Similarity-based Ordering of Instances for Efficient Concept Learning

Erik Weitnauer; Paulo F. Carvalho; Robert L. Goldstone
({eweitnau,pcarvalh,rgoldsto}@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E 10th St
Bloomington, IN 47405 USA

Helge Ritter (helge@techfak.uni-bielefeld.de)
CITEC, Bielefeld University, Universitätsstr. 21-23,
33615 Bielefeld, Germany

Abstract

Theories in concept learning predict that interleaving instances of different concepts is especially beneficial if the concepts are highly similar to each other, whereas blocking instances belonging to the same concept provides an advantage for learning low-similarity concept structures. This suggests that the performance in concept learning tasks can be improved by grouping the instances of given concepts based on their similarity. To explore this hypothesis, we use Physical Bongard Problems, a rich categorization task with an open feature space, to analyze the combined effects of comparing dissimilar and similar instances within and across categories. We manipulate the within- and between-category similarity of instances presented close to each other in blocked, interleaved and simultaneous presentation schedules. The results show that grouping instances to promote dissimilar within- and similar between-category comparisons improves the learning results, to a degree depending on the strategy used by the learner.

Keywords: category learning; order effects; similarity

Introduction

In inductive learning, one abstracts from trained examples to derive a more general characterization that can lead to both seeing the familiar in new and the new in familiar situations. One particularly powerful technique for inductive learning of difficult, relational concepts is the comparison of multiple cases (Kurtz, Boukrina, & Gentner, 2013; Loewenstein & Gentner, 2005; Gick & Holyoak, 1983). The benefit of comparison goes beyond establishing similarities between the inputs, it frequently promotes noticing the commonalities and differences between the compared instances, which helps constructing useful generalizations and can change one's representation and understanding of what is compared (Hofstadter, 1996; Medin, Goldstone, & Gentner, 1993; Mitchell, 1993). In this paper, we look into how the type and order comparisons influences category learning.

In a category learning setting in which category labels are provided, two important types of comparisons are possible: comparisons between instances from the same concept and comparisons of instances from different concepts. The existing research literature does not only suggest different roles for those two types of comparisons, but also makes different predictions as to the factors that influence their effectiveness.

Since comparing instances of the same concept can serve to highlight commonalities between them, it may be beneficial to compare instances that share as few features that are irrelevant for the characterization of the concept as possible. This notion, called "conservative generalization" by Medin and Ross (1989), is that people will generalize as minimally as possible, preserving shared details unless there is compelling reason to discard them. As within-category objects become more similar, their superficial similarities might be mistaken as defining ones and might lead to too narrow a category representation, for example when learning to discriminate pairs of similar-sounding words (Rost & McMurray, 2009), or when learning about which methods to use in exploratory data analysis (Chang, Koedinger, & Lovett, 2003). By varying the irrelevant features possessed by examples with a single category, the relatively stable, deep commonalities stand out and can make hard learning tasks like learning relational syntax rules from examples feasible (Gómez, 2002).

All of the studies mentioned above find advantages of low similarity for learning a concept using within comparisons. Kotovsky and Gentner (1996) add an important constraint to this. They argue that a meaningful comparison of structured instances requires first successfully aligning them and this can be too difficult a task for instances that are very different on the surface. Using the notion of "progressive alignment," they demonstrate that especially at the beginning of a learning process, comparing high-similarity instances of the same category can be essential (Gentner, 2010).

For the case of comparing instances between categories, the predictions of the influence of similarity on the learning progress are more univocal. In order to learn how to tell two categories apart, one should best compare the most similar instances of the two categories with each other, or, more precisely, the instances that have the smallest number of non-discriminative differences. This has the advantage of decreasing the likelihood of spurious differences being chosen as the basis for discriminating the categories, as Winston (1970) described using the notion of "near misses." An additional advantage of high similarity for between-category comparisons is the observation that when learning to distinguish between several similar concepts, one major difficulty lies in identifying the subtle differences between them. One proposal is that interleaving similar categories results in increased between-category contrast and discriminability, which in turn enhances learning (Carvalho & Goldstone, 2013; Birnbaum, Kornell, Bjork, & Bjork, 2012; Kornell & Bjork, 2008; Kang & Pashler, 2012).

In summary, the two lines of arguments described above

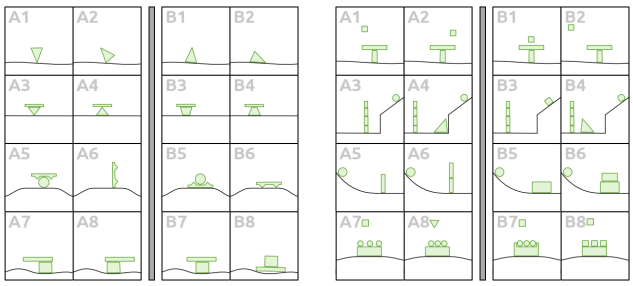
predict that between-category comparisons should best be made using similar instances, while within-category comparisons should be made using dissimilar instances, as long as the instances are still similar enough to allow for meaningful alignment. Both types of comparisons are potentially important in learning concepts and the best weighting between them will be different across learning situations, depending on the specific task, context, experience of the learner, and structure of concepts (Goldstone, 1996).

In previous research, we pitted the predicted influences of similarity in within- vs. between-comparisons against each other by grouping all instances by similarity or by dissimilarity (Weitnauer, Carvalho, Goldstone, & Ritter, 2013). The choice to use a single similarity factor to manipulate both within and between similarities made it difficult to draw strong conclusions from the results. We now present two new experiments that shed more light on how similarity effects the efficiency of comparisons. The first experiment is a replication of the one in our previous paper, but with the within and between similarities disentangled. The second experiment extends the paradigm to a more natural way of presentation, in which instead of showing a sequence of instance pairs, all instances are available to the learner at once.

In the experiments, we use Physical Bongard Problems (PBPs), which were recently introduced by Weitnauer and Ritter (2012), as our problem domain. Each PBP consists of two sets of 2D physical scenes representing two concepts that must be identified. The scenes of the first concept are on the left side, the scenes of the second concept are on the right. Figure 1 shows two example problems. What makes PBPs particularly interesting as a domain for concept learning is their open-ended feature space. People do not know in advance which features a solution might be based on (or indeed what the features are), and while some of the problems rely on features that are readily available such as shape or stability, others rely on relationships between the objects or require the construction of features as a difficult part of the solution (e.g., the time an object is airborne or the direction a particular object in the scene is moving in). This intricate situation in which both features and concepts have to be identified at the same time is quite common in real life and people deal with it impressively well, while it is still considered a very hard problem in the Artificial Intelligence community.

Experiment 1

In this experiment, we analyze the effects of within-category and between-category comparisons of similar and dissimilar PBP scenes. Participants are presented with a sequence of screens, where on each screen exactly two of the PBP scenes are visible while all other scenes are covered. Their task is to derive the concept the scenes on the left belong to and the concept the scenes on the right belong to. We vary the order in which the scenes are shown and which scenes are shown together. This allows us to a) manipulate the *within similarity*, which is the average similarity of scenes from one category that are shown together, b) manipulate the *between similarity*, which is the average similarity of scenes from different categories that are shown together, and finally c) promote either within- or between-category comparisons by showing the scenes using one of two *presentation schedules*.



(a) PBP 08 (b) PBP 33

Figure 1: The task in Physical Bongard Problems is to identify the two concepts A and B. The concepts labels are not shown during a study. See the end of the paper for the solution.

In the first presentation schedule, the *blocking* schedule, scenes that are shown simultaneously are taken from the same category (AA-BB-AA-BB-AA-BB-AA-BB). In the second, the *interleaved* schedule, simultaneously show scenes are taken from different categories (AB-AB-AB-AB-AB-AB-AB-AB). See Figure 4 for a graphical explanation. In this design, the blocked condition facilitates within-category comparisons, while between-category comparisons can still be made across successive scene pairs, but involve higher memory demands. Analogously, the interleaved condition enhances between-category comparisons but still allows for within-category comparison across successive scene pairs.

Participants

We conducted the experiment on Amazon Mechanical Turk[1]. One hundred and eighty-eight participants, all US-citizens, took part in the experiment in return for monetary compensation. Of these, we excluded 90 who did not finish all problems or did not get at least one solution correct across the entire task. Most of the excluded participants finished less than 4 problems. The data from the remaining 98 participants was used in the following analysis.

Material

We are using 22 PBPs, each with 20 scenes organized in five groups of four similar scenes. Scenes across different groups are relatively dissimilar to each other. We use 16 scenes from four of the five groups as training scenes, which allowed for the highest contrast in the similarity of close scenes between the different conditions (see Figure 3). The remaining four scenes of each problem were used as test scenes together with two randomly selected training scenes.

[1]See Mason and Suri (2012) for an introduction to using Mechanical Turk as a platform for research.

Design

We used a 2 × 2 × 2 factorial design. The study condition *presentation schedule* ∈ {*blocked*, *interleaved*} × *within-category similarity* ∈ {*similar*, *dissimilar*} × *between-category similarity* ∈ {*similar*, *dissimilar*} was randomly chosen for each problem in a within-subject manner and was balanced for each subject.

Procedure

The participants were first presented with a short introduction to the domain of PBPs and how they work, together with a solved example problem. They then had to solve a series of 22 PBPs presented in an order which was designed to minimize context effects between consecutive problems. For each problem, the participants were presented a sequence of scene pairs through which they could cycle in their own pace and as often as they wanted. This is essential to enable reinterpreting old scenes in the light of a new solution hypothesis (Finks, Pinker, & Farah, 1989). Which scenes were paired was determined by the presentation schedule and the two similarity conditions, see Figures 3 and 4 for details. When the participants had seen all scenes at least once, a button appeared on the screen that gave them the option to finish the training for the current problem whenever they liked. The button took them to a page where they had to classify six test scenes, one at a time, as belonging to the left or the right category. The participants were then prompted to type in a free text description of what defined both concepts. After submitting their solution and before continuing with the next problem, they were shown the current problem with all scenes at once and its correct solution. The original experiment is available online at `http://goo.gl/0TrVtB`.

Results

The written solutions of all participants were categorized as correct or incorrect by two trained coders blind to the experimental hypothesis. Cases in which the description of the left and right side was swapped as well as cases with a valid solution different from the official one were counted as correct (Cronbach's $\alpha = 0.87$). All cases of disagreement were resolved by a third trained coder, also blind to the experimental hypothesis.

We applied a 2 × 2 × 2 repeated measures ANOVA with *schedule condition*, *within similarity condition* and *between similarity condition* as factors and the proportion of correct answers (accuracy) as the dependent variable. This revealed a significant effect of presentation schedule, $F(1,97) = 16.78$, $p < .001$, and of between-category similarity, $F(1,97) = 11.05$, $p = .001$, as well as an interaction between the two factors, $F(1,97) = 19.70$, $p < .001$. There were no other significant effects ($p > .05$).

A second repeated measures ANOVA with the number of scene pairs cycled through as the dependent measure was applied. We chose the number of scene pairs a participant cycled through over the total training time, since it is less affected by participants taking a short break during an experiment. The analysis revealed a significant main effect of between similarity, $F(1,414) = 14.93$, $p < 0.001$, and a significant interaction between presentation schedule and between similarity, $F(1,414) = 7.28$, $p < 0.01$. See Figure 2b. All other effects were not reliable ($p > .05$).

To rule out the varying difficulty of PBPs as a potential confound, we included the average accuracy per problem as a covariate in an additional analysis of Experiment 1 and 2. All ANOVA's yielded the same qualitative results.

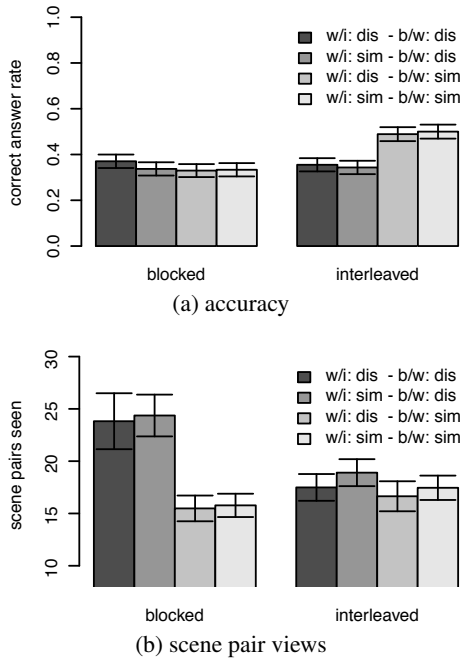


Figure 2: Top: Significantly more correct solutions in conditions with interleaved presentation and high between-category similarity. Bottom: For solved problems, participants looked at more scene pairs during blocked presentation and low between-category similarity. Error bars represent standard errors.

Discussion

The results above and an inspection of Figure 2a shows that it is the combination of the interleaved schedule and the high between similarity that is correlated with a higher rate of correct answers. This is in line with our predictions: The comparison of similar exemplars from different categories, which is by far easiest to do during interleaved presentation of high between-category similarity pairs, leads to significantly better learning results.

Despite the expected benefits of low within-category similarity for generalization, we did not find an effect of low within-category similarity, even for the blocked schedule. One possible explanation is that the necessary alignment of the scenes in each pair was too difficult to do for dissimilar scenes. A second possible explanation is that although the blocked presentation introduced a strong bias towards within-category comparisons, the participant might still have tried to build an interrelated characterization of the categories and therefore focused on discriminating between the categories

by looking for differences between successive scene pairs. The pattern in the number of scene views for solved problems in blocked schedule supports this hypothesis: There was a significant benefit of between-, but no influence of within-category similarity.

Experiment 2

The first experiment provided insight into the effect that different types of comparisons and the similarity of the compared scenes have on learning performance. We promoted either within-category or between-category comparisons by showing a sequence of pairs of specifically selected scenes. We now use a more natural way of presentation in which all scenes of the PBP are shown simultaneously. We still manipulate the similarity structure of comparisons – in a slightly more indirect way – by arranging the scenes differently across conditions.

In order to allow a direct comparison to the previous sequential scene presentation, we included the best condition from Experiment 1 additionally to four similarity conditions in the new simultaneous presentation schedule.

Participants

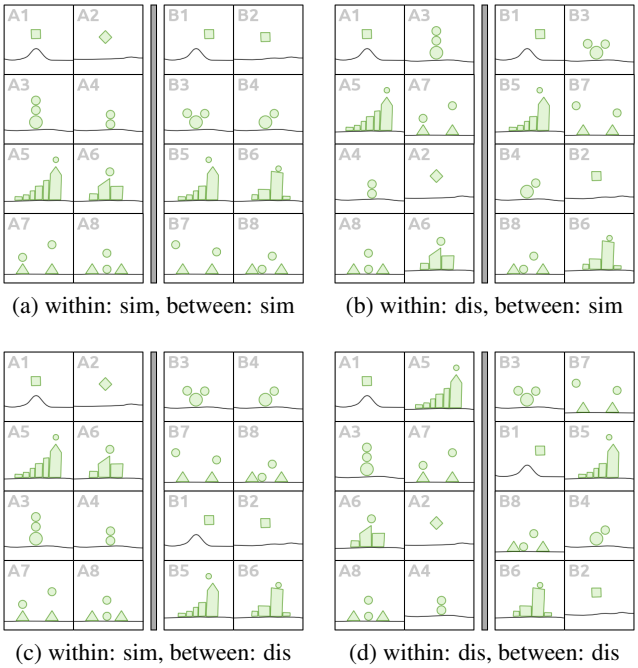
We conducted the experiment on Amazon Mechanical Turk. One hundred forty-three participants, all US-citizens, took part in the experiment in return for monetary compensation. Of these, we excluded 52 who did not finish all problems or did not get at least one solution correct across the entire task. The data from the remaining 91 participants was used in the following analyses. On average, participants solved 11.5 out of the 22 problems presented.

Material

We used the same problems in the same order as in Experiment 1. All training scenes were shown at the same time in one of four spatial arrangements, according to the condition. The scenes were aligned so that for the high within-category similarity condition, similar scenes of the same category were placed spatially close to each other, while for the low within-category similarity condition, they were placed far from each other. Analogously, adjacent scenes of different categories were similar for the high between-category similarity condition and dissimilar for the low between-category similarity condition. The different spatial alignments are shown exemplarily for PBP 24 in Figure 3. There was a fifth condition that replicated exactly the best condition of the previous experiment, "interleaved-sim-sim", in order to allow a direct comparison of performance for simultaneous versus sequential scene presentation. See Figure 4 for the timing of training scene display in both the simultaneous and the sequential presentation schedule. The test scenes were selected and presented as in Experiment 1.

Design

We used a 2×2 factorial design for the simultaneous schedule plus one additional sequential condition. One of these five conditions was chosen for each problem presented to a subject in a balanced and within-subject manner.



(a) within: sim, between: sim (b) within: dis, between: sim

(c) within: sim, between: dis (d) within: dis, between: dis

Figure 3: Arrangement of scenes of PBP 24 for the simultaneous presentation schedule. Each arrangement varies on how similar the scenes close to each other are.

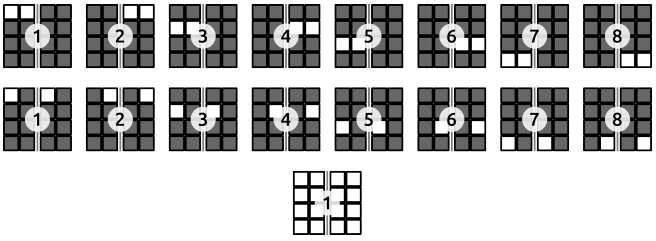


Figure 4: Presentation schedules. The positions and timing at which the scenes are shown for blocked (top), interleaved (center) and simultaneous (bottom) presentation. For the first two, the participants proceed manually through the eight states as often as they want. White squares represent visible, gray squares represent hidden scenes.

Procedure

The course of the experiment was identical to that of Experiment 1, except that for the simultaneous presentation schedule, participants could not cycle through scenes pairs but were rather shown all scenes at once and could directly proceed to the next stage. The original experiment is available online at `http://goo.gl/066U59`.

Results

The written solutions were coded as in Experiment 1. (Cronbach's $\alpha = 0.79$.)

We applied a 2×2 repeated measures ANOVA on all problem instances presented with the simultaneous schedule to analyze the effect of the factors *within similarity condition*

and *between similarity condition* on the proportion of correct answers (accuracy). We found a significant effect of *within-category similarity* $F(1,90) = 8.07$, $p < .01$. The *between-category similarity* had no significant effect and there was no significant interaction. In a separate ANOVA, we found no significant effects of the similarity conditions on the reaction time ($p > .05$).

We used four planned t-tests to compare the accuracy in the interleaved condition with the performances in each of the four simultaneous conditions. For the condition "simultaneous-sim-dis", which is the most difficult of the simultaneous conditions, we got a significant difference $t(90) = 3.1$, $p = 0.01$, where p was corrected using the Bonferroni method to correct of multiple comparisons. There was no significant difference in reaction times between the interleaved and the simultaneous presentation schedules ($p > .05$).

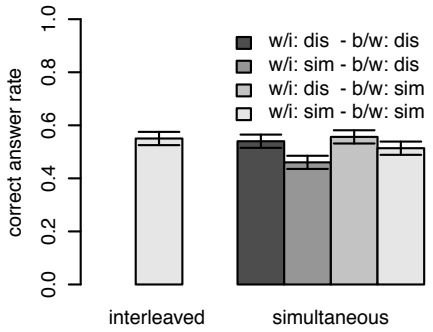


Figure 5: There were significantly more correct solutions for low vs. high within-category similarity for simultaneous presentation. Error bars represent standard errors.

Discussion

This experiment resulted in two important findings. First, for the simultaneous presentation, low within-category similarity was correlated with better classification performance, which is in line with the body of research that shows variance (low similarity) to be beneficial for category learning. We did not find the expected positive effect of high between-category similarity, though. This is different but complementary to the results of Experiment 1 and we will look into an explanation for this in the general discussion below.

Second, the interleaved presentation of PBP scenes in the high within, high between similarity condition was significantly better suited for learning than the simultaneous presentation with high within and low between similarity. Although we are comparing the best of the sequential conditions with the worst of the simultaneous conditions, this is still a surprising result: If not for generating solution hypotheses, then at least for validating them the simultaneous condition should provide an advantage over the sequential one, in which one never gets to see how all pieces fit together. That there nevertheless is a simultaneous condition that is better than a sequential one strongly suggests that the process of selecting which scenes should be compared is a substantial and non-trivial part of the learning task. Presenting just two scenes at a time, which are chosen to be beneficial to the learning when compared, can promote efficient perception and reasoning strategies which lead to better learning results.

General Discussion

The two main results presented in this paper are first, the benefit of comparing similar scenes of different categories in Experiment 1 and second, the benefit of comparing dissimilar scenes of the same category in Experiment 2.

The first result follows naturally from the prediction based on the notion of discriminative contrast, as discussed in Carvalho and Goldstone (2013) and Kang and Pashler (2012), which states that direct comparison of instances from different categories highlights their differences, together with the insight that comparing similar instances is especially effective since there are fewer superficial differences and the alignment of instances is easier (see Winston (1970) on "near misses" as well as Markman and Gentner (1993)).

The second result is predicted by theories like "conservative generalization" by Medin and Ross (1989) that attribute the advantage of low similarity within-concept comparisons to having less superficial similarities that can be mistaken for the defining similarities.

The question remains of why we did not find both effects in both experiments. Our answer is based on the insight that there are different approaches people use when learning concepts. Both Goldstone (1996) and Jones and Ross (2011) argue that learners might either focus primarily on what a category is like (using "inference learning" to build a "positive" or "isolated characterization") or focus on how a category is different from other categories (using "classification learning" to build an "interrelated characterization"). What kind of comparisons are most informative depends strongly on which of these approaches is pursued by the learner. While for building a positive characterization within-category comparisons are of central importance, for building an interrelated characterization the between-category comparisons are more useful.

An interpretation of the results consistent with the different findings across the two experiments is that in Experiment 1, participants were focusing on between-category comparisons because they were trying to build an interrelated characterization of the concepts. Naturally, participants would make few within-category comparisons, explaining why the within-similarity condition did not play a significant role. In Experiment 2, subjects might have tried to build a positive characterization instead, and consequently did not pay much attention to between-category comparisons, explaining the effect of within-category similarity and the lack of an effect of between-category similarity.

There are a couple of reasons why it is plausible to assume that participants chose to look for differences in the sequential presentation and for commonalities in the simultaneous presentation. In the simultaneous presentation, all instances of one category were grouped together on one side of the scene, a layout that allows for quickly scanning all instances to ef-

ficiently check for reoccurring patterns and shared features. When presented with just two scenes at a time, however, looking for differences might appear as the more efficient strategy: due to the open ended feature space of the PBP domain, participants had to identify or construct relevant feature dimensions as a major part of the task. Comparing similar scenes from different concepts highlights such feature dimensions, an advantage that comparing dissimilar scenes within one concept does not provide (see also Carvalho & Goldstone, 2013).

In Weitnauer et al. (2013), we introduced the grouping of instances by similarity as a dimension along which presentation order can be manipulated to optimize learning. The results of the current experiments provide a deeper insight into when and how to group instances. For between-concept comparisons, instances should be similar, while for within-concepts comparisons it is beneficial to look at dissimilar cases. How big the influence of the similarity of the compared scenes on the learning performance is, depends on which kind of comparisons are favored by the learning strategy that is used. How this strategy is chosen will depend, among other factors, on the task, the category structure and the way of presentation.

Acknowledgments

This research was supported by Grant R305A1100060 from the IES Dep. of Education and Grant 0910218 from the NSF REESE program. PFC was also supported by Graduate Training Fellowship SFRH/BD/78083/2011 from the Portuguese Foundation for Science and Technology. We would like to thank Joshua de Leeuw, the author of the jsPsych library (Leeuw, 2014) which we used in our Mechanical Turk experiment. We thank our coders Abigail Kost, Alifya Saify and Shivani Vasudeva.

Solutions to the problems

PBP 08: unstable vs. stable

PBP 24: several possible outcomes vs. one possible outcome

PBP 33: the construction gets destroyed vs. it stays intact

References

Birnbaum, M., Kornell, N., Bjork, E., & Bjork, R. (2012). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, 1–11.

Carvalho, P. F., & Goldstone, R. L. (2013). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & cognition*.

Chang, N., Koedinger, K. R., & Lovett, M. C. (2003). Learning spurious correlations instead of deeper relations. *Proceedings of the 25th Cognitive Science Society. Boston, MA: Cognitive Science Society*, 228–233.

Finks, R. A., Pinker, S., & Farah, M. J. (1989). Reinterpreting visual patterns in mental imagery. *Cognitive Science*, *13*(1), 51–78.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34*(5), 752–775.

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

Goldstone, R. (1996). Isolated and interrelated concepts. *Memory & Cognition*, *24*(5), 608–628.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436.

Hofstadter, D. (1996). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. Basic Books.

Jones, E. L., & Ross, B. H. (2011). Classification versus inference learning contrasted with real-world categories. *Memory & cognition*, *39*(5), 764–777.

Kang, S., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103.

Kornell, N., & Bjork, R. (2008). Learning concepts and categories is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*(6), 2797–2822.

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.

Leeuw, J. de. (2014). jsPsych [Computer software manual]. Bloomington, IN. Available from `https://github.com/jodeleeuw/jsPsych`

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, *50*(4), 315–353.

Markman, A., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431–431.

Mason, W., & Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods*, *44*(1), 1–23.

Medin, D., Goldstone, R., & Gentner, D. (1993). Respects for similarity. *Psychological review*, *100*(2), 254.

Medin, D., & Ross, B. (1989). The specific character of abstract thought: Categorization, problem solving, and induction. *Advances in the psychology of human intelligence*, *5*, 189–223.

Mitchell, M. (1993). *Analogy-making as perception: A computer model*. MIT Press.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349.

Weitnauer, E., Carvalho, P. F., Goldstone, R. L., & Ritter, H. (2013). Grouping by similarity helps concept learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *35th annual conference of the cognitive science society.*

Weitnauer, E., & Ritter, H. (2012). Physical bongard problems. *Artificial Intelligence Applications and Innovations*, 157–163.

Winston, P. (1970). *Learning structural descriptions from examples* (Tech. Rep.). DTIC Document.