Grouping by Similarity Helps Concept Learning

Erik Weitnauer (eweitnau@techfak.uni-bielefeld.de)
CITEC, Bielefeld University, Universitätsstr. 21-23,
33615 Bielefeld, Germany

Paulo F. Carvalho (pcarvalh@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E 10th St
Bloomington, IN 47405 USA

Robert L. Goldstone (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

In inductive learning, the order in which concept instances are presented plays an important role in learning performance. Theories 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 leaves open the question of the relative influence of similarity on interleaved versus blocked presentation. To answer this question, we pit within- and between-category similarity effects against each other in a rich categorization task called Physical Bongard Problems. We manipulate the similarity of instances shown temporally close to each other with blocked and interleaved presentation. The results indicate a stronger effect of similarity on interleaving than on blocking. They further show a large benefit of comparing similar between-category instances on concept learning tasks where the feature dimensions are not known in advance but have to be constructed.

Keywords: category learning; order effects; similarity

Introduction

Inductive learning is an essential cognitive ability which, by abstracting from specific examples, allows the transfer of experience to new, similar situations. There is a significant body of evidence from cognitive psychology suggesting that comparison of multiple cases represents a particularly promising avenue for inductively learning difficult, relational concepts (Loewenstein & Gentner, 2005). Comparison not only takes representations as inputs to establish similarities, but also uses perceived similarities to establish new representations (Hofstadter, 1996; Medin, Goldstone, & Gentner, 1993; Mitchell, 1993). When we compare entities, our understanding of the entities changes, and this may turn out to be a far more important consequence of comparison than simply deriving an assessment of similarity. In this paper, we are interested in identifying optimal ways of organizing these comparisons, and the kinds of cases that should be optimally compared.

One major line of argument is that comparing instances of a concept with very dissimilar features should lead to the best induction and generalization for the concept. If comparison serves to highlight commonalities between instances of the same concept while de-emphasizing differences, comparing instances that share irrelevant features could result in those features being retained in a learner's mental representation. 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 a compelling reason to discard them. This, in turn, could limit generalizability to new, dissimilar cases. Some research is consistent with this conclusion. For example, Halpern, Hansen, and Riefer (1990) asked students to read scientific passages that included either "near" (superficially similar) or "far" (superficially dissimilar) analogies. The passages that included far analogies led to superior retention, inference and transfer compared to those featuring superficially similar comparison, which showed no benefit at all.

The conservative generalization principle predicts that increasing the similarity of simultaneously presented instances from one category will inhibit people's ability to discover the rule that discriminates between the two categories. The true, discriminating rule will need to compete with many other possible hypotheses related to the many other features shared by the compared instances. By this account, decreasing the similarity of the compared instances that belong within a category will make it more likely that the proper grounds for generalization are inferred, by eliminating misleading common features that lead to incorrect categorization rules.

Results of Rost and McMurray (2009) on young infants learning to discriminate pairs of similar words point into the same direction. These authors found that increasing the within-category variability of the to-be-learned words by having different speakers repeat them increases the infants' ability to discriminate between the words. One of the potential explanations they give for their results is that young infants might still be unsure about what feature dimensions are relevant for the task and the variability in the irrelevant dimensions helps the infants to focus on the relevant, stable ones.

Another line of argument is that concepts which are highly similar to each other are better learned when instances of different concepts are interleaved. When learning to distinguish between several similar concepts, one major difficulty lies in identifying the subtle differences between them. Birnbaum, Kornell, Bjork, and Bjork (2012) suggested, in their discriminative contrast hypothesis, that interleaving instances of different concepts enhances the discriminative contrast between them and therefore helps with the task of spotting their differences, see also (Carvalho & Goldstone, 2012; Kornell & Bjork, 2008; Kang & Pashler, 2012). Additionally, comparing very similar instances from different categories has the advantage that there are fewer random, irrelevant differences that compete for attention with the defining difference (see Winston, 1970, on "near misses").

In summary, the two lines of arguments described above predict that high similarity supports between-category comparison, while low similarity supports within-category comparison. Both types of comparisons are potentially important in learning concepts, but one might be more effective than the other for a specific learning task, depending on the specific task, context, experience, and structure of concepts (Goldstone, 1996).

In this paper, we compare the effect that similarity has on learning performance in blocked and interleaved presentation schedules. Carvalho and Goldstone (2012) recently conducted an experiment with a similar purpose. They manipulated the category structures in a perceptual categorization task towards more or less similarity, both within and between categories, and found this modulates the advantage of blocking and interleaving in the expected directions.

Our approach is different in three important ways. First, we manipulate similarity by grouping concept instances into either similar or dissimilar comparison, instead of switching between two separate sets of categories. Second, we designed the blocked and interleaved schedules in a way that they would enhance within- and between-category comparison, respectively, while still allowing for both types of comparisons. Therefore, the two argument lines above make opposite predictions on whether high similarity of instances shown closely together should help or hurt the induction and will allow for a direct comparison of effect strengths. Third, we use an inductive learning task, Physical Bongard Problems (PBPs), with a much larger feature-space.

This problem domain is inspired by the Bongard problems (Hofstadter, 1979; Bongard, 1970) and was recently introduced by Weitnauer and Ritter (2012) to study concept learning and categorization of dynamic, physical situations. Each problem consists of two sets of 2D physical scenes representing two concepts that must be identified. The scenes of one concept are on the left side, the scenes of the other concept on the right side. Figure 1, 2 and 3 show three 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 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)[1]. 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.

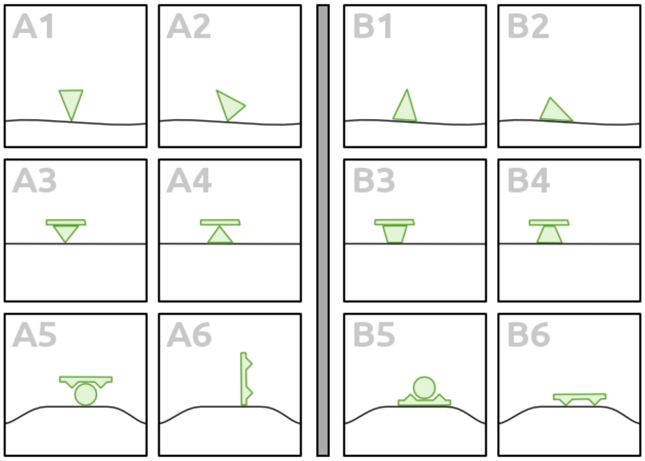

Figure 1: PBP 08. The task is to identify the two concepts A and B represented by the scenes on the left and on the right side, respectively. This is the similarity version in which similar scenes are grouped by rows. The concept labels were not shown during the study. See the end of paper for the solution.

Experiment

In this experiment we analyze the effects of different presentation schedules and similarity groupings on concept learning performance. We selected 22 PBPs and extended them by additional scenes so that the problems consist of sixteen training scenes and 8 test scenes each. Half of the scenes are shown on the left side and belong to category A (we name them A1, ..., A10) while the other half of the scenes are shown on the right side and belong to a different category B (we name them B1, ..., B10). All scenes were designed to fit into five similarity groups {A1, A2, B1, B2}, {A3, A4, B3, B4}, {A5, A6, B5, B6}, {A7, A8, B7, B8} and {A9, A10, B9, B10}, so that within-group similarity between the scenes is high, whereas between-group similarity is low.

During presentation, two scenes are always displayed simultaneously so that for each problem a sequence of six train-

[1] Solutions can be based on a great variety of features and feature combinations, as geometrical or physical object features, the way a physical scene evolves over time, relations between the objects, or even potential interactions with the scene. Additionally, focusing on a subset of objects and aligning the scenes with each other is required to find some of the solutions.

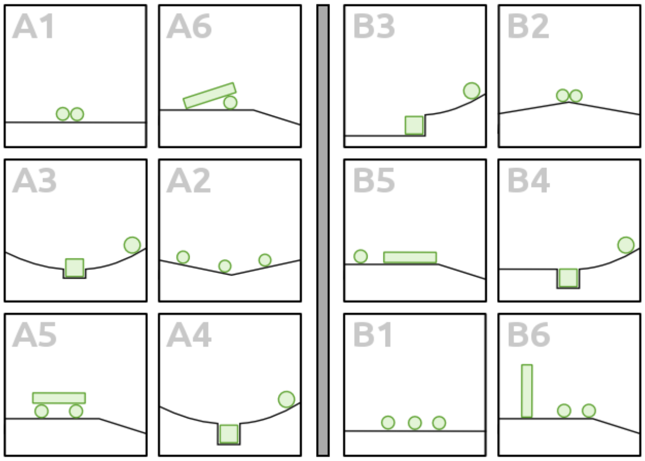

Figure 2: PBP 18. This is the dissimilarity version in which similar scenes are positioned far from each other. See the end of paper for the solution.

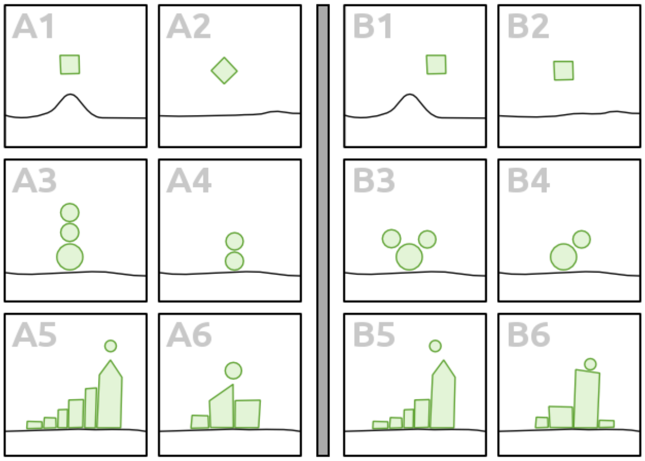

Figure 3: PBP 24. This is the similarity version in which similar scenes are grouped by rows. See the end of paper for the solution.

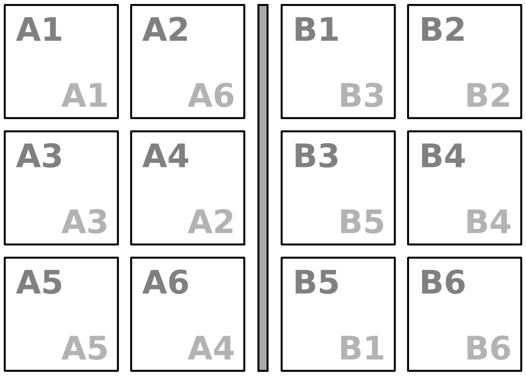

Figure 4: Positions of the 12 training scenes for the conditions *grouped by similarity* (upper left corners) and *grouped by dissimilarity* (lower right corners).

ing scene pairs is shown to the participant. We vary the presentation order of scenes along two dimensions with two values each, resulting in four conditions. The first dimension, similarity grouping, controls whether similar scenes are shown temporally close to each other (“111122223333”) or temporally far from each other (“132121323213”). We will refer to the former as “grouped by similarity” or “similar” and to the latter as “grouped by dissimilarity” or “dissimilar”. Figure 4 depicts how scenes are positioned for both cases.

The second dimension, presentation schedule, controls whether the scenes that are shown simultaneously are from the same or from different categories (AA-BB-AA-BB-AA-BB vs. AB-AB-AB-AB-AB-AB, see Figure 5). We will refer to the former as “blocked” condition[2] and to the latter as “interleaved” condition. In the blocked condition while within-category comparisons are facilitated by presenting scenes from the same category simultaneously, between-category comparisons can still be made between 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.

We expected to find that grouping by similarity should improve learning performance for the interleaved condition and grouping by dissimilarity should improve performance for the blocked condition.

Subjects

We conducted the experiment on Amazon Mechanical Turk[3]. Sixty-seven participants, all US-citizens, took part in the experiment in return for monetary compensation. Of these, we excluded 27 who did not finish all problems (most of them dropped out after seeing only a few) and another two that did not get at least one solution correct across the entire task. There was no need to use catch trials, because the subjects were required to write down the solutions as free text. Any cheating or automated answers would have become immediately apparent during our hand-coding of the solutions. The data from the remaining 38 participants was used in the following analyses. On average, participants solved 8.6 out of the 22 problems presented.

Material

For each of the 22 problems, the training scenes were arranged in three rows, each with four scenes. We prepared two versions of each problem by placing the scenes at different positions. In the “grouped by similarity” version, the scenes were arranged in such a way that the scenes inside

[2] We use the term “blocked” to refer to a slightly different presentation schedule than it is usually done. Instead of showing all instances of one category before switching to the next, we only block two instances of one category and interleave these blocked pairs.

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

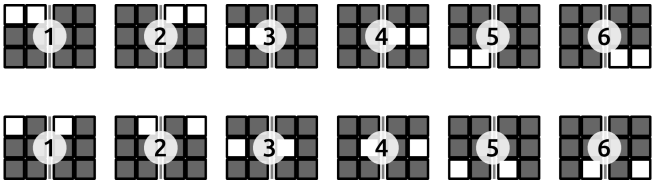

Figure 5: The scene presentation schedule for blocked (top) and interleaved (bottom) presentation. The participant manually proceeds through the six states. In each state, two scenes (in white) are shown while the other scenes (in gray) are hidden.

each row are similar to each other. In the “grouped by dissimilarity” version, similar scenes were distributed over all rows. Figures 1 and 2 show an example of a dissimilarity and similarity version, respectively.

Design

We used a 2 x 2 factorial design. The study condition (presentation schedule: {blocked, interleaved} × similarity grouping: {similar, dissimilar}) was randomly chosen for each problem in a within-subject manner.

Procedure

The participants were first given a brief introduction to PBPs including an example problem with a solution. During the experiment, they could proceed through the scene pairs of each problem at their own pace by pressing a key. After they had viewed all scenes once, they were asked whether they thought they had found a solution. Then they needed to classify six test scenes which were randomly drawn from the eight available test scenes. The test scenes were shown one by one. Finally they had to type in a description of their solution or their best guess. Before moving on to the next problem, they were shown the problem with all training scenes at once together with the official solution. There was no time limit to the task. At the end of the experiment participants were debriefed on the study objectives and variables. The original experiment is available online at Weitnauer (2013).

Results

We used two separate measures to evaluate learning success. First, we hand-coded the accuracy of each textual solution given by the participants. Some of the participants had difficulties remembering which side was left and which side was right, so they provided a correct solution but with sides swapped (e.g., writing “left: all objects are squares” and “right: all the objects are circles” when in fact the left-side objects were all circles and the right-side objects were all squares). These cases were counted as correct solutions.

The second measure is based on the proportion of test scenes that were classified correctly. Using this directly would be misleading for cases in which participants mixed up the sides. We therefore developed a consistency measure instead. This consistency measure is defined as $\max(c, 6-c) - 3$, where c is the number of correctly classified scenes being minimally zero and maximally six. The consistency can take values between zero and three, where the latter corresponds to cases where either all test scenes were classified correctly or were all (consistently) classified wrongly. Figures 6 and Figure 7 show the average of these two measures for all four conditions.

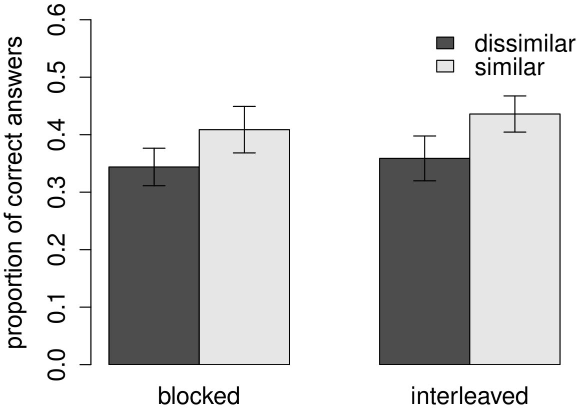

Figure 6: Mean proportion of correct answers for blocked and interleaved presentation schedules and grouping of scenes by similarity or dissimilarity. There is a significant effect of similarity.

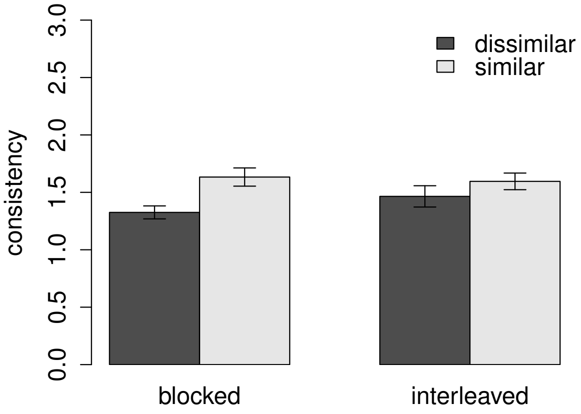

Figure 7: Mean consistency of test scene classifications for blocked and interleaved presentation schedule and grouping of scenes by similarity or dissimilarity. There is a highly significant effect of similarity.

We applied two separate 2 x 2 repeated measures ANOVAs with presentation schedule (blocked vs. interleaved) and similarity grouping (similar vs. dissimilar) as factors to the proportion of correct responses and consistency measures. These analyses revealed a significant effect of similarity condition, $F(1,37) = 5.32$, $p = .03$ for the proportion of correct answers measure and $F(1,37) = 15.7$, $p = .0003$ for the consistency measure. There was no effect of schedule of presentation, or interaction between the two factors for any of the measures (all $p > .05$).

Discussion

The data analysis revealed a positive effect of grouping scenes by similarity, independent of whether they were presented in a blocked or an interleaved schedule. We argue that this is explained by a strong positive effect of similarity on interleaving which more than compensates for any possible negative effect that similarity had on blocking.

The advantage of similarity for interleaving is in line with our expectations. Goldstone (1996) and the discriminative contrast hypothesis of Birnbaum et al. (2012) predict that direct comparison of instances from different categories highlights their differences (see also Carvalho & Goldstone, 2012). Identifying differences between highly similar scenes is especially effective, as there are fewer superficial differences to compete with the defining one. This insight is already present in the desirable "near misses" in Winston (1970) work, where instances from different concepts that differ by just one feature are ideal for his algorithmic learner. Near misses provide clear evidence about what features are critical, concept-defining ones. Another possible contributing effect is that it is easier to structurally align two similar scenes than two very different scenes and this alignment process promotes noticing differences (Markman & Gentner, 1993).

What might seem surprising at first is that similarity also improves learning performance in the blocked condition, given that theories like "conservative generalization" by Medin and Ross (1989) predict that similarity for blocked scenes will lead to many superficial similarities and therefore inferior performance compared to dissimilar scenes. However, the results can be explained in a way compatible with these theories. We designed both scheduling conditions in a way that allows for within- *and* between-category comparisons. Given this, negative effects of similarity on the former and positive effect of similarity on the latter will compete with each other. In the blocking condition, within-category comparisons were facilitated by showing scenes of the same category simultaneously, while scenes of different categories had to be compared sequentially.

Still, a strong positive effect of similarity on between-category comparison could mask a small negative effect of similarity on within-category comparison and lead to the overall improvement due to similarity that we found. What is indeed surprising is that, although learners were pushed towards attending to similarities with a paired comparison, they still exploited between-pair differences to find the solution.

We believe that one important reason for this might be found in the type of categorization task that was used. Due to its open ended feature space, participants had to identify or construct relevant feature dimensions as a major part of the challenge. Comparing similar scenes from different concepts provides the additional advantage of highlighting such feature dimensions, an advantage that blocking of dissimilar scenes does not provide.

Implications for an Algorithmic Learner An interesting question is how the presented results could inform the implementation of a computational model of concept learning in open feature-spaces. A general observation is the fact that presentation order matters at all. This means that attending to the first scenes changes the way the following scenes are perceived and solution hypotheses that are formed. The limited memory capacity of humans makes it impossible to keep a detailed representation of all instances or a large number of hypotheses in mind and forces a decision on which aspects of an instance one should concentrate on and which information should be retained. The big challenge is that these decisions have to be made before knowing the answer to the problem and therefore before knowing what aspects are actually important. In open-ended feature spaces algorithmic learners could face similar problems because the *a-priori* construction of all possible features might be infeasible due to a combinatorial explosion, so dynamic processes that discover feature dimensions and concepts at the same time might be necessary.

The main insight from the present experiment is that between-category comparisons of similar instances are especially beneficial, as they promote learning, both by making new, potentially relevant feature dimensions more salient and by increasing the likelihood that a perceived difference is a defining one. Between-category comparisons should therefore play a privileged role in how active learning algorithms choose the next training example.

Pedagogical Implications Birnbaum et al. (2012) showed the benefit of interleaving for several concept learning tasks, and Carvalho and Goldstone (2012) proposed that this benefit is modulated by how similar the concepts are, so that in low-similarity cases blocking can be better. The current work provides a slightly different perspective. Our results suggest no direct advantage of interleaved or blocked presentation, but instead a greater potential of between- compared to within-category comparisons. This holds even for situations in which the between-comparison relies on sequentially shown instances while within-comparison can be made on the basis of simultaneously shown instances. A result that might directly inform the design of learning material is the big benefit of comparing similar scenes from different categories. The grouping of instances by similarity - instead of relying on a single similarity measure for a whole set of concepts - is a new, interesting dimension along which presentation order can be manipulated to optimize learning.

Acknowledgments

This research was supported by Grant R305A1100060 from the Institute of Education Sciences Department of Education and Grant 0910218 from the National Science Foundation REESE program. PFC was also supported by Graduate Training Fellowship SFRH/BD/78083/2011 from the Portuguese Foundation for Science and Technology (FCT). We would like to thank Joshua de Leeuw, the author of the jsPsych library (Leeuw, 2013) which we used in our Mechanical Turk experiment and Spenser Benge for his help with stimuli preparation and response coding.

Solution to the problems

PBP 08: unstable vs. stable

PBP 18: objects eventually touch vs. objects are eventually separated

PBP 24: several possible outcomes vs. one possible outcome

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.

Bongard, M. M. (1970). *Pattern recognition*. Rochelle Park, N.J.: Hayden Book Co., Spartan Books.

Carvalho, P., & Goldstone, R. (2012). Category structure modulates interleaving and blocking advantage in inductive category acquisition. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the thirty-fourth annual conference of the cognitive science society* (pp. 186–191). Austin, TX: Cognitive Science Society.

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

Halpern, D., Hansen, C., & Riefer, D. (1990). Analogies as an aid to understanding and memory. *Journal of Educational Psychology*, *82*(2), 298.

Hofstadter, D. (1979). *Gödel, escher, bach: an eternal golden braid*. Harvester Press.

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

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.

Leeuw, J. de. (2013). *A javascript library for running behavioral experiments on the web.* 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. (2013). *Mechanical turk study on scene ordering in PBPs.* Available from http://perceptsconcepts.psych.indiana.edu/experiments/ew/pbp3/

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.