Human Relational Concept Learning on the Synthetic Visual Reasoning Test

Andrew J. Lee[1]
andrewlee0@g.ucla.edu

Hongjing Lu[1,2]
hongjing@ucla.edu

Keith J. Holyoak[1]
holyoak@lifesci.ucla.edu

[1]Department of Psychology
[2]Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

Humans exhibit a remarkable ability to learn relational concepts from a small number of examples. On the Synthetic Visual Reasoning Test (SVRT), a collection of 23 problems that require learning relational concepts, people typically discover the relational rules from a handful of examples. An important question is what learning mechanisms underlie the human ability to acquire relational concepts so quickly. Previous work has demonstrated that comparison of examples via analogical mapping underlies rapid relational concept acquisition. Here, we examine whether learners switch to learning strategies that do not involve comparison when cognitive load is high. We conducted two experiments that varied the display format and problem order for the SVRT. When problems are presented in an easy-to-hard order, people learn more efficiently when prior examples are displayed in spatially segregated sets, consistent with the use of analogical mapping as a learning strategy. However, when the problems are presented in a random order, the advantage of spatially segregated displays is eliminated. We propose that when hard problems are encountered early in a problem sequence, analogical mapping becomes too demanding, causing people to fall back on a less efficient learning strategy that does not require the comparison of multiple examples.

Keywords: relations; concepts; learning; analogical comparison; mapping; abstraction

Introduction

The rapidity of human concept learning is particularly apparent for concepts that are primarily defined by *relations* between entities, rather than solely by attributes of individual entities. Many everyday concepts are defined by relational structures connecting entities (Gentner & Kurtz, 2005; Asmuth & Gentner, 2017; Goldwater & Schalk, 2016). For example, a "barrier" is something that prevents the achievement of some goal. Different instances of relational concepts can be highly variable in their attributes (e.g., a barrier could be a roadblock or poverty). Learning such concepts requires identifying shared relational structures connecting objects, rather than focusing solely on features of individual objects (Corral, Kurtz, & Jones, 2018).

A relatively simple laboratory task that involves learning relational concepts is the Synthetic Visual Reasoning Test (SVRT). This task (see Figure 1) consists of a set of 23 categorization problems, for each of which the goal is to correctly sort novel images into those that fit a particular category versus those that do not (Fleuret et al., 2011).

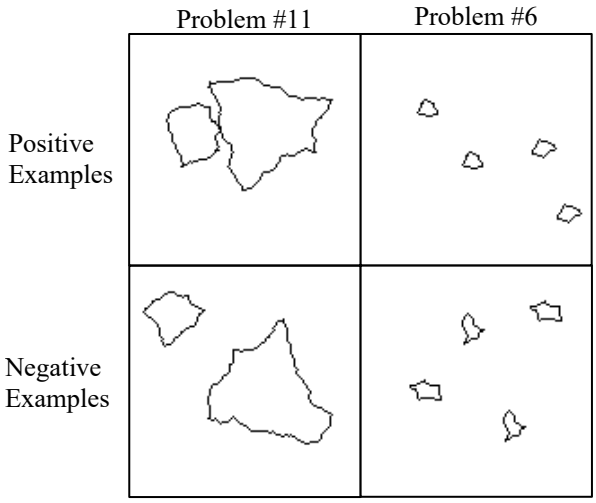

Figure 1: Examples of SVRT problems #11 and #6, respectively the easiest and hardest problems based on human performance from Fleuret et al. (2011). Top: positive examples of the categories. Bottom: negative examples of the categories.

Categories in SVRT problems are defined by visuospatial relations between shapes (e.g., *inside-of*, *larger-than*). Although SVRT images are perceptually simple, the spatial relations underlying a category can be subtle. Humans can nonetheless solve many SVRT problems from a handful of examples, whereas models that solved the ImageNet challenge (Krizhevsky, Sutskever, & Hinton, 2017) require several hundreds of thousands of SVRT training examples, and for some problems, fail to generalize to a similar task (Kim, Ricci, & Serre, 2018; Messina et al., 2021).

The deep learning models that have been applied to the SVRT are trained in an end-to-end fashion from pixel-level inputs of images; no prior knowledge of visual features or relations is assumed. But for simple geometric forms of the sort used in the SVRT, people likely come equipped with basic representational elements, including both features of objects (e.g., size, shape) and basic visuospatial relations. Several models developed in cognitive science, each equipped with such building blocks, suggest ways in which people might learn SVRT concepts from relatively few examples. Here we will consider three general approaches.

The first and perhaps simplest possibility is that people may adopt a learning mechanism based on accumulation of

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

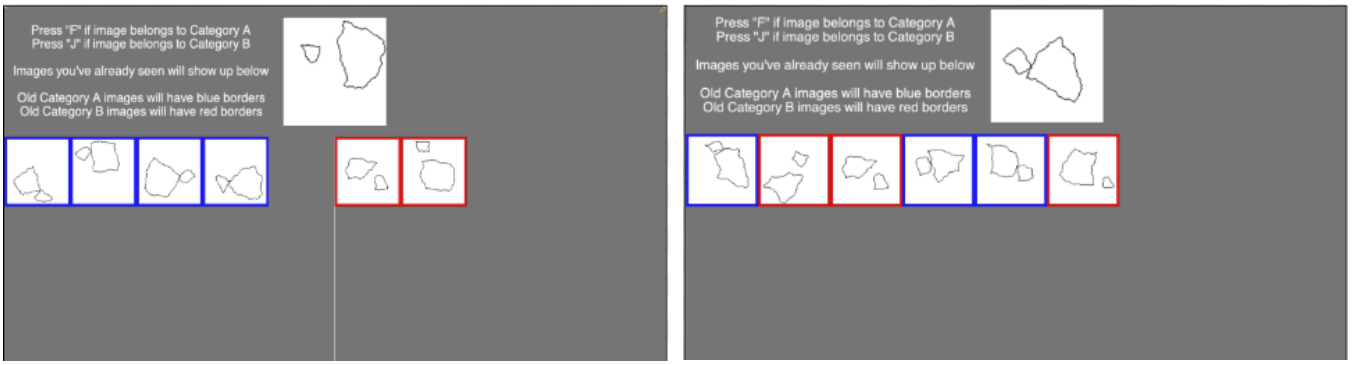

Figure 2: Left: sorted display in which previous instances are separated into positive vs. negative examples (blue or red frame). Right: shuffled display in which previous instances are intermixed in a randomized presentation order.

information about the statistical associations between features of objects and category labels, in combination with hypothesized rules and storage of individual exemplars (e.g., Erickson & Kruschke, 1998; Nosofsky & Palmeri, 1998). Although such *statistical learning models* have not been directly applied to the SVRT, they have been used to successfully predict human data on the acquisition of concepts defined by simple visual forms. It seems reasonable that a visual relation, such as *A has same shape as B*, could serve as the basis of a rule to predict category membership based on statistical associations.

A second possible approach that can achieve rapid concept learning is *program synthesis* (combined with Bayesian inference), in which representations of concepts are similar to computer programs that can each reproduce a concept to varying degrees of success. By iteratively combining and rearranging a few basic functions, (assumed to be available prior to the concept learning task), the program synthesis approach can generate a whole space of possible concept representations (Ellis, Solar-Lezama, & Tenenbaum, 2015). After narrowing this space via Bayesian inference, program synthesis can—with sometimes as little as one example—recreate handwritten characters (Lake, Salakhutdinov, & Tenenbaum, 2015), causal structures (Lake & Piantadosi, 2019), and visuospatial concepts including those used in the SVRT (Ellis et al., 2015).

A third possible approach, which focuses most directly on relational representation, involves learning concepts by *analogical mapping*. Analogical mapping—the process of identifying relational correspondences between examples—is most often considered as a mechanism for transferring knowledge from one domain to another. However, mapping can also serve as a mechanism for induction, as comparison can induce an abstraction of shared relational structures that guides subsequent transfer (Gick & Holyoak, 1983). A number of computational models have used analogical mapping as a guide for visual concept induction (e.g., McLure, Friedman, & Forbus, 2010; see Forbus, Ferguson, Lovett, & Gentner, 2017). At least one learning model based on analogical mapping has been applied to the SVRT problems (Shurkova & Doumas, 2022).

A crucial distinction between analogical mapping and both the statistical approach and program synthesis is that mapping depends on *explicit* comparison of one example to another, whereas the other two approaches operate by processing each individual example sequentially. There is evidence that humans learn to discriminate different visual categories by *selectively attending* to features of a concept that are indicative or diagnostic of category membership (Rehder & Hoffman, 2005; Zaki & Salmi, 2019). Analogical mapping between positive examples of a category (within-category comparisons) can focus attention on shared relations, whereas mapping a positive example to a negative "near miss" that lacks a single critical relation (between-category comparisons) can similarly focus attention to a relation necessary for category membership (Winston, 1975).

Of course, neither the statistical approach nor program synthesis strictly prohibit comparison-based learning. In fact, for at least one model of the statistical approach, a limited form of comparison is assumed (SAT-M; Carvalho & Goldstone, 2022). A key finding in work on concept learning is that selective attention is a product of discovering similarities or differences between recently seen examples, depending on the order in which they are presented (Carvalho & Goldstone, 2014; Zaki & Salmi, 2019). In *interleaved* orderings, unique differences between categories become salient to the observer, facilitating discovery of category boundaries, whereas *blocked* orderings highlight same-category similarities and reveal category-specific information (Carvalho & Goldstone, 2014, 2017). To account for these learning differences between sequence types, the model proposed by Carvalho and Goldstone (2022) differentially weights the encoding strengths of an example's features based on similarities and differences to features of the preceding example.

In analogical mapping, in contrast, two presented examples are compared via the formation of one-to-one relational correspondences that reveal shared structure. In mapping models, relations assume a distinct representational status from their arguments, traditionally in the form of role-filler bindings (e,g., Falkenhainer, Forbus, & Gentner, 1989; Hummel & Holyoak, 1997). Models of statistical learning typically do not separate relational from non-relational information when obtaining features of concepts, resulting in the use of relationally-entangled representations of concepts. Such entangled features are often expressed as a

multidimensional vector, which makes comparison of relations implicit in an overall calculation of distance.

Although analogical mapping can foster acquisition of relational categories (Halford, Bain, Maybery, & Andrews, 1998; Halford & Busby, 2007; Christie & Genter, 2010; Kurtz, Boukrina, & Gentner, 2013; Jung & Hummel, 2015), there is also evidence that the mapping process places a considerable burden on working memory and related executive processes (e.g., Waltz, Lau, Grewal, & Holyoak, 2000; Philips, Takeda, & Sugimoto, 2007). It is possible that people have multiple strategies for learning relational categories and will be more likely to use analogical mapping when the learning situation imposes less cognitive load. However, it remains unclear whether people can learn relational categories using alternative strategies that do not involve explicit comparison of relational structure (Corral et al., 2018; Goldwater, Don, Krusche, & Livesey, 2018). In addition, it is unclear whether people can switch learning strategies in response to changes in cognitive load during the course of learning.

The SVRT is a promising testbed for probing these questions, as the standard method for administering the 23 learning problems includes a procedure that seems likely to aid in comparing examples. As illustrated in Figure 2 left, the display used by Fleuret et al. (2011) maintained a visual record of all the examples previously presented, with positive and negative examples sorted into two spatially segregated groups that appear below the example presented on the current trial. This *sorted* display format likely encourages analogical comparisons between positive examples (which appear together) to extract common relational structures, similar to *blocked* sequences which encourage within-category comparison (Corral et al., 2018). However, like *interleaved* sequences, sorted displays may also support comparisons between positive and negative examples to differentiate the relational structures involved in each category (which although spatially separated, are each grouped to make systematic comparisons relatively easy). Thus, sorted displays may facilitate a systematic combination of within- and between-category comparisons.

To determine whether the display format may in fact impact learning on the SVRT, we performed two experiments in which the cumulative record of previous examples was either sorted (as in the original study) or *shuffled*, with examples recorded in the same random order as that in which they had been presented. If people use analogical mapping to learn the concepts, acquisition should be more efficient when examples are sorted rather than shuffled. Both experiments test the hypothesis that sorted displays facilitate rapid learning, while Experiment 2 varies another procedural factor—the order of the 23 problems with respect to their difficulty—that seems likely to influence cognitive load. When the learning situation imposes greater cognitive load by introducing difficult problems toward the start of the experiment, participants may forgo analogical mapping as a learning strategy, in which case the advantage of sorted displays may disappear.

Experiment 1

To discriminate between analogical mapping and learning mechanisms that do not involve comparison, we modified the original SVRT paradigm. For each individual SVRT problem, participants were presented with a series of trials in which positive and negative examples of the to-be-learned category were presented, one at a time in random order. On each trial, participants classified the novel instance into one of two categories defined by negative and positive examples, after which they received feedback.

Crucially, as participants viewed each novel instance, they also continued to see all the instances shown on previous trials. In Experiment 1, we displayed these instances in one of two spatial organizations. In a *sorted* display (Figure 2 left), the examples are segregated into two sets, with positive examples on the left and negative examples on the right (the same display type used in the original study by Fleuret et al., 2011). In a *shuffled* display (Figure 2 right), the examples appear in the same random order in which they had been presented. In both displays each example was shown with a colored border (blue or red) that distinguished positive from negative instances.

Although the information provided by the sorted display was redundant given the color coding, it seems likely the spatial grouping makes it easier to perform systematic analogical mappings between examples from either the same category (within a spatially-defined set) or from different categories (across sets). When the display is instead shuffled, with all previous instances randomly intermixed on the screen, comparisons are expected to be more difficult and less systematic. Shuffled displays do enhance between-category comparisons which reveal category differences, but such differences are meaningful only against the backdrop of a common relational structure (i.e., alignable differences); discovering a common relational structure is more likely facilitated by sorted displays. The analogical mapping hypothesis, therefore, predicts that sorted displays will lead to faster concept learning. In contrast, approaches that do not involve comparison predict that the two displays will lead to equivalent learning rates.

Previous work has established that the 23 different SVRT problems vary in overall difficulty (Fleuret et al., 2011). Based on human results reported for individual problems, we divided the problems into two subsets, using a natural break based on overall difficulty, to form a set of 13 easy and 10 hard problems. In accord with evidence that in general an "easy-to-hard" ordering of problems supports more efficient overall learning (Pashler & Mozer, 2013), the easy subset of problems was presented before the hard subset.

Method

Participants 64 undergraduates from the University of California, Los Angeles (UCLA) participated for course credit (46 female, 18 male; mean age = 20.1). Sample sizes were equal for the two display conditions (32 each).

Materials, Design, and Procedure The SVRT is a collection of 23 concept learning problems, each of which consists of two categories: one defined by common spatial relations and the other defined by negative examples that do not instantiate those relations. Participants were not informed that one of the categories is defined by negative examples. They were instructed to categorize novel instances into either category A (always the positive examples) or category B (always the negative examples) by pressing "f" or "j" on the keyboard, respectively. Participants received a maximum of 34 novel instances per problem (17 positive, 17 negative).

On each trial, a novel instance, chosen randomly from either category, was presented on the screen. After a categorization decision was made (without speed pressure), feedback was presented for 1s ("Correct!" or "Incorrect!"). The current instance then moved to the bottom of the screen, with a smaller image size of 0.64 the original width, surrounded by a colored frame to distinguish categories (blue for category A, red for category B). In a sorted display, the novel instance appeared either on the left (positive examples) or right (negative examples), separated by a white line. In a shuffled display, previously encountered instances accumulated in order from left to right. In both conditions, no more than 10 previously encountered instances accumulated in each row; if necessary, a second row was added below the first. Previous instances were juxtaposed right next to each other to maximize the size of each image.

Half the participants were randomly assigned to a sorted display, whereas the other half were assigned to a shuffled display. All participants first solved the set of 13 easy problems, randomized in order, and then the 10 hard problems, also randomized in order. (Participants were not told the order of the problems.) Presentation of examples for each problem continued until the participant reached a criterion of 7 correct in a row, or until a maximum of 34 instances had been shown. If the problem was a failure (criterion not reached), then trials to criterion was set to the maximum value of 34. Otherwise, trials to criterion was set to the total number of trials in the problem minus 7, so that the 7 correct in a row did not count toward trials to criterion.

Results and Discussion

For each problem, two dependent variables were measured: trials to criterion (the number of trials before achieving a criterion of 7 correct in a row), and proportion of failures (criterion not achieved within the maximum allotment of 34 learning trials). For data analyses, mean trials to criterion was obtained by averaging each participant's trials to criterion separately for easy and hard problems. For the failure measure, we first summed each participant's number of failures to obtain a total number of failures for easy problems and a total number of failures for hard problems. We then normalized both sums by dividing each by the total number of problems (13 for easy, 10 for hard). Finally, we averaged across participants' mean trials to criterion and proportion of failures, separately for easy and hard problems and for each display condition. Note that lower trials to criterion and lower proportion of failures indicate better learning performance. The resulting means for each dependent measure and condition are depicted in Figure 3.

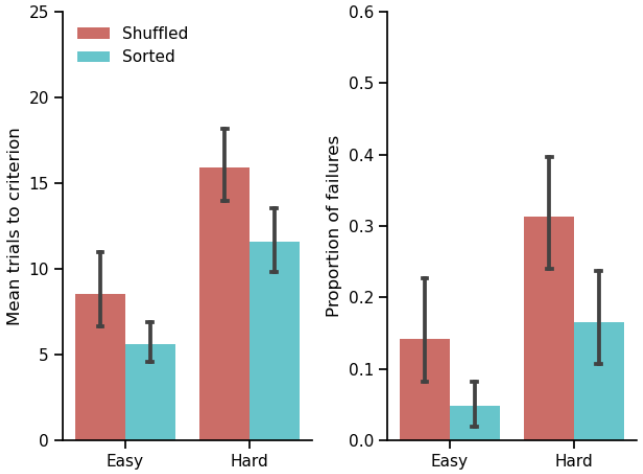

Figure 3: Learning performance in Experiment 1. Lower trials to criterion and lower proportion of failures indicate better learning performance. Sorted displays yielded better learning performance than shuffled displays, with lower trials to criterion and a smaller proportion of failures. Error bars represent 95% confidence intervals of the mean. Trials to criterion do not include the 7 correct in a row to achieve criterion.

We used trials to criterion and proportion of failures as dependent measures in two separate mixed-factors ANOVAs with a between-subjects factor (sorted vs. shuffled) and a within-subjects factor (easy vs. hard problems). These analyses revealed a main effect of display type: for the sorted displays, trials to criterion was lower (M = 8.61; $F(1, 62)$ = 7.92, $p < 0.01$) with a lower proportion of failures (M = 0.11; $F(1, 62)$ = 7.92, $p < 0.01$) than for shuffled displays (trials to criterion: M = 12.21; proportion of failures: M = 0.23). As expected, there was also a main effect of easy/hard difficulty: compared to easy problems (trials to criterion: M = 7.07; proportion of failures: M = 0.095), hard problems led to higher trials to criterion (M = 13.75; $F(1, 62)$ = 154.89, $p <$ 0.001) and higher proportion of failures (M = 0.24; $F(1, 62)$ = 58.05, $p < 0.001$). The interaction between sorted/shuffled display and easy/hard problems was not significant for either trials to criterion ($F(1, 62)$ = 1.77, p = 0.19) or proportion of failures ($F(1, 62)$ = 1.97, p = 0.17).

An interaction effect based on magnitudes of scores may not reflect whether the advantage of sorted displays is different between easy and hard problems, since a difference of a few trials/failures reflects a smaller learning disparity for harder problems. Accordingly, we normalized the sorted condition by the scores on the shuffled display, so that for each problem the mean scores of the sorted condition were divided by the corresponding mean score of the shuffled condition. An independent samples t-test revealed no difference in standardized trials to criterion of sorted displays between easy problems (M = 0.65) and hard problems (M = 0.71; $t(21)$ = 0.78, p = 0.45), and no difference in

standardized proportion of failures of sorted displays between easy problems (M = 0.30) and hard problems (M = 0.46; $t(21) = 1.29$, $p = 0.21$).

Consistent with a comparison-based learning strategy, learning in Experiment 1 was faster and more successful given sorted rather than shuffled spatial displays of accumulated examples. This pattern supports the hypothesis that analogical mapping mediates human relational concept learning on the SVRT.

Experiment 2

Experiment 1 used an easy-to-hard ordering of the SVRT problems. It seems probable that this type of ordering may aid learning overall, because on the easy early trials people will be led to focus on individual relations (e.g., same versus different shapes) that will later be relevant on harder problems involving greater visual complexity (Pashler & Mozer, 2013).

To assess the generality of the influence of spatial organization on learning that we observed in Experiment 1, in Experiment 2 we explicitly varied the order in which the 23 SVRT problems were administered. The ordering was either fixed from easiest to hardest problem based on the data reported by Fleuret et al. (2011), or fully randomized for each participant. We predicted that the easy-to-hard ordering would lead to more efficient learning overall. Moreover, it is possible that when the order is randomized, so that people often encounter hard problems early, analogical mapping may be discouraged due to early failures on problems for which the mapping is complex. If so, it is possible that the spatial organization of the accumulated examples will have less impact when problem order is randomized, because people will be less likely to use analogical mapping as their primary learning strategy.

Method

Participants 125 UCLA undergraduates participated for course credit (94 female, 26 male, 4 nonbinary, 1 declined to answer; mean age = 20.2). Sample sizes for each of the four between-subjects conditions were: sorted/easy-to-hard (n = 32), shuffled/easy-to-hard (n = 33), sorted/randomized (n = 29), shuffled/ randomized (n = 31).

Materials, Design, and Procedure The methodology was nearly identical to that of Experiment 1, except participants received either a fixed easy-to-hard ordering of problems, or else a fully randomized sequence. Participants were not told the order of the problems they would encounter.

Results and Discussion

Data were scored in the same way as in Experiment 1 (see Figure 4). ANOVAs with two between-subjects factors (sorted vs. shuffled; easy-to-hard vs. random order) revealed a two-way interaction between displays (sorted/shuffled) and problem order (easy-to-hard/randomized) for both dependent measures: trials to criterion ($F(1, 121) = 4.67$, $p = 0.033$) and proportion of failures ($F(1, 121) = 4.62$, $p = 0.034$).

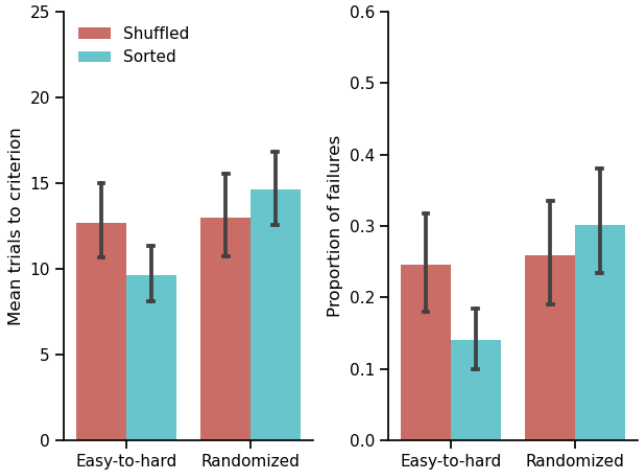

Figure 4: Learning performance in Experiment 2. The influence of problem order interacted with sorted/shuffled display type. Error bars represent 95% confidence intervals of the mean.

Participants showed better learning performance for the easy-to-hard order (trials to criterion: M = 11.19; proportion of failures: M = 0.19) than randomized order (trials to criterion: M = 13.78; proportion of failures: M = 0.28), with a main effect of problem order for both trials to criterion ($F(1, 121) = 5.83$, $p = 0.017$) and proportion of failures ($F(1, 121) = 6.61$, $p = 0.011$). However, there was no main effect of sorted versus shuffled displays (trials to criterion: $F(1, 121) = 0.38$, $p = 0.54$; proportion of failures: $F(1, 121) = 0.85$, $p = 0.36$).

Tests of simple effects revealed that for participants who received the easy-to-hard problem order, sorted displays led to lower trials to criterion (M = 9.65) and reduced proportion of failures (M = 0.14) relative to shuffled displays (trials to criterion: M = 12.69; $F(1, 121) = 4.03$, $p = 0.047$; proportion of failures: M = 0.25; $F(1, 121) = 4.92$, $p = 0.028$). These findings replicate the pattern observed in Experiment 1, which also used easy-to-hard problem orders.

In contrast, when problem order was fully randomized, no advantage was obtained for sorted (trials to criterion: M = 14.66; proportion of failures: M = 0.30) versus shuffled displays (trials to criterion: M = 12.97; $F(1, 121) = 1.14$, $p = 0.29$; proportion of failures: M = 0.26; $F(1, 121) = 0.72$, $p = 0.40$). Thus, sorted displays facilitated learning only when problems were presented in the easy-to-hard order.

General Discussion

The two experiments reported here investigated the impact of alternative spatial displays of accumulated examples on efficiency of visual concept learning in the SVRT. Analogical mapping, unlike either a statistical learning approach or program synthesis, predicts that efficiency will be higher when displays sort positive and negative examples into spatially segregated subsets, facilitating systematic comparisons. An advantage of sorted over shuffled displays was indeed found when problems were presented in an easy-to-hard ordering (Experiment 1, and the comparable condition in Experiment 2). However, when the problem

order was fully randomized so that hard problems were possibly encountered early in the sequence (Experiment 2), learning was less efficient overall and the sorting advantage was eliminated.

The interaction we observed in Experiment 2 between display organization and problem order is consistent with the possibility that people have multiple potential strategies for learning relational concepts. Analogical mapping, which is known to create a high cognitive load, is more likely to be recruited consistently when problems are ordered easy-to-hard. In this situation, on easy early trials mapping is likely to succeed in both solving the problem and in identifying specific relations that will be relevant for later, more complex problems. When mapping is used consistently, sorted displays are useful in guiding systematic comparisons of individual examples.

In contrast, when the problem order is fully randomized, analogical mapping is likely to fail on some hard problems that are presented early. The mapping strategy may then be abandoned, in which case sorted displays no longer convey an advantage. Rather than comparing examples, as required for analogical mapping, people may elect to use a learning strategy that focuses on individual examples. Previous work has also found evidence that people can be oriented toward different learning strategies for relation-based category learning (Goldwater et al., 2018). Although the present study does not identify what alternative strategy may have been used when problem order was randomized, either a statistical approach or program synthesis are viable possibilities. Future work should explore these possibilities. Whatever the exact nature of the alternative strategy, it reduces the overall efficiency of learning relative to the mapping strategy.

Another useful direction for future research would be to use eye-tracking methods to provide more detailed analyses of how people perform comparisons with sorted versus shuffled displays, as has been done in similar work on interleaved and blocked sequences (Zaki et al., 2019). Investigating the aspects of sorted and shuffled displays that impact learning may clarify their relationship to the seemingly-related distinction between interleaved and blocked sequences of examples. Do sorted displays combine the strengths of both sequence types by systematically facilitating both within- and between-category comparisons? Do people make frequent short-distance saccades within a category of examples to first discover relational structure, and then shift over to the other category to locate critical differences? Do shuffled displays reduce comparison overall by requiring longer-distance eye movements?

In sum, the current study provides preliminary evidence that analogical mapping may underlie rapid relational concept learning in humans, at least when problems are presented in ways that foster systematic comparisons between examples while minimizing cognitive load. Further work is required to probe the learning mechanisms that allow humans to learn concepts defined by visual relations from modest amounts of training data.

Acknowledgements We thank Rishi Deorah, Ziqi Zheng, Zixuan Zhou, Ashley Choy, and Yining Liang for data collection and helpful discussions, and also Nick Ichien for his insightful comments. Preparation of this paper was supported by NSF Grant IIS-1956441 awarded to H.L.

References

Asmuth, J., & Gentner, D. (2017). Relational categories are more mutable than entity categories. *Quarterly Journal of Experimental Psychology, 70*(10), 2007-2025.

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.

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.

Carvalho, P. F., & Goldstone, R. L. (2022). A computational model of context-dependent encodings during category learning. *Cognitive Science*, *46*(4), e13128.

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

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.

Ellis, K., Solar-Lezama, A., & Tenenbaum, J. (2015). Unsupervised learning by program synthesis. *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107-140.

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

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*, *USA*, *108*(43), 17621-17625.

Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*, *41*(5), 1152-1201.

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.

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

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.

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*(7), 729-757.

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.

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.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427.

Jung, W., & Hummel, J. E. (2015). Making probabilistic relational categories learnable. *Cognitive Science, 39*(6), 1259-1291.

Kim, J., Ricci, M., & Serre, T. (2018). Not-So-CLEVR: learning same–different relations strains feedforward neural networks. *Interface Focus*, *8*(4), 20180011.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

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.

Lake, B. M., & Piantadosi, S. T. (2020). People infer recursive visual concepts from just a few examples. *Computational Brain & Behavior*, *3*, 54-65.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332-1338.

McLure, M., Friedman, S., & Forbus, K. (2010). Combining progressive alignment and near-misses to learn concepts from sketches. *Proceedings of the 24th International Workshop on Qualitative Reasoning*. Portland, OR.

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.

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.

Pashler, H., & Mozer, M. C. (2013). When does fading enhance perceptual category learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1162-1173.

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*. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01310/full.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology, 51*(1), 1–41.

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.

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

Winston, P. H. (1975). Learning structural descriptions from examples. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 157-209). New York: 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.