Journal of Experimental Psychology: Learning, Memory, and Cognition

Enhancing Visuospatial Mapping in Relational Category Learning

Andrew J. Lee, Keith J. Holyoak, and Hongjing Lu

Online First Publication, July 3, 2025. https://dx.doi.org/10.1037/xlm0001514

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology:
Learning, Memory, and Cognition

ISSN: 0278-7393

https://doi.org/10.1037/xlm0001514

Enhancing Visuospatial Mapping in Relational Category Learning

Andrew J. Lee[1], Keith J. Holyoak[1], and Hongjing Lu[1, 2]
[1] Department of Psychology, University of California, Los Angeles
[2] Department of Statistics, University of California, Los Angeles

Visual relational concepts—defined by patterns of relationships between entities—are thought to require structured, compositional representations with explicit role information about each entity. Analogical mapping over compositional representations is a key strategy for acquiring such concepts, but in complex situations with many entities and relations, this process can be cognitively demanding. As a result, learning may occur over feature-based representations, where exemplars are encoded as unstructured lists of entities and relations, losing crucial role information and limiting generalizability. To reduce the cognitive load of analogical mapping, we explored the effectiveness of two visuospatial training aids: (a) spatially organizing exemplars by category to facilitate comparisons and (b) using color coding to highlight the roles of entities within each exemplar. Across three experiments, we examined whether these visuospatial aids improve learning rates on the Synthetic Visual Reasoning Test (SVRT), a collection of 23 problems that require learning relational concepts. Our results showed that displays of previous instances that spatially sorted them into positive and negative sets led to faster concept learning. Learning was faster overall when problems were ordered easy-to-hard rather than randomly, but sorted displays were more effective in either case. Color coding proved beneficial only when colors unambiguously and nonredundantly linked entities that played corresponding roles; when color coding did not support a clear mapping, it interfered with learning. These findings suggest that rapid learning of relational concepts can be facilitated by display characteristics that support analogical mapping by comparisons.

Keywords: categorization, analogy, mapping, relations

Visual categories have traditionally been defined as sets of exemplars that share similar features with either a prototype or each other (Estes, 1986; Tversky, 1977). Historically, what counts as a "feature" has been construed broadly, encompassing virtually any aspect of a stimulus from low-level visual details to high-level semantics. Many perceptual learning experiments have focused on predefined features of artificial stimuli (e.g., Erickson & Kruschke, 1998; Nosofsky & Palmeri, 1998; Zaki & Salmi, 2019). Some human modeling studies have used the complex distributed patterns of activities extracted by convolutional neural networks trained on naturalistic images (Battleday et al., 2020, 2021). Other work has explored the hypothesis that features can be created by the category learning goal itself (Goldstone, 1998, 2000; Schyns et al., 1998). Despite their unbounded range of potential content, features act as independent elements in classification models; no single feature possesses a uniquely different form or representational type than another.

Although a variety of feature-based models have been applied to human category learning, it is widely believed that *relational* categories require a fundamentally different representational format. Many everyday concepts are defined by patterns of relationships between parts, rather than the individual parts or their properties (e.g., *inside* is a relational concept that holds when any object is enclosed within another; Asmuth & Gentner, 2017; Gentner & Kurtz, 2005; Goldwater & Schalk, 2016). As such, relational categories naturally lend themselves to structured, compositional representations in which relations and individual entities are encoded as separable components that combine to form relational structures (Doumas et al., 2022; Hummel & Holyoak, 1997, 2003; Kurtz et al., 2013; Shurkova & Doumas, 2022). A compositional approach to

Andrew L. Cohen served as action editor.
Andrew J. Lee https://orcid.org/0000-0002-1708-2538
A partial report of this work was presented at the 43rd Annual Meeting of the Cognitive Science Society in 2023. Preparation of this article was supported by National Science Foundation Grant IIS-1956441 to Hongjing Lu.

The authors thank Rishi Deorah, Ziqi Zheng, Zixuan Zhou, Ashley Choy, Yining Liang, and Rui Yu for data collection and helpful discussions, and Barbara Knowlton, Nick Ichien, Emily Grossman, and Alice Xu for insightful comments.

Andrew J. Lee played a lead role in data curation, formal analysis, investigation, project administration, software, supervision, validation, visualization, writing–original draft, and writing–review and editing, a supporting role in funding acquisition and resources, and an equal role in conceptualization and methodology. Keith J. Holyoak played a lead role in writing–review and editing, a supporting role in data curation, formal analysis, funding acquisition, resources, and visualization, and an equal role in conceptualization, investigation, methodology, project administration, supervision, and writing–original draft. Hongjing Lu played a lead role in funding acquisition, resources, and writing–review and editing, a supporting role in data curation, formal analysis, project administration, software, validation, visualization, and writing–original draft, and an equal role in conceptualization, investigation, methodology, and supervision.

Correspondence concerning this article should be addressed to Andrew J. Lee, Department of Psychology, University of California, Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095-1563, United States. Email: andrewlee0@g.ucla.edu

relational categories is consistent with evidence that visual perception is organized by relations. Some basic visuospatial relations may be perceived quickly and automatically (Hafri & Firestone, 2021), and object representations appear to encode the hierarchical structure of the scenes in which they are embedded (Turini & Võ, 2022).

Mechanisms for Relational Category Learning

One long-posited mechanism for acquiring relational representations of categories is analogical mapping (Gick & Holyoak, 1983). Analogical mapping involves aligning two or more exemplars so as to identify correspondences between entities that play similar roles within their respective relational structures. Once an optimal alignment is found, a relational category is formed by abstracting the common structure. This structured, relationally sensitive style of comparison is widely regarded as the mechanism by which humans learn relational categories (Christie & Gentner, 2010; Halford & Busby, 2007; Halford et al., 1998; Jung & Hummel, 2015; Kittur et al., 2004; Kurtz et al., 2013) and appears to occur automatically when tasked to compare two exemplars without explicit instruction to focus on relational similarities (Gentner & Markman, 1997; Markman & Gentner, 1993a, 1993b). But although spontaneous and unprompted comparisons are a frequent part of everyday cognition, it is also the case that analogical mapping requires substantial cognitive effort, as aligning and maintaining multiple relational structures places heavy demands on working memory and its executive functions, such as inhibitory control (e.g., Phillips et al., 2016; Waltz et al., 2000).

Partly because of the cognitive demands apparently imposed by analogical mapping, recent work has explored the possibility of alternative representational formats. Corral et al. (2018) suggested that learners may opt for simpler "featural" representations when these are less demanding and adequate for a same-different judgment task. Here, "features" include not only visual elements but also the *relation* itself. When categories differ only with respect to the presence or absence of a relation (e.g., a ball inside vs. outside of a box), the roles of the individual objects become unnecessary, and learners only need to notice the difference in the relation to make a similarity judgment. Rather than representing the category compositionally as *inside (ball, box)*, learners may use the unstructured "flattened" list of concept-level features *ball*, *box*, and *inside*, where the relation *inside* is encoded but the roles of *ball* and *box* are not. The choice of representation may depend on cognitive load and task demands, with learners favoring a less complex shortcut when it sufficiently supports the task (Ichien et al., 2024). However, Corral et al. point out that such featural representations, while less demanding, are limited in supporting generalization and inference precisely because they fail to capture specific object roles. The limitation arises because these representations cannot distinguish between, for instance, a category defined by a square inside a triangle versus a triangle inside a square. Role-filler bindings are crucial for generalization as they specify how elements relate to one another and ensure that relational concepts transfer properly to new contexts with different elements (Hummel et al., 2004).

Another alternative learning strategy proposes that relational categories can be represented as generative programs capable of recreating category members (Ellis et al., 2015). Like relational structures, generative programs are compositional representations composed of separately encoded parts and relations. However, unlike the purely static connections of relational structures, relations in programs can be action-oriented (i.e., *draw [x]*, *move [y, 3 steps]*), specifying a sequence of operations in addition to semantic, spatial, and comparative relations. These structured operations, much like a computer program, unfold in a sequence to generate category members. Models based on generative programs often create a vast space of possible program representations by recombining existing elements and functions (Lake & Piantadosi, 2020; Lake et al., 2015). The most plausible program is then selected using Bayesian inference, an exhaustive search method that ranks the likelihood of each program generating exemplars correctly. Although generative programs offer a different form of compositional representation, it is unclear whether this approach circumvents the level of computational demands imposed by analogical mapping.

Overview of the Present Study

An important objective in cognitive science is to identify task constraints that shape the form of conceptual representations. A second objective is to develop learning methods that foster knowledge capable of broad generalization (Goldwater & Schalk, 2016). If analogical mapping—which relies on compositional representations based on role-filler bindings—is hindered by cognitive load, it is essential to explore training conditions that may reduce this load. To reduce the cognitive load associated with analogical mapping and encourage learners to engage with compositional representations, we consider two potential training aids: (a) spatially organizing exemplars by category and (b) using color to highlight the roles of entities within each exemplar.

Sorting exemplars spatially (into groups by category) can reduce cognitive load by easing a tendency to *compare* examples by highlighting similar relations within categories and different relations across categories—a critical step of analogical mapping. Such comparison could involve representations of relations as independent features (without role-filler bindings) rather than as part of a compositional structure. However, evidence suggests that comparison of scenes with shared relational structures induces an alignment of those scenes based on relational roles beyond simple visual similarity (Gentner & Markman, 1997; Markman & Gentner, 1993a, 1993b). Nevertheless, color coding of exemplars may more explicitly highlight role-filler bindings than a mere sorting of exemplars and could thereby direct attention to overall compositional structure. In educational settings, color coding can be used to highlight analogous parts of scientific and mathematic diagrams (Gray & Holyoak, 2021). However, the possible impact of color coding on visual relational category learning has yet to be explored experimentally.

In this study, we test the impact of organized sorted displays and color-coded exemplars on relational category learning using the Synthetic Visual Reasoning Test (SVRT; Fleuret et al., 2011). The SVRT consists of 23 visual relational concepts, each defined by an abstract pattern of visuospatial relations over objects. Each concept includes a large pool of positive and negative examples, in which positive images depict the defining relational pattern (e.g., two objects touching at a point) and negative images deviate from it (e.g., two objects that do not touch). Each image is perceptually simple, with a few black pixelated shapes that resemble islands populating a white background. The goal of the SVRT is to learn the abstract

relational pattern that defines each concept and correctly categorize new images into the positive or negative set. Various models have been applied to the SVRT using program synthesis, connectionist-style analogical mapping, or convolutional and transformer-style deep learning (Ellis et al., 2015; Messina et al., 2021, 2022; Shurkova & Doumas, 2022).

Several properties make the SVRT particularly suitable for studying relational category learning. First, SVRT problems vary widely in difficulty, and this range reflects the complexity and type of relations involved. Humans require, on average, 5.7 classification attempts to achieve seven correct classifications of novel images in a row, with a range from 2.4 to 9.3 attempts across problems (reconstructed from Fleuret et al., 2011). At least 10 of the 23 concepts are based on sameness or difference of shape; the rest involve other spatial relations (Kim et al., 2018). Most concepts are based on a single relation, but several involve conjunctions of relations (e.g., two objects have the same shape, and one is inside the other) or complex patterns over multiple objects (e.g., the distance between objects in a pair is the same across pairs). See Table 1 for a complete description of the relational concepts underlying each problem.

Second, the SVRT's design compels learners to identify relations between objects rather than simple visual features of the objects. This is because negative examples are "hard negatives"—they contain the same number of objects with similar visual features as positive examples but differ in the relational patterns formed by the objects. Moreover, negative examples often exhibit their own coherent relational patterns. In problem 11, for example, negative images are not merely the negation of "two objects touching at a point"; instead, they consistently show two objects separated by roughly equal distance. This "positive" property of negative images makes it challenging to distinguish positives and negatives on the basis that negatives can depict *anything* besides the positive concept, such as an unequal variance of pixel locations (Fleuret et al., 2011). Thus, the inclusion of hard negatives with their own "positive"

Table 1
Description of Positive and Negative Concepts for Each Problem, With the Relation(s) That Need to Be Identified to Distinguish Concepts

Problem	Relations needed	Positive concept	Negative concept
1	Same/different shape	Two same shapes of same medium size	Two different shapes of similar medium size
2	Center/boundary	Small shape inside and at center of large one	Small shape inside but near boundary of large one
3	Touching/not touching	Three out of four different medium-sized shapes touching at a point	Two pairs of all different medium-sized shapes, with each pair touching at a point
4	Inside/outside	Small shape inside large one	Small shape outside large one
5	Same/different shape	Two different pairs of same medium-sized shapes	Four different shapes of similar medium sizes
6	Equal/unequal distance Same/different shape	Two different pairs of same small shapes, with equal distances within pairs	Two different pairs of same small shapes, with unequal distances within pairs
7	Same/different shape	Three different pairs of same medium-sized shapes	Two different triplets of same medium-sized shapes
8	Same/different shape Inside/outside	Medium-sized shape inside large one, and both have same shape	Medium-sized shape outside large one, or inside large one but with different shape
9	Between/trailing	Medium-sized shape is in between two smaller ones	Medium-sized shape is leading or trailing two smaller ones
10	Square/not square	Four small shapes form the points of a square	Four small shapes in random locations
11	Touching/not touching	A large and a medium-sized shape touching at a point	A large and a medium-sized shape not touching
12	Equal/unequal distance	Two small shapes equally distant to a medium-sized one	Two small shapes unequally distant to a medium-sized one
13	Same/different shape	Two metashapes are identical (metashape consists of a pair of a large and small shape)	All four of the shapes are randomly located
14	Line/not line	Three small identical shapes arranged in a line	Three small identical shapes are do not form a line
15	Same/different shape	Four small identical shapes arranged in a square	Four small different shapes arranged in a square
16	Mirrored/not mirrored	Three small same shapes reflected and mirrored across vertical bisector	Three small same shapes reflected across vertical bisector but not mirrored
17	Equal/unequal distance Same/different shape	Three small identical shapes equally distant to a different small shape	Three identical small shapes randomly located
18	Mirrored/not mirrored	Three small identical shapes reflected but not mirrored across the vertical bisector	Three small identical shapes located randomly
19	Same/different shape	Two same shapes where one may vary in size	Two different shapes of varying sizes
20	Same/different shape	Two small identical shapes mirrored along some line of reflection in the image	Two different small shapes
21	Same/different shape	Two same shapes where one may vary in size and rotation	Two different shapes of varying sizes
22	Line/not line	Three small identical shapes in a line	Three small different shapes in a line
23	Inside/outside	Two different small shapes inside a large one	Two different small shapes—one is inside and another is outside the large one

concept ensures that categorization cannot rely on object recognition or salient low-level visual differences.

We note, however, that simply identifying relations that differentiate positive and negative sets does not ensure a *compositional* relational concept representation. As previously discussed, if positive and negative sets differ only in the presence or absence of a few relations, one can distinguish sets by encoding each set as a separate noncompositional list of concept-level features—capturing individual objects and relations without their relational roles—and then comparing these lists to identify differences in the relations (Corral et al., 2018). All SVRT problems can, in principle, be solved with a noncompositional approach based on flat feature lists (see Table 1). For Problem 1, for example, the positive examples exhibit the relation *same*, whereas the negative examples exhibit the relation *different*. Since this simpler strategy is cognitively less demanding, learners are likely to favor it when cognitive load is relatively high. This, in turn, allows us to assess any performance gains that may emerge specifically due to the use of mapping, as facilitated by our manipulations to reduce cognitive load.

Finally, the SVRT provides a promising testbed for probing representation learning, as its standard method of presenting examples uses category-organized displays. Fleuret et al. (2011) displayed all previously presented examples in a *sorted* format, where positive and negative examples are spatially separated below the current trial's image (see Figure 1, left). This arrangement may encourage comparisons within categories, facilitating the extraction of common relational structures; it may also promote comparisons between positive examples and the negative "near misses" (i.e., hard negatives) to highlight relational contrasts (Winston, 1975) or "alignable differences" (Gentner & Markman, 1994). Previous research has shown that simultaneous comparison of similar objects across categories enhances category learning by highlighting critical distinguishing properties (Gentner & Markman, 1994; Jee et al., 2013). However, direct experimental evidence that the organized property of sorted displays benefits learning relational categories by promoting systematic comparisons—contrary to the potential cognitive disruption of *shuffled* displays—is lacking.

To determine whether display format impacts learning on the SVRT, we conducted experiments in which the cumulative record of previous examples was either sorted (as in the original study) or shuffled, with prior examples recorded in the same random order as they had been presented. We hypothesized that if sorted displays reduce the cognitive load of comparisons, learning speeds should be faster with sorted displays. Experiment 1 tests this hypothesis. In Experiments 2 and 3, we introduced color coding in conjunction with sorted versus shuffled displays. We hypothesized that appropriate highlighting of role-filler bindings through color coding might reduce the cognitive load of analogical mapping, thereby enhancing learning efficiency.

Experiment 1

Method

Participants

Sixty-four undergraduates from the University of California, Los Angeles participated for course credit (46 female, 18 male; M_{age} = 20.1), with equal numbers assigned to the sorted and shuffled display conditions (32 each).

Stimuli

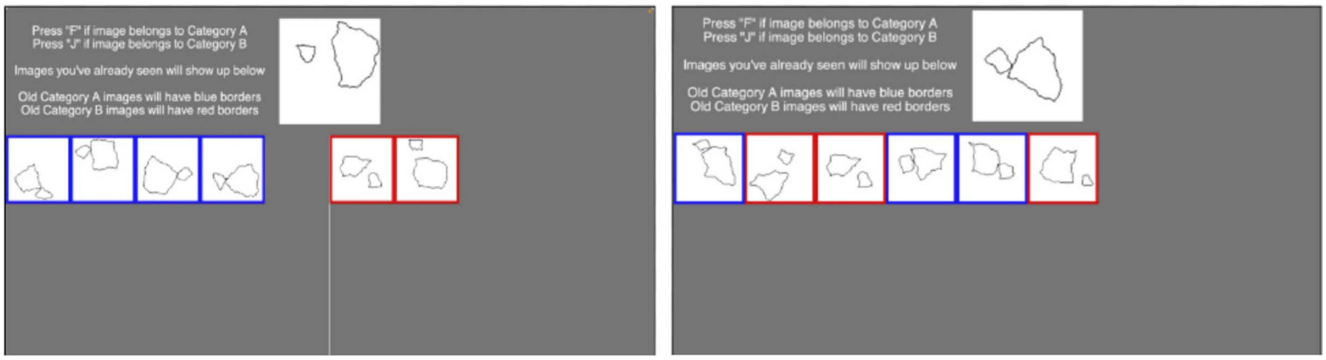
The 23 SVRT problems have been previously categorized into different types. In one study, the authors classified the problems into either same-different shape or other spatial relation problems (Kim et al., 2018). In Table 1, we provide a brief description of each problem's positive and negative sets, along with the relations that need to be utilized to distinguish the sets. Our classification of problems mirrors that of Kim et al. (2018) with some exceptions. For example, we include Problem 6 as requiring same-different shape detection because, even though equal/unequal distance distinguishes the sets, this relation operates at the level of same-shape pairs. Noticing this higher order relation requires initially detecting two pairs of same shapes.

Materials and Procedure

Participants attempted each of the 23 SVRT problem one-by-one in a quasi-randomized easy-then-hard order. The experiment began

Figure 1

An Illustration of Displays With Visual Records of Category Exemplars



Note. Left: sorted display in which previously presented instances are separated into positive versus negative examples (blue vs. red frame). Right: shuffled display in which previously presented instances are intermixed in their randomized presentation order, with positive versus negative examples distinguished by blue versus red frame. See the online article for the color version of this figure.

with a subset consisting of 13 easy problems, followed by the remaining subset of 10 harder problems, with order randomized within subsets separately for each participant. This division between easy and hard problems was based on the rank order of problem difficulty as defined by trials-to-criterion in the original study (Fleuret et al., 2011); a small gap separated the 13 easy problems from the 10 harder ones. In accord with evidence that an "easy-then-hard" training schedule can support more efficient learning (Hornsby & Love, 2014; McLaren & Suret, 2000), the easy subset of problems was presented before the hard subset. The easy-then-hard ordering was intended to aid in motivating participants by promoting their success on early problems.

On each trial of an SVRT problem, a previously unseen image was randomly selected from either the positive or negative category and presented at the center-top of the screen. Participants were instructed to classify the image (without speed pressure) into one of two categories, simply termed A and B, after which feedback was displayed for 1 s after their decision ("Correct!" or "Incorrect!"). We did not inform participants that either A or B referred to the positive or negative set, or that categories may be construed as two positive concepts. Trials for each problem continued until the participant reached a strict learning criterion of seven correct trials in a row, or a maximum of 34 trials (to control the duration of the experiment). If criterion was reached within 34 attempts, trials-to-criterion was scored as the total number of attempts minus 7. Otherwise, the problem was considered a "fail," and trials-to-criterion was set to the ceiling of 34.

Below each image to-be-classified, participants could see all images they had previously classified for the current problem (following the procedure used by Fleuret et al., 2011). In this "visual record or history" of category exemplars, every image was surrounded by a colored border denoting its true category membership (blue for category A, red for category B). Participants saw one of two types of visual records throughout the 23 problems (Figure 1). In a *sorted* display, previous exemplars were spatially separated by category, accumulating left-to-right in trial order on either the left side (category A) or right side (category B) of the visual record. In a *shuffled* display, previous images accumulated in trial order from left to right without spatial category separation; this display effectively intermixed categories (since successive trials were randomly positive or negative).

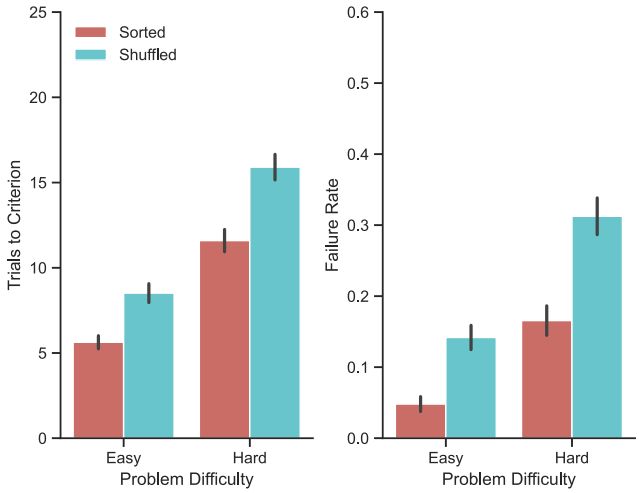
Because categories of images in the visual record were always distinguished by the color of their border, the segregated grouping of categories in the sorted display provided a redundant cue to category membership of each image. For both display types, no more than 10 images accumulated per row; whenever necessary, subsequent rows were added. Images in the record were scaled down by a factor 0.64 relative to the width at which images were initially presented, and then placed side-by-side to maximize visibility. Materials, data, and code for all experiments are available on Open Science Framework at https://osf.io/q6uad/ (Lee et al., 2024).

Results and Discussion

Two dependent variables were measured for each problem: trials-to-criterion (the number of classification attempts before seven correct in a row) and problem failure rate (based on whether or not criterion was met within the maximum allotment of 34 attempts). Figure 2 depicts the grand means of each dependent measure and condition, calculated by averaging over problems and participants. Note that lower trials-to-criterion and failure rates indicate superior performance.

Figure 2

Trials-to-Criterion (Left) and Failure Rates (Right) in Experiment 1 for Alternative Displays of the Visual Record (Sorted vs. Shuffled) and Problems of Varying Difficulty Levels



Note. Error bars represent standard errors of the mean (trials-to-criterion) and binomial standard errors (failure rate) at the individual problem level. See the online article for the color version of this figure.

To account for the nested structure of our data set, where each participant completed 23 problems, we employed mixed-effects regression models. These models were fitted separately for trials-to-criterion (using linear regression) and problem failure (using logistic regression) with the *lmer* and *glmer* functions of *lme4* Version 1.1.35.3 in R. For each dependent variable, our full model included *display type* (sorted vs. shuffled), *problem difficulty* (easy vs. hard), and their *two-way interaction* as fixed effects. We incorporated *participant ID* and *problem ID* as random intercepts to capture variations across participants and problems. In addition, *problem difficulty* was included as a random slope for *participant ID* to account for its within-subject nature.

To test the interaction between display type and problem difficulty, we used a likelihood-ratio test of the full model relative to a reduced model that lacked only the effect of interest but was otherwise equivalent to the full model. An interaction was not significant: a reduced model without the interaction term did not increase regression error for either trials-to-criterion, $\Delta AIC = 0$, $\chi^2(1) = 1.77$, $p = .18$, or problem failures, $\Delta AIC = 1.26$, $\chi^2(1) = 0.74$, $p = .39$. Next, we tested our primary prediction that sorted displays lead to improved performance compared to shuffled displays. A contrast of display type in the full model (collapsing over easy and hard problems) suggests that sorted displays led to fewer trials-to-criterion, $M = 8.22$; $t(62) = 2.81$, $p = .0065$, and a lower failure rate, $M = 0.099$; $z = 3.51$, $p = .0004$, compared to shuffled displays (trials-to-criterion: $M = 11.73$; failure rate: $M = 0.22$). Finally, to check that our selection of easy and hard problems based on the original experiment's trials-to-criterion data was, in fact, easy and hard for participants, we also tested a

contrast of problem difficulty, collapsing over display type. Our designation of easy problems indeed yielded fewer trials-to-criterion, $M = 7.07$; $t(21.6) = 3.27$, $p = .0036$, and a lower failure rate ($M = 0.095$; $z = 3.11$, $p = .0018$) than hard problems (trials-to-criterion: $M = 13.75$; failure rate: $M = 0.24$).

Although the *magnitude* of sorted-shuffled dependent variable differences may be similar across problem difficulty, it remains possible that hard problems, which have a poorer performance baseline than easy problems, demonstrate a larger *percentage* improvement for sorted displays, compared to hard problems. To account for such baseline differences, we normalized the dependent variables by expressing both trials-to-criterion and problem failure as a percentage change of its respective easy or hard marginal mean, where each marginal mean was a simple average over participants in either display type (as sorted and shuffled display conditions have equal sample sizes). Using percent-changes as the dependent variables, sorted displays resulted in improved performance compared to shuffled displays, trials-to-criterion: $t(62) = 2.74$, $p = .0080$; problem failures: $t(62) = 2.49$, $p = .016$. Thus, the overall advantage of sorted over shuffled displays clearly held for both easy and hard problem subsets in an easy-then-hard order.

The results indicate that sorted displays lead to faster learning and fewer problem failures across levels of task difficulty compared to shuffled displays. This performance advantage reflects the organizational benefits of sorted displays, which eases the difficulty of performing comparisons within and across categories. It is possible that systematically organizing the SVRT examples enhanced learning by specifically facilitating analogical mappings, but we remain agnostic about the exact mechanism.

Experiment 2

Given the promising role of sorting exemplars in supporting comparisons for relational category learning, Experiment 2 explored whether color coding could further enhance learning by explicitly reducing the cognitive load of analogical mapping. Color coding is used to teach analogies (Gray & Holyoak, 2021) with the assumption that color eases the mapping process by highlighting corresponding object roles. This assumption has not been empirically tested. In addition, we aimed to replicate the impact of display format for prior instances (sorted vs. shuffled) while changing the order of the presented problems from easy-then-hard to fully randomized.

Method

Participants

One hundred thirty-six University of California, Los Angeles undergraduates participated for course credit (105 female, 26 male, four nonbinary; $M_{age} = 20.0$). Sample sizes for each of the four between-subjects conditions were: sorted/color-mapped ($n = 35$), shuffled/color-mapped ($n = 31$), sorted/uncolored ($n = 34$), and shuffled/uncolored ($n = 36$).

Materials

For the uncolored conditions, we used the same stimuli as in Experiment 1. For the color-mapped conditions, color codes for each problem were generated based on the number of colors required. If an analogical mapping in an SVRT problem could be achieved with two colors, then a random color was chosen from red–green–blue (RGB) value 0–63 or 192–255, and its opposite was calculated (255 minus the RGB value). These RGB ranges were selected to exclude blue or red hues and minimize confusability with the colored category borders. If the analogical mapping required three or four colors, then colors were selected at random from the same RGB ranges. Images were colored with Adobe Photoshop's paint-bucket tool using point-and-click. Each SVRT problem used a different set of colors. Positive and negative images of the same SVRT problem used the same color code (i.e., same colors and same number of colors).

To avoid having colors correlated with visual features of the objects, half of the problems with a large or centrally placed object, or left-positioned objects, were colored with the darkest of each problem's color code, whereas for each problem in the other half these objects were colored with the lightest of its code. For problems defined by same-different shape, each object was given a different color from the code, regardless of whether they had the same or different shape from other objects. Furthermore, objects were randomly assigned to colors with respect to positioning, so that the locations of these colors were evenly distributed across images.

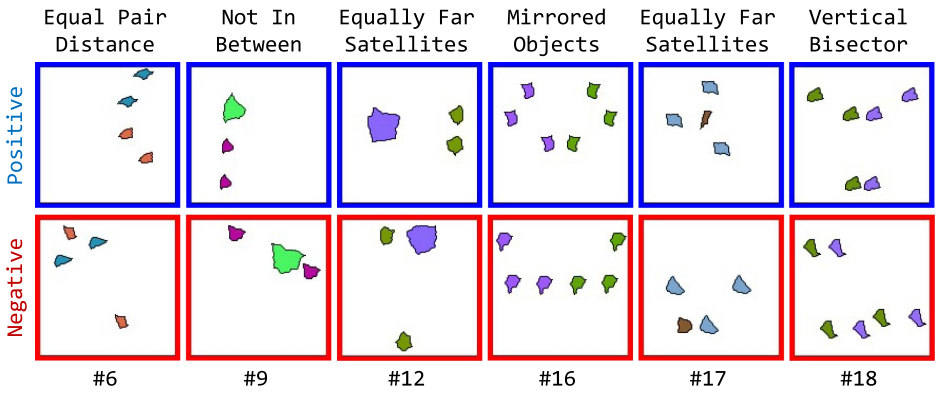
We examined each of the 23 SVRT problems and attempted to generate color codes that highlight the roles of each object as defined by the relational concept. This task was difficult because we aimed to highlight the appropriate mapping for both positive and negative instances (which formed two distinct categories), while using the same number of colors (so that number of colors would not become a superficial cue to category membership). We were able to create optimal color codes for six of the problems (Figure 3) by meeting the following five criteria (which we explicitly defined post hoc, after the color coding was constructed and data were collected):

1. *Color consistency across sets:* Both positive and negative instances used the exact same color code so that number of colors alone (a nonrelational feature) cannot distinguish sets.
2. *Multiple colors for both sets:* More than one color should be utilized by the color code, since positive and negative sets cannot be distinguished by a single color.
3. *No role-color conflicts:* The color coding should not introduce conflicts such that objects filling the same role are represented by different colors.
4. *Conceptually reinforcing roles:* The roles of objects as determined by the color coding should highlight, rather than obscure, the relational concept of the problem.
5. *Minimal perceptual redundancy:* An optimal color code should enhance role information that is not already easily determinable from visual features. We allowed for minimal size differences that correlate with role information, as in problems 9 and 12 (Figure 3).

The first criterion is essential to avoid confounding positive/negative class rules with nonrelational visual features. Two problems (3 and 7) could not meet this criterion. Data from these problems were therefore dropped from all analyses for both this experiment and Experiment 3.

Figure 3

Example Instances for Each of the Six Optimal Color-Mapped Problems

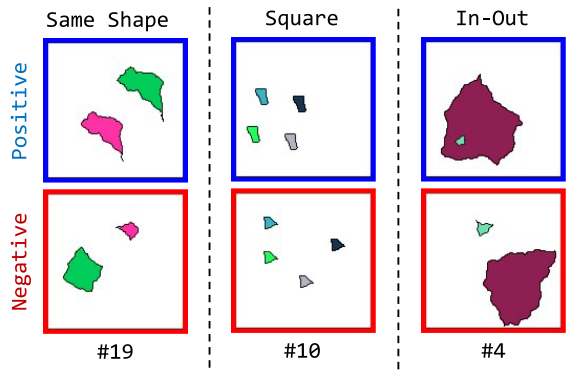


Note. Labels at the top indicate problem concept. Labels at the bottom indicate problem ID from Fleuret et al. (2011). See the online article for the color version of this figure.

For the remaining 15 problems we were able to satisfy the first and necessary criterion, but unable to satisfy all others. These 15 problems can be grouped into three challenges, as illustrated in Figure 4. The first group, consisting of nine problems, are based on same-different shape (Figure 4, left column), where positive images contain objects of the same shape and negative images contain differing shapes. If positive and negative images share the same single color to clarify the positive concept, then sameness of color would not distinguish positive and negative sets (Criterion 2) and furthermore would conflict with the difference of shape in negative images (Criterion 3). We therefore assigned a unique color to each object, while using the same set of colors for positive and negative images. However, this assignment creates an ambiguity for the positive set, since objects with the same shape are assigned different colors (violating Criterion 3).

Figure 4

Three Representative Problems of the 15 Suboptimal Color-Mapped Problems, Each Exhibiting by a Distinct Color Coding Challenge (See Main Text)



Note. Each column displays the best color code we were able to create for a problem with a distinct challenge. Labels at the top indicate problem concept. Labels at the bottom indicate problem ID from Fleuret et al. (2011). See the online article for the color version of this figure.

A second group of two problems involve a set of objects that form a geometric configuration, arranged in either a line or a square (vs. a disorganized group; Figure 4, middle column). If all objects shared the same color so as to highlight the positive concept of a geometric arrangement, then sameness of color would not distinguish sets (Criterion 2). We therefore colored each object differently, such that objects with the same relative location in the arrangement had the same color (e.g., upper right corner of square). Although this may clarify a mapping based on relative object position, coding the role of each object by difference in color may sharpen attention to the individual objects themselves rather than their larger configuration (violating Criterion 4).

In the last group of four problems, images contain a large object and one or two small objects. The concepts are defined by the rules that the objects are touching (vs. not touching), the small object is inside the large one (vs. outside), the small object is near the large object's center (vs. near its border), and that both small objects are either inside or outside the large object (Figure 4, right column). For all four problems, we were able to construct a color code that satisfies Criteria 1–3: all the large objects share the same color, and all the small objects share a different color. However, because the large size difference of objects is a strong indicator of role (as opposed to smaller size differences as in problem 9), color coding may not add relevant information that is not already apparent in uncolored images (Criterion 4). Thus, while the mapping is unambiguous, the color coding is suboptimal.

In sum, we created six optimal color-mapped problems (6, 9, 12, 16, 17, 18) and 15 suboptimal color-mapped problems (1, 2, 4, 5, 8, 10, 11, 13, 14, 15, 19, 20, 21, 22, 23). We determined this distinction post hoc. The optimal color-mapped problems have a higher average difficulty (mean trials-to-criterion = 9.66, mean failure rate = 0.312; reconstructed from Fleuret et al., 2011) compared to the suboptimal problems (mean trials-to-criterion = 5.15, mean failure rate = 0.07). Furthermore, based on our classification of easy and hard problems in Experiment 1, four of the six optimal problems are deemed "hard" compared to only five of the 15 suboptimal problems. We account for this difference of difficulty in our analyses with a regression term for *problem ID*.

It is not simply coincidental that the optimal problems happen to be more difficult than the suboptimal problems. The optimal problems, by nature of our selection criteria, have less salient object size cues that reflect those roles (Criterion 5). These problems consist of several objects (at least three) that are smaller, with concepts mostly based on noticing sameness of these (small) shapes.

Design and Procedure

The experiment used a 2 (sorted vs. shuffled displays) × 2 (color-mapped vs. uncolored) between-subjects design. Initial displays were identical to those used in Experiment 1 (i.e., all instances were first presented uncolored). Target images were initially presented uncolored for both color-mapped and uncolored conditions in order to control for information in the target image. This aspect of the design ensured that color could not be directly used to classify a novel image.

In the sorted condition, the positive and negative sets were spatially segregated; in the shuffled condition, all previous instances appeared in the same random order in which they had been presented. For the color-mapped condition, the previous instances were colored in the manner described above; in the uncolored condition they continued to appear in black-and-white (as in Experiment 1). Figure 5 depicts an example of a display in the sorted, color-mapped condition. As in Experiment 1, previous instances appeared within a blue or red border that distinguished the positive and negative sets.

In Experiment 2, the order of all problems was randomized (rather than using an easy-to-hard order as in Experiment 1). Each participants received an individual randomized problem order, which was matched across conditions. Specifically, every *i*th participant in each condition shared the same problem order. All other aspects of the procedure were the same as Experiment 1.

Results and Discussion

We first analyzed the data for all problems by fitting linear and logistic mixed-effects regressions to trials-to-criterion and problem failures. The full model for each analysis included *display type* (sorted vs. shuffled), *color type* (color-mapped vs. uncolored), and their *interaction* as fixed effects, as well as *participant ID* and *random problem order seed* as random effects. We attempted to include *problem ID* as a random effect, but the logistic models failed to converge despite 100 million iterations and various optimizers. Since it is desirable to account for differences in problem difficulties, we included *problem ID* as a fixed effect, specifically as a categorical predictor with 23 levels. To simplify the highly parameterized model, we excluded *problem difficulty* (easy vs. hard), as we did not find an interaction with display type in Experiment 1.

Across all problems, there was no interaction between display type and color type: dropping the interaction term in the full model did not result in worse regression error, trials-to-criterion: $\Delta AIC = 2$, $\chi^2(1) = 0.63$, $p = .43$; problem failures: $\Delta AIC = 2$, $\chi^2(1) = 0.0001$, $p = .98$. There was a main effect of display type: sorted displays resulted in improved performance compared to shuffled displays, trials-to-criterion: $t(99.9) = 2.43$, $p = .017$; problem failures: $z = 2.34$, $p = .019$, replicating the finding of Experiment 1.

For uncolored displays, we examined the impact of problem presentation order (random vs. easy-then-hard) on learning by comparing performance in the uncolored condition of this experiment with Experiment 1, which presented problems in an easy-then-hard order. A full model was defined with *experiment* (Experiment 1 vs. Experiment 2) and *problem ID* as fixed effects, and *participant ID* as a random effect. Indeed, the uncolored condition of Experiment 2 showed weaker learning than Experiment 1 across all problems, trials-to-criterion: $t(2362) = 7.32$, $p < .0001$; problem failures: $z = 6.74$, $p < .0001$). This finding suggests that a randomized problem order increases the overall contextual difficulty of category learning relative to easy-then-hard.

Interestingly, results in Experiment 2 did not show a significant main effect of color type: color-mapping across all problems did not enhance learning than uncolored, trials-to-criterion: $t(101) = -0.77$, $p = .44$; problem failures: $z = -0.70$, $p = .49$. However, as outlined in the Method, there is color coding complexity at the individual level of problem items. We distinguished two sets of problems post

Figure 5
Example of a Sorted Display With Color-Mapping Used in Experiment 2

Note. Only images of instances from previous trials were colored; the novel image displayed at the top was black-and-white in both color-mapped and uncolored conditions. See the online article for the color version of this figure.

hoc: *optimal* color-mapped problems, for which we deem the color-mapping procedure to adequately highlight otherwise unclear relational roles, and *suboptimal* color-mapped problems, for which our color codes would not clearly enhance category learning. If this distinction holds, optimal color-coded problems should exhibit a beneficial effect of color coding (otherwise masked by the overall analysis), whereas suboptimal color-coded problems may show no effect (or even impair learning) relative to uncolored versions. To test this possibility, we again fit mixed-effects models to analyze trials-to-criterion and problem failure, but now separately for the optimal and suboptimal problems. The full models were the same as before. Note that the *problem ID* random effect is now particularly important because it accounts for the inherent difficulties of optimal and suboptimal problems, which differ substantially on average (see Materials).

We first report analyses of the optimal problems, then the suboptimal problems. For the optimal problems (Figure 6, left), again there was no interaction between display type and color type, trials-to-criterion: $\Delta AIC = 0.6$, $\chi^2(1) = 0.58$, $p = .45$; problem failures: $\Delta AIC = 2$, $\chi^2(1) < 0.0001$, $p = .98$. However, color coding was more effective than uncolored images, trials-to-criterion: $t(103) = 2.05$, $p = .043$; problem failures: $z = 2.11$, $p = .035$, suggesting that color coding improves performance when it clearly highlights relational roles. Sorted displays yielded trends toward better learning, but did not reliably improve performance compared to shuffled displays, trials-to-criterion: $t(101) = 1.43$, $p = .16$; problem failures: $z = 1.49$, $p = .14$.

The suboptimal problems yielded a complementary pattern of results (Figure 6, right). There was again no interaction between display type and color type, trials-to-criterion: $\Delta AIC = 2$, $\chi^2(1) = 0.56$, $p = .45$; problem failures: $\Delta AIC = 1.8$, $\chi^2(1) = 0.21$, $p = .65$, but color coding did not enhance learning relative to uncolored images, trials-to-criterion: $t(100) = 0.16$, $p = .88$; problem failures: $z = 0.45$, $p = .65$, indicating that color coding does not improve performance when it does not optimally highlight relational roles. The suboptimal problems yielded a reliable advantage for sorted relative to shuffled displays, trials-to-criterion: $t(99.7) = 2.68$, $p = .0087$; problem failures: $z = 2.74$, $p = .0061$.

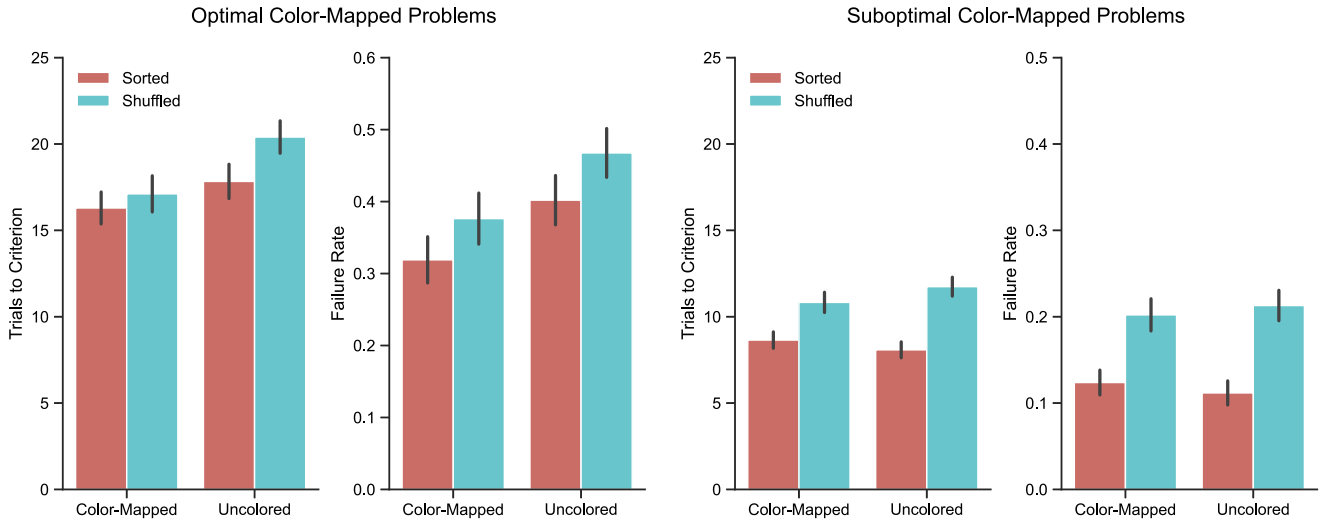
Although a reliable sorted advantage was obtained for the suboptimal problems, the trend was not reliable for the optimal problems. Notably, the mean trials-to-criterion for the optimal problems was numerically closer overall to the ceiling of 34 attempts than for the same problems in Experiment 1. The optimal problems were more difficult (mean trials-to-criterion = 9.66, mean failure rate = 0.312; reconstructed from Fleuret et al., 2011) than the suboptimal problems (mean trials-to-criterion = 5.15, mean failure rate = 0.07). Difficulty may likely have increased due to the fully randomized problem order. Thus, the lack of statistical significance for the sorted effect in the optimal problems could be explained by poor performance for these very challenging problems, combined with the relatively small number of problems included in the set.

Experiment 3

Our proposed explanation for the advantage of color mapping on the optimal problems in Experiment 2 is that the color code highlights relational similarities across images. However, an alternative possibility is that colored stimuli simply increase visual saliency and attention toward individual images without enhancing relational comparisons between images. To assess this alternative, in Experiment 3 we introduced a *color-scrambled* condition in which we randomized the assignment of colors so that the color code did not convey a systematic mapping. If color simply increases visual saliency is some way, the color-scrambled condition would enhance performance; but if color operates by aiding mapping, the color-scrambled condition will not benefit learning, and may in fact impair it.

Figure 6

Trials-to-Criterion and Failure Rates in Experiment 2 for Optimal (Left) and Suboptimal (Right) Color-Mapped Problem Sets



Note. Error bars represent standard errors of the mean (trials-to-criterion) and binomial standard errors (failure rate) at the individual problem level. See the online article for the color version of this figure.

The sorted advantage was not reliable for optimal problems in Experiment 2 perhaps because of the additional overall difficulty created by the fully randomized problem order. In Experiment 3 we used a variant of the easy-then-hard presentation (similar to Experiment 1), aiming to return to a reduced overall difficulty. In addition, we manipulated sorted versus shuffled displays as a within-subject variable, aiming to increase statistical power.

Method

Participants

Experiment 3 employed a more complex design than the first two experiments. The sorted/shuffled factor was manipulated as a within-subjects variable, while color remained between-subjects, with an added third level (color-scrambled). While manipulating sorted/shuffled as a within-subjects factor was expected to reduce variance attributable to individuals, it may not compensate for the increased between-group variance created by adding a third level to color type. To ensure adequate statistical power—particularly for detecting interactions, including a potential three-way interaction between display type, color type, and display order (sorted first vs. shuffled first)—sample sizes were increased for each between-subjects group (three groups vs. four in Experiment 2). A total of 205 University of California, Los Angeles undergraduates participated for course credit (152 female, 49 male, four nonbinary; M_{age} = 20.8). Sample sizes for each of the three between-subjects conditions were: color-mapped (n = 69), color-scrambled (n = 67), and uncolored (n = 69).

Materials

To control for the stimuli used across all conditions, we edited the color-mapped stimuli in Experiment 2. To create the scrambled-color condition, we randomized the assignment of colors to objects (Figure 7). We defined the color coding of each problem as a list of colors of each object. We then shuffled the list 34 times such that for each image in the total set of 34 images for the problem, objects were colored from left to right in the order of the colors of the randomized list.

Design and Procedure

We employed a 3 (color-mapped/color-scrambled/uncolored) × 2 (sorted/shuffled) mixed-factors design, with color type as a between-subjects factor and display type a within-subjects factor. Participants underwent two phases, each based on one display condition (sorted or else shuffled). Within each phase, problems were presented using an easy-to-hard order. Problems were first ordered from easy to hard according to the trials-to-criterion means reported by Fleuret et al. (2011). Then, for every two problems starting with the easiest, one of the two was randomly assigned to the sorted condition and the other to the shuffled condition; this procedure randomly assigned the optimal and suboptimal problems to the first and second phases. The two problems (No. 3 and No. 7) excluded from Experiment 2 (because the number of colors could not be matched between positive and negative sets) were also excluded in Experiment 3. To establish an even split of ten problems for each easy-to-hard list, we also removed the hardest problem (6). This problem happens to be an optimal color-mapped problem, resulting in five optimal problems for Experiment 3. The order of display conditions was counterbalanced, and the ith participant in each color condition received the same random problem assignment to create the easy-to-hard problem lists. Figure 8 depicts an example display for a trial in the color-scrambled condition.

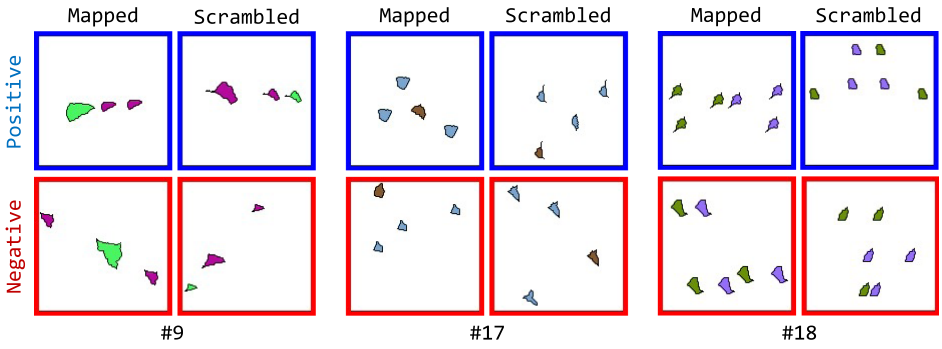
Results and Discussion

We first analyzed the entire set of problems. The full model included *display type* (sorted vs. shuffled), *color type* (color-mapped vs. color-scrambled vs. uncolored), and *experiment phase* (first half vs. second half), along with their *three-way interaction* and all possible *two-way interactions* as fixed effects. Random intercepts were included for *participant ID, problem order seed*, and *problem ID*, and a random slope of *display type* was included for *participant ID*. We removed the random slope of *display type* whenever setting *problem ID* as a fixed categorical predictor did not help the model converge. We continued to exclude *problem difficulty*.

There was no three-way interaction between color, display, and experiment half, and no two-way interactions between these effects (all $p > .10$). There was a main effect of display type, with sorted displays leading to improved learning performance over shuffled

Figure 7

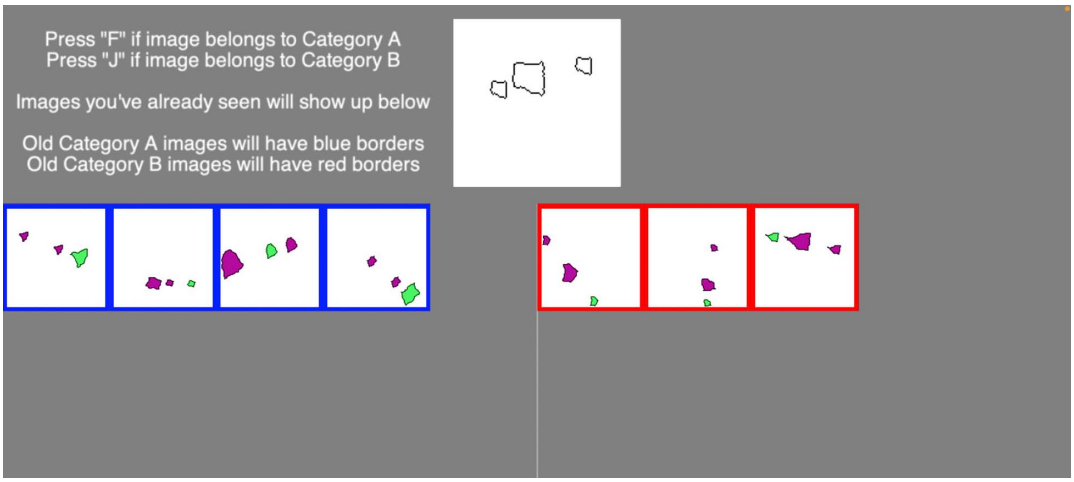
Examples of Color-Scrambled Problems (Experiment 3)



Note. Each group displays a problem with color-mapped and color-scrambled variants. Rows separate positive and negative instances. See the online article for the color version of this figure.

Figure 8

Example of a Sorted Display With Color-Scrambling Used in Experiment 3



Note. Only instances from previous trials were colored; the novel image displayed at the top was black-and-white in all conditions (color-mapped, color-scrambled, and uncolored). See the online article for the color version of this figure.

displays, trials-to-criterion: $t(197) = 3.132$, $p = .0020$; problem failures: $z = 3.288$, $p = .0010$. There was also a main effect of experiment phase for trials-to-criterion only, with the first half being more difficult than the second, $t(210) = 2.21$, $p = .028$; problem failures: $z = 1.17$, $p = .24$. Furthermore, uncolored problems resulted in improved performance over color-scrambled versions, trials-to-criterion: $t(167) = 2.42$, $p = .044$; problem failures: $z = 2.50$, $p = .033$. Similar to the results of Experiment 2, color-mapped performance was not significantly different from uncolored performance, trials-to-criterion: $t(168) = 1.46$, $p = .31$; problem failures: $z = 1.850$, $p = .15$ and even color-scrambled performance, trials-to-criterion: $t(168) = 0.98$, $p = .60$; problem failures: $z = 0.67$, $p = .78$.

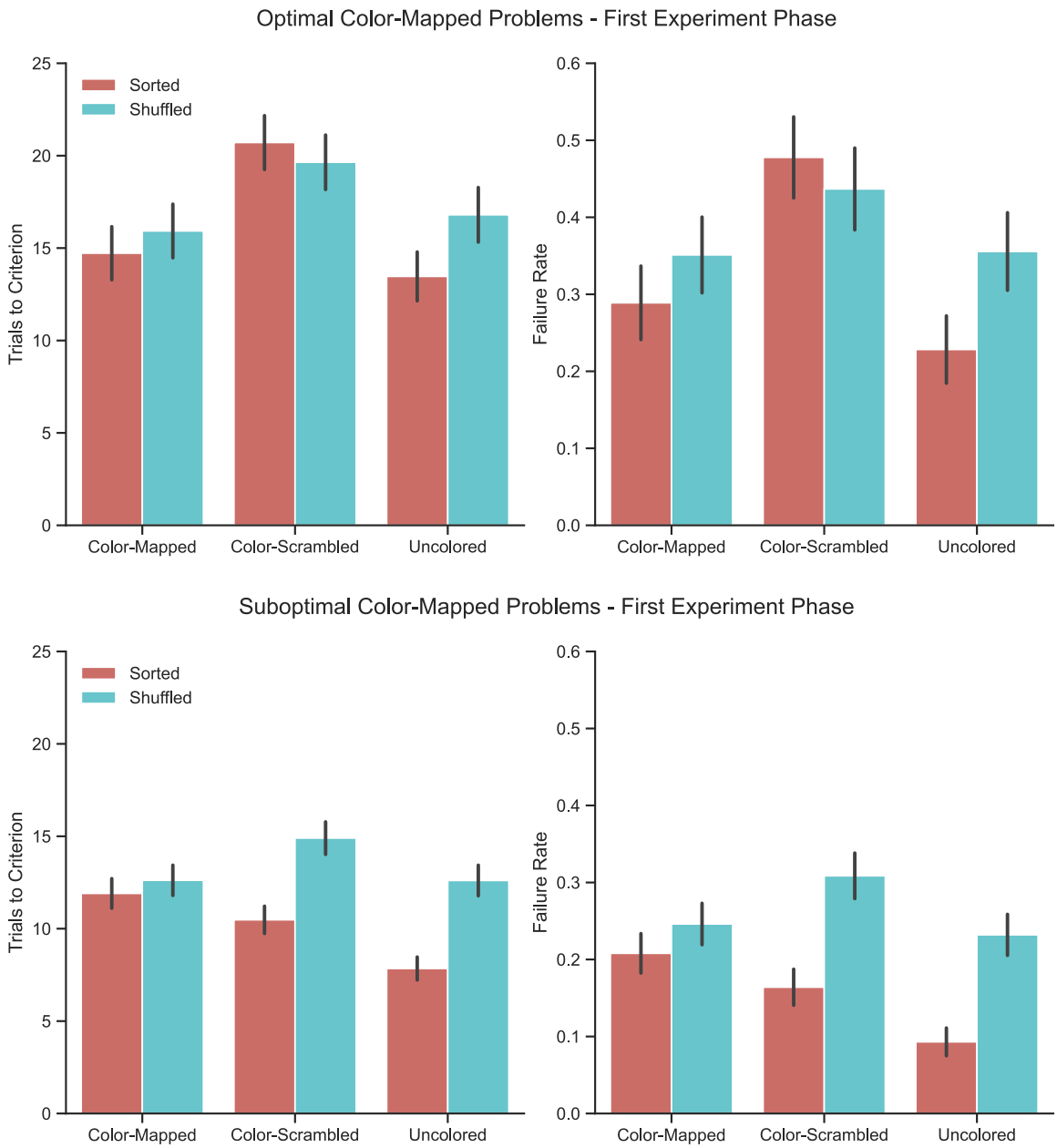
Although performance with uncolored displays was superior to that with color-scrambled displays, the lack of an overall color-mapped advantage suggests the possibility that color-mapping may only be beneficial when it does not hinder or repeat otherwise salient information. We therefore split the problems into an optimal (Figure 9, top) and a suboptimal set (Figure 9, bottom). Although a trend toward a three-way interaction between color, display, and optimality of a problem is apparent, this interaction was not significant, trials-to-criterion: $\Delta AIC = 3$, $\chi^2(2) = 0.63$, $p = .73$; problem failures: $\Delta AIC = 3.4$, $\chi^2(2) = 0.62$, $p = .73$. Nevertheless, given previously stated reasons that effects of color may vary based on our criteria of color coding efficacy, we analyzed the optimal and suboptimal problems separately, as in Experiment 2. We also analyzed the data separately for the two phases, given the difference in learning performance across experiment phases in trials-to-criterion. Such performance differences between the two phases could be driven by multiple factors, such as a carry-over learning effect from the early phase to the later phase, or some kind of disruption in the switch in display formats between the two phases (from sorted to shuffled or vice versa). Hence, our analysis focuses on only the first phase of the optimal and suboptimal problems to avoid confounds with carry-over effects.

Figure 9 reports the first phase results for the optimal problems in the top panel, then for the suboptimal set in the bottom panel. For the optimal problems, there was no interaction between display type and color type, trials-to-criterion: $\Delta AIC = 2.3$, $\chi^2(2) = 1.68$, $p = .43$; problem failures: $\Delta AIC = 2.1$, $\chi^2(2) = 1.87$, $p = .39$. Sorted displays again did not improve performance relative to shuffled displays for the optimal problems, perhaps owing to a ceiling effect, trials-to-criterion: $t(151) = 0.96$, $p = .34$; problem failures: $z = 0.98$, $p = .33$. Color-scrambling led to poorer performance than both color-mapped, trials-to-criterion: $t(151) = -3.53$, $p = .0016$; problem failures: $z = 2.87$, $p = .012$ and uncolored conditions, trials-to-criterion: $t(151) = 3.49$, $p = .0018$; problem failures: $z = 3.29$, $p = .0029$. However, color-mapping did not improve performance compared to uncolored problems, trials-to-criterion: $t(151) = 0.038$, $p = 1.00$; problem failures: $z = 0.47$, $p = .89$. As is apparent from a comparison of the data in Figure 9 (Experiment 3) versus Figure 6 (Experiment 2), learning rate for the optimal color-coded condition was very similar in the two experiments, but learning rate for the uncolored condition was faster in Experiment 3. The easy-to-hard problem order used in Experiment 3 (as compared to the randomized order in Experiment 2) appears to have eliminated the benefit of adding color to support mapping.

The suboptimal problems again depicted a complementary pattern of results. There was no interaction between display type and color type, trials-to-criterion: $\Delta AIC = 2.0$, $\chi^2(2) = 1.79$, $p = .41$; problem failures: $\Delta AIC = 3.2$, $\chi^2(2) = 0.79$, $p = .67$. Sorted displays resulted in improved performance compared to shuffled displays, trials-to-criterion: $t(166) = 2.96$, $p = .0036$; problem failures: $z = 3.19$, $p = .0014$. There was no effect of color coding; all pairwise comparisons of color types were not significant (all $p > .061$). These results reinforce the hypothesis that color coding does not enhance performance when it does not highlight relational roles and suggests

Figure 9

Trials-to-Criterion and Failure Rates for the First Experiment Phases of Optimal Color-Mapped (Top) and Suboptimal Color-Mapped Problems (Bottom) in Experiment 3



Note. Error bars represent standard errors of the mean (trials-to-criterion) and binomial standard errors (failure rate) at the individual problem level. See the online article for the color version of this figure.

that color-scrambling has a similar effect as color-mapping when suboptimal.

General Discussion

A prominent theoretical account of relational category learning emphasizes the importance of analogical mapping—a comparison of compositional representations of exemplars that reveals their shared relational structure (e.g., Doumas et al., 2022; Kurtz et al., 2013; Lovett & Forbus, 2017; Shurkova & Doumas, 2022). Although research suggests that mapping can occur spontaneously even when the comparison is not guided by explicit relational cues (Gentner & Markman, 1997; Markman & Gentner, 1993a, 1993b), it is also cognitively demanding, which may discourage its use in relational category learning when negative exemplars (i.e., those that deviate from the concept) differ from positives only in the

pattern of relations formed by the constituent objects and otherwise share similar object-level features. In the case where the only difference is the presence or absence of a particular relation(s), a less demanding strategy is to encode exemplars as lists of objects and relations without specifying relational roles (Corral et al., 2018). A simple comparison of these lists can differentiate categories, but results in a noncompositional concept (i.e., one that lacks the structure afforded by role-filler bindings). Thus, if analogical mapping underlies relational category learning, learners may only favor it when cognitive demands are low.

A central goal of research on category learning has been to identify task constraints that shape the content and form of conceptual representations. Specifically, it is important to identify the constraints that determine which conceptual representations are compositional in order to foster generalization across domains. Using the SVRT (Fleuret et al., 2011), we explored factors that might impact the ease of mapping between exemplars. Critically, all SVRT problems can be solved with a noncompositional feature-list "shortcut" because the negative images for every problem share the same objects and general visual attributes as the positive images, differing only in the relations formed by the objects. Since all SVRT problems can be solved using a less cognitively demanding representational approach, learners are not forced by the stimuli to acquire compositional representations, which our visual aids aim to facilitate. In the original SVRT study, previously presented exemplars accumulated in the visual display. In three experiments, we manipulated characteristics of this visual record in ways intended to impact the ease of mapping.

Sorted Versus Shuffled Displays of Previous Instances

In all three experiments, previous items either accumulated in the order they were presented (which was randomized; *shuffled* display), or else positive and negative instances were spatially segregated (*sorted* display). Our expectation was that the sorted display would ease the cognitive load of mapping by facilitating comparisons both within the positive set and the negative set, and between the two sets. In each experiment, sorted displays indeed led to overall faster learning, as measured by both trials-to-criterion and failure rates (i.e., trials on which the criterion was not reached). An advantage for sorted displays was observed both when the set of SVRT problems was presented in an easy-to-hard order (Experiments 1 and 3) and when problem order was randomized (Experiment 2). Overall, learning rates for individual problems were faster when the 23 problems were presented in an easy-to-hard order (Experiment 1) rather than a randomized order (Experiment 2); however, the sorted advantage was not reliable for the color-mapped optimal set in Experiment 2. The major exception to the sorted advantage was observed in Experiment 3, when the display type was changed from sorted to shuffled, or vice versa, between the first and second phases of the experiment. No reliable differences between conditions were observed in the second phase, but performance in the second phase improved for one dependent variable and remained consistent for the other. This finding implies that in teaching visuospatial categories from examples, it is important to consider both the consistency of display formats and the potential for learners to adjust or benefit from prior experience.

The contrast between shuffled and sorted displays bears a superficial resemblance to that between distributed versus massed presentation orders (i.e., orders in which instances of different categories are interleaved, vs. orders in which instances of a single category are presented in sequence). Several studies have found an advantage for distributed over massed presentation order in visual category learning (e.g., Carvalho & Goldstone, 2014, 2017; Kang & Pashler, 2012; Kornell & Bjork, 2008). However, the shuffled/sorted distinction in our study applies to the visual record of *previous* instances, rather than the initial presentation order of instances. In all our experiments, for each problem, the initial order of presentation randomly interleaved positive and negative category instances (i.e., was distributed). Few previous studies of category learning have used displays similar to those used here. In the context of science learning, Meagher et al. (2017) found that presenting instances of multiple categories in coherent, spatially organized displays enhanced learning relative to sequential presentation. The present findings illustrate how a distributed presentation order can be combined with a record of prior examples that is spatially organized to segregate positive and negative instances. Future work can investigate whether this combined presentation mode in fact provides the learning advantages of both distributed presentation and an organized spatial record of instances.

Color Coding Images

In Experiments 2 and 3, we also explored the potential impact of color as a cue to mapping of objects across images in the SVRT. Although we are not aware of previous experimental studies that specifically investigated the use of color as a cue to mapping, color is commonly used to highlight mappings in analogies presented in educational settings (Gray & Holyoak, 2021). Our initial aim was to use color to highlight the roles of each object as defined by the relational concept. However, we immediately confronted difficulties due to the fact that a categorization task (unlike a typical mapping between two specific analogs) inherently involves at least *two* categories (in the SVRT, a positive and a negative set of instances). Ideally (to avoid confounding), a color code should satisfy several basic constraints—in particular, using the same colors to represent objects playing the same role, with the specific color assignments held constant across the positive and negative sets. We were only able to satisfy these constraints for six of the 23 SVRT problems, which we dubbed the "optimal" color-mapping set. Two problems had to be set aside because it was not possible to equate the number of colors across the positive and negative instances, and the remaining 15 problems permitted only a "suboptimal" color mapping (e.g., a color mapping that was redundant with size, another salient visual cue).

In general, the impact of color mapping was weak. In Experiment 2 (in which SVRT problems were presented in random order), problems in the optimal set were learned more quickly in the color-mapped than in the standard uncolored condition. However, this advantage of the color-mapped condition was not found in Experiment 3 (in which problems were presented in the more favorable easy-to-hard order). In Experiment 3, for the optimal set, scrambling the colors (so color was not a cue to mapping) impaired learning relative to the color-mapped and uncolored conditions. However, we found no reliable influence of color coding for the suboptimal set in either experiment. It is possible that the influence of color was mitigated because in all conditions, each instance was initially presented uncolored; the colors were only added when the

instance appeared in the visual record of presented instances. This design feature was intended to equate the initial presentation format across conditions and to make it impossible to use color directly to classify newly presented instances. However, having images change from uncolored to colored in the course of a trial may have discouraged participants from attending to the color cues.

Further research is required to explore other possible display characteristics that may impact relational comparisons during visuospatial category learning. The present findings call attention to the need to carefully consider the fact that category learning inherently involves at least two categories: the positive instances, and the negative instances that do not fit the rule defining the positive set. Spatial segregation of positive and negative instances in a visual record is an effective aid to learning because it supports comparisons both within and between the two sets. It is more difficult to design a color mapping that can apply unambiguously to both sets. Other visuospatial cues that might support mapping, such as spatial alignment (Simms et al., 2023), should also be investigated as potential devices to facilitate category learning.

Representational Pluralism

This work explores ways to reduce the cognitive load of analogical mapping in a visual relational category learning task, in which learners may rely on either compositional representations for analogical mapping or simpler feature-based representations for discrimination. While we did not explicitly demonstrate that color coding or sorted displays enhance the use of analogical mapping, our findings suggest that in several cases these visual aids improve learning efficiency. Since these aids help reveal compositional structure, and previous work suggests that spontaneous comparisons of scenes evoke mapping (Gentner & Markman, 1997; Markman & Gentner, 1993a, 1993b), these performance gains may reflect a tendency for learners to opt for mapping as a consequence of reduced cognitive load. It remains possible, however, that learners did not engage in mapping even with visual aids, or that they naturally employed this strategy even without visual aids (albeit less effectively). An informative next step would be to test whether learners encode and recall relational roles differently depending on the constraints imposed during learning. If a learner forms a compositional representation that preserves role information of objects, they should be more likely to recall or recognize the specific relational roles occupied by objects (e.g., identifying which object was the agent vs. recipient in a causal relation). In contrast, if a learner relies on noncompositional encoding based on feature lists, they may recall object identity or category membership without preserving role structure. These alternatives could be assessed using postlearning memory probes or forced-choice recognition tasks that test role binding (e.g., "Which object was the one that caused the other?").

Regardless of the representational format of relational categories, the downstream reasoning mechanism used to induce the concept from the exemplars is likely a kind of comparison process, whether a simple list comparison or the structured kind of analogical mapping. Our findings of learning differences attributable to sorted versus shuffled displays, and color coding versus color-scrambled exemplars, coupled with previous empirical evidence (e.g., Christie & Gentner, 2010; Kittur et al., 2004; Kurtz et al., 2013), consistently demonstrate a performance advantage for conditions in which comparison is employed. In particular, Goldwater et al. (2018) found that blocking trials—presenting exemplars from the same category consecutively—was more effective than interleaving trials—presenting exemplars in an order that mixed categories—for self-reported *rule-based* learners, who tend to identify patterns across examples that distinguish categories. However, self-reported *exemplar-based* learners, who focus on memorizing individual items rather than comparing them, showed no benefit from blocking. Critically, within the blocked condition, performance on relational category learning was positively correlated with a greater tendency toward rule-based learning (on a Self-Reported Scale from exemplar-based to rule-based). These findings—that individuals predisposed to comparison not only benefit more from learning conditions that support it, such as blocked presentation, but also outperform those less inclined to compare under those same conditions—provide compelling evidence that comparison functions as a core mechanism through which relational structures are abstracted.

Evidence for the central role of comparison in category learning challenges the assumptions of current program induction models. Although the creation of programs in program induction models does not *forbid* comparison, existing program induction models typically rely on meta-programs that generate *multiple* candidate concepts (in the form of programs) from a *single* exemplar. In these models, inductive reasoning operates as a selection process: given a large space of programs, which one best captures the exemplar? By contrast, comparison-based approaches produce a *single* concept—rather than a whole set—from at least *two* exemplars. Thus, whether relational category representations are encoded as programs or relational graphs, the process by which they are learned likely depends on comparison—perhaps through a novel form of analogical mapping that operates over programs rather than traditional relational graphs. This possibility echoes a proposal by Christie and Gentner (2010), who suggested that Bayesian inference might rely on analogical mapping to construct candidate concepts.

More broadly, these considerations point to the value of distinguishing between (a) the mechanism of concept formation (e.g., comparison vs. meta-program), (b) the format of the resulting representation (e.g., graph, feature list, or program), and (c) the selection processes that operate when multiple representations are available (e.g., Bayesian inference). Embracing this kind of representational pluralism reframes relational category learning not as the discovery of a single optimal representation and mechanism, but as a coordination problem involving multiple representational possibilities and cognitive strategies. This perspective suggests that learners might dynamically shift between strategies—sometimes engaging in analogical comparison to construct structured representations, other times relying on feature-list discrimination when comparison is difficult (e.g., resource-intensive). This flexibility could account for individual differences, as well as task demands that modulate strategy use. In sum, pluralism allows researchers to ask not just what representation is formed, but how, why, and under what conditions.

References

Asmuth, J., & Gentner, D. (2017). Relational categories are more mutable than entity categories. *The Quarterly Journal of Experimental Psychology*, *70*(10), 2007–2025. https://doi.org/10.1080/17470218.2016.1219752

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*(1), Article 5418. https://doi.org/10.1038/s41467-020-18946-z

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2021). From convolutional neural networks to models of higher-level cognition (and back again). *Annals of the New York Academy of Sciences*, *1505*(1), 55–78. https://doi.org/10.1111/nyas.14593

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481–495. https://doi.org/10.3758/s13421-013-0371-0

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699–1719. https://doi.org/10.1037/xlm0000406

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, *11*(3), 356–373. https://doi.org/10.1080/15248371003700015

Corral, D., Kurtz, K. J., & Jones, M. (2018). Learning relational concepts from within- versus between-category comparisons. *Journal of Experimental Psychology: General*, *147*(11), 1571–1596. https://doi.org/10.1037/xge0000517

Doumas, L. A. A., Puebla, G., Martin, A. E., & Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. *Psychological Review*, *129*(5), 999–1041. https://doi.org/10.1037/rev0000346

Ellis, K., Solar-Lezama, A., & Tenenbaum, J. (2015). Unsupervised learning by program synthesis. *Advances in neural information processing systems*. Curran Associates.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140. https://doi.org/10.1037/0096-3445.127.2.107

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, *18*(4), 500–549. https://doi.org/10.1016/0010-0285(86)90008-3

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., & Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(43), 17621–17625. https://doi.org/10.1073/pnas.1109168108

Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W.-K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (pp. 151–175). American Psychological Association.

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science*, *5*(3), 152–158. https://doi.org/10.1111/j.1467-9280.1994.tb00652.x

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45–56. https://doi.org/10.1037/0003-066X.52.1.45

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1–38. https://doi.org/10.1016/0010-0285(83)90002-6

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*(1), 585–612. https://doi.org/10.1146/annurev.psych.49.1.585

Goldstone, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 86–112. https://doi.org/10.1037/0096-1523.26.1.86

Goldwater, M. B., Don, H. J., Krusche, M. J. F., & Livesey, E. J. (2018). Relational discovery in category learning. *Journal of Experimental Psychology: General*, *147*(1), 1–35. https://doi.org/10.1037/xge0000387

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*(7), 729–757. https://doi.org/10.1037/bul0000043

Gray, M. E., & Holyoak, K. J. (2021). Teaching by analogy: From theory to practice. *Mind, Brain and Education*, *15*(3), 250–263. https://doi.org/10.1111/mbe.12288

Hafri, A., & Firestone, C. (2021). The perception of relations. *Trends in Cognitive Sciences*, *25*(6), 475–492. https://doi.org/10.1016/j.tics.2021.01.006

Halford, G. S., Bain, J. D., Maybery, M. T., & Andrews, G. (1998). Induction of relational schemas: Common processes in reasoning and complex learning. *Cognitive Psychology*, *35*(3), 201–245. https://doi.org/10.1006/cogp.1998.0679

Halford, G. S., & Busby, J. (2007). Acquisition of structured knowledge without instruction: The relational schema induction paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 586–603. https://doi.org/10.1037/0278-7393.33.3.586

Hornsby, A. N., & Love, B. C. (2014). Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*, *3*(2), 72–76. https://doi.org/10.1016/j.jarmac.2014.04.009

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466. https://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*(2), 220–264. https://doi.org/10.1037/0033-295X.110.2.220

Hummel, J. E., Holyoak, K. J., Green, C. B., Doumas, L. A., Devnich, D., Kittur, A., & Kalar, D. J. (2004, October). A solution to the binding problem for compositional connectionism. *AAAI technical report (3)* (pp. 31–34).

Ichien, N., Lin, N., Holyoak, K. J., & Lu, H. (2024). Cognitive complexity explains processing asymmetry in judgments of similarity versus difference. *Cognitive Psychology*, *151*, Article 101661. https://doi.org/10.1016/j.cogpsych.2024.101661

Jee, B. D., Uttal, D. H., Gentner, D., Manduca, C., Shipley, T. F., & Sageman, B. (2013). Finding faults: Analogical comparison supports spatial concept learning in geoscience. *Cognitive Processing*, *14*(2), 175–187. https://doi.org/10.1007/s10339-013-0551-7

Jung, W., & Hummel, J. E. (2015). Making probabilistic relational categories learnable. *Cognitive Science*, *39*(6), 1259–1291. https://doi.org/10.1111/cogs.12199

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103. https://doi.org/10.1002/acp.1801

Kim, J., Ricci, M., & Serre, T. (2018). Not-So-CLEVR: Learning same–different relations strains feedforward neural networks. *Interface Focus*, *8*(4), 20180011. https://doi.org/10.1098/rsfs.2018.0011

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). Erlbaum.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Kurtz, K. J., Boukrina, O., & Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1303–1310. https://doi.org/10.1037/a0031847

Lake, B. M., & Piantadosi, S. T. (2020). People infer recursive visual concepts from just a few examples. *Computational Brain & Behavior*, *3*(1), 54–65. https://doi.org/10.1007/s42113-019-00053-y

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Lee, A. J., Holyoak, K., & Lu, H. (2024). *Enhancing visuospatial mapping in relational category learning*. Open Science Framework. https://doi.org/10.17605/OSF.IO/Q6UAD

Lovett, A., & Forbus, K. (2017). Modeling visual problem solving as analogical reasoning. *Psychological Review*, *124*(1), 60–90. https://doi.org/10.1037/rev0000039

Markman, A. B., & Gentner, D. (1993a). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, *32*(4), 517–535. https://doi.org/10.1006/jmla.1993.1027

Markman, A. B., & Gentner, D. (1993b). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*(4), 431–467. https://doi.org/10.1006/cogp.1993.1011

McLaren, I. P. L., & Suret, M. B. (2000). Transfer along a continuum: Differentiation or association? In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the twenty-second annual conference of the Cognitive Science Society* (pp. 340–345). Cognitive Science Society.

Meagher, B. J., Carvalho, P. F., Goldstone, R. L., & Nosofsky, R. M. (2017). Organized simultaneous displays facilitate learning of complex natural science categories. *Psychonomic Bulletin & Review*, *24*(6), 1987–1994. https://doi.org/10.3758/s13423-017-1251-6

Messina, N., Amato, G., Carrara, F., Gennaro, C., & Falchi, F. (2021). Solving the same-different task with convolutional neural networks. *Pattern Recognition Letters*, *143*, 75–80. https://doi.org/10.1016/j.patrec.2020.12.019

Messina, N., Amato, G., Carrara, F., Gennaro, C., & Falchi, F. (2022). Recurrent vision transformer for solving visual reasoning problems. *Image Analysis and Processing—ICIAP 2022: 21st international conference, Lecce, Italy, May 23–27, 2022, proceedings, part III* (pp. 50–61). https://doi.org/10.1007/978-3-031-06433-3_5

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, *5*(3), 345–369. https://doi.org/10.3758/BF03208813

Phillips, S., Takeda, Y., & Sugimoto, F. (2016). Why are there failures of systematicity? The empirical costs and benefits of inducing universal constructions. *Frontiers in Psychology*, *7*, Article 1310. https://doi.org/10.3389/fpsyg.2016.01310

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, *21*(1), 1–17. https://doi.org/10.1017/S0140525X98000107

Shurkova, E., & Doumas, L. A. A. (2022). Toward a model of visual reasoning. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th annual meeting of the Cognitive Science Society*. Cognitive Science Society.

Simms, N. K., Matlen, B. J., Jee, B. D., & Gentner, D. (2023). Spatial alignment supports comparison of life science images. *Journal of Experimental Psychology: Applied*, *29*(4), 747–760. https://doi.org/10.1037/xap0000471

Turini, J., & Võ, M. L. H. (2022). Hierarchical organization of objects in scenes is reflected in mental representations of objects. *Scientific Reports*, *12*(1), Article 20068. https://doi.org/10.1038/s41598-022-24505-x

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Waltz, J. A., Lau, A., Grewal, S. K., & Holyoak, K. J. (2000). The role of working memory in analogical mapping. *Memory & Cognition*, *28*(7), 1205–1212. https://doi.org/10.3758/BF03211821

Winston, P. H. (1975). Learning structural descriptions from examples. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 157–209). McGraw Hill.

Zaki, S. R., & Salmi, I. L. (2019). Sequence as context in category learning: An eyetracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(11), 1942–1954. https://doi.org/10.1037/xlm0000693

Received October 3, 2024
Revision received May 5, 2025
Accepted May 27, 2025 ■