.197	.18	.328	.231	.316			
WLT strategy								
r	.132	.041	−.052	.043	.012	−.003	—	
p	.393	.793	.738	.78	.936	.984		
RPM								
r	.057	.165	−.328*	.255	.327*	.07	.06	—
p	.715	.284	.03	.095	.03	.65	.699	

* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

(*Appendix continues*)

Table A8
Correlation Matrix for Participants in the Half Color Saturation Condition of Experiment 3

Variable	Learning accuracy	Baseline trials	Feature trials	Relation trials	Cross-mapped trials	Far transfer	WLT strategy	RPM
Learning accuracy								
r	—							
p								
Baseline trials								
r	.886**	—						
p	<.001							
Feature trials								
r	.193	.217	—					
p	.216	.163						
Relation trials								
r	.440**	.403**	−.617**	—				
p	.003	.007	<.001					
Cross-mapped trials								
r	.202	.191	−.812**	.804**	—			
p	.195	.219	<.001	<.001				
Far transfer								
r	.393**	.393**	.007	.307*	.098	—		
p	.009	.009	.962	.045	.531			
WLT strategy								
r	.269	.297	.012	.264	.117	.22	—	
p	.081	.053	.937	.087	.454	.157		
RPM								
r	.420**	.349*	.16	.08	−.09	.253	.122	—
p	.005	.022	.305	.609	.564	.102	.437	

* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

Table A9
Correlation Matrix for All Participants in Experiment 4

Variable	Learning accuracy	Baseline trials	Feature trials	Relation trials	Cross-mapped trials	Far transfer	MSLQ: Elaboration	RPM
Learning accuracy								
r	—							
p								
Baseline trials								
r	.052	—						
p	.645							
Feature trials								
r	.312**	.407**	—					
p	.005	<.001						
Relation trials								
r	−.09	.519**	−.374**	—				
p	.425	<.001	.001					
Cross-mapped trials								
r	−.268*	.049	−.814**	.765**	—			
p	.016	.662	<.001	<.001				
Far transfer								
r	.008	.224*	−.205	.394**	.358**	—		
p	.945	.044	.067	<.001	.001			
MSLQ: Elaboration								
r	.089	.048	−.114	.075	.174	−.1	—	
p	.43	.672	.312	.505	.12	.375		
RPM								
r	−.028	.107	.044	.133	.096	.256*	−.059	—
p	.802	.342	.698	.237	.394	.021	.602	

* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

(*Appendix continues*)

Table A10
Correlation Matrix for Participants in the Classification Learning Group in Experiment 4

Variable	Learning accuracy	Baseline trials	Feature trials	Relation trials	Cross-mapped trials	Far transfer	MSLQ: Elaboration	RPM
Learning accuracy								
r	—							
p								
Baseline trials								
r	.205	—						
p	.199							
Feature trials								
r	.295	.398**	—					
p	.061	.01						
Relation trials								
r	−.026	.390*	−.520**	—				
p	.871	.012	<.001					
Cross-mapped trials								
r	−.208	−.135	−.898**	.761**	—			
p	.192	.401	<.001	<.001				
Far transfer								
r	.055	.204	−.069	.12	.132	—		
p	.731	.201	.667	.454	.411			
MSLQ: Elaboration								
r	−.031	−.157	−.497**	.252	.506**	.06	—	
p	.847	.325	.001	.112	.001	.71		
RPM								
r	.161	.207	.043	.157	.158	.167	.087	—
p	.314	.194	.791	.328	.324	.298	.587	

* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

Table A11
Correlation Matrix for Participants in the Inference-Learning Group in Experiment 4

Variable	Learning accuracy	Baseline trials	Feature trials	Relation trials	Cross-mapped trials	Far transfer	MSLQ: Elaboration	RPM
Learning Accuracy								
r	—							
p								
Baseline trials								
r	−.17	—						
p	.294							
Feature trials								
r	.066	.332*	—					
p	.684	.036						
Relation trials								
r	−.044	.703**	−.156	—				
p	.786	<.001	.337					
Cross-mapped trials								
r	−.103	.433**	−.566**	.822**	—			
p	.525	.005	<.001	<.001				
Far transfer								
r	.136	.361*	−.157	.555**	.495**	—		
p	.401	.022	.332	<.001	.001			
MSLQ: Elaboration								
r	.079	.097	.096	−.008	−.011	−.151	—	
p	.626	.553	.554	.963	.946	.353		
RPM								
r	−.253	.029	.011	.127	.056	.392*	−.229	—
p	.116	.859	.947	.433	.731	.012	.155	

* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

(*Appendix continues*)

Table A12
Correlation Matrix Within the Word-Learning Task in Experiment 2

Variable	WLT learning strategy	WLT transfer strategy	WLT rule awareness	WLT rule knowledge	WLT training average	WLT trained test	WLT ambiguous test	WLT exemplar test	WLT rule test
WLT learning strategy									
r	—								
p									
WLT transfer strategy									
r	.536**	—							
p	<.001								
WLT rule awareness									
r	.431**	.466**	—						
p	<.001	<.001							
WLT rule knowledge									
r	.126	.279**	.415**		—				
p	.247	.009	<.001						
WLT training average									
r	.104	.111	.578**	.418**					
p	.338	.307	<.001	<.001					
WLT trained test									
r	.048	−.032	.239*	.189	.387**	—			
p	.66	.768	.026	.081	<.001				
WLT ambiguous test									
r	.462**	.530**	.779**	.311**	.368**	.153	—		
p	<.001	<.001	<.001	.004	<.001	.161			
WLT exemplar test									
r	−.286**	−.458**	−.315**	.042	.078	.064	−.440**	—	
p	.008	<.001	.003	.702	.476	.56	<.001		
WLT rule test									
r	.168	.229*	.440**	.484**	.301**	.253*	.342**	−.053	—
p	.122	.034	<.001	<.001	.005	.019	.001	.625	

Note. Rule knowledge refers to average accuracy for questions 4 and 5 in the WLT questionnaire.
* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

(*Appendix continues*)

Table A13
Correlation Matrix Within the Word-Learning Task in Experiment 3

Variable	WLT learning strategy	WLT transfer strategy	WLT rule awareness	WLT rule knowledge	WLT training average	WLT trained test	WLT ambiguous test	WLT exemplar test	WLT rule test
WLT learning strategy									
r	—								
p									
WLT transfer strategy									
r	.383**	—							
p	<.001								
WLT rule awareness									
r	.485**	.698**	—						
p	<.001	<.001							
WLT rule knowledge									
r	.181	.168	.323**	—					
p	.093	.12	.002						
WLT training Average									
r	.470**	.296**	.502**	.275**	—				
p	<.001	.005	<.001	.01					
WLT trained test									
r	.287**	.067	.159	.174	.535**	—			
p	.007	.54	.141	.108	<.001				
WLT ambiguous test									
r	.295**	.566**	.711**	.332**	.248*	.017	—		
p	.005	<.001	<.001	.002	.021	.874			
WLT exemplar test									
r	−.196	−.317**	−.378**	−.125	−.096	.086	−.582**	—	
p	.068	.003	<.001	.248	.376	.429	<.001		
WLT rule test									
r	.146	.003	.112	.031	.134	−.006	−.112	.192	—
p	.176	.982	.3	.773	.215	.959	.3	.074	

Note. Rule knowledge refers to average accuracy for questions 4 and 5 in the WLT questionnaire.
* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

Received October 21, 2016
Revision received October 5, 2017
Accepted October 11, 2017 ■