Similarity-Based Reasoning is Shaped by Recent Learning Experience

Paul H. Thibodeau (paul.thibodeau@oberlin.edu)
David Jaz Myers (david.myers@oberlin.edu)
Oberlin College, Department of Psychology
120 W. Lorain St., Oberlin, OH 44074, USA

Stephen J. Flusberg (stephen.flusberg@purchase.edu)
SUNY Purchase College, Department of Psychology
735 Anderson Hill Road, Purchase, NY 10577, USA

Abstract

Popular approaches to modeling analogical reasoning have captured a wide range of developmental and cognitive phenomena, but the use of structured symbolic representations makes it difficult to account for the dynamic and context sensitive nature of similarity judgments. Here, the results of a novel behavioral task are offered as an additional challenge for these approaches. Participants were presented with a familiar analogy problem (A:B::C:?), but with a twist. Each of the possible completions (D1, D2, D3), could be considered valid: There was no unambiguously "correct" answer, but an array of equally good candidates. We find that participants' recent experience categorizing objects (i.e., manipulating the salience of the features), systematically affected performance in the ambiguous analogy task. The results are consistent with a dynamic, context sensitive approach to modeling analogy that continuously updates feature weights over the course of experience.

Keywords: similarity; analogy; statistical learning; relational reasoning; categorization

Introduction

But soft! What light through yonder window breaks?
It is the east, and Juliet is the sun.

Romeo & Juliet, Act 2, Scene 2

It is one of the most recognizable metaphors in the Western canon: a love-struck Romeo spies Juliet at her window and compares her to the star that nourishes the earth with light and heat. Though it seems like a straightforward, if not clichéd figure of speech, there are in fact several commonalities between Capulet's daughter and the sun; they are both rising in the east, they are both beautiful, and they are both golden (for the sake of illustration, we assume Juliet is wearing a yellow dress). When we are confronted with several possible matches in a similarity-based comparison such as this, what drives us to select one interpretation in particular[1]?

In the past three decades, a great deal of research has examined the nature and development of similarity-based reasoning, which is believed to play a major role in everything from problem solving to creativity to scientific discovery (Holyoak & Thagard, 1996). Cognitive scientists have argued that *analogy* is at the heart of this process, and computational models of analogy have successfully captured a range of cognitive and developmental phenomena that require similarity-based reasoning (Gentner & Forbus, 2011; Hummel & Holyoak, 2005). The dominant approach to modeling analogy assumes that highly structured symbolic (or hybrid) representations and a particular suite of cognitive machinery are necessary for analogical mapping (e.g., SME: Falkenhainer, Forbus, & Gentner, 1989; LISA: Hummel & Holyoak, 2003; Hummel, 2010).

One key observation is that children undergo a perceptual-to-relational shift in reasoning over the course of development: younger children typically match based on perceptual similarities ("they're both yellow!"), while older children may start to match based on relational similarities ("they are both *rising* in the east!"; Gentner, 1988; Piaget, 1952). This developmental trajectory could be taken as evidence for structured models of analogy like SME and LISA as these approaches are well equipped to model this phenomenon. Once children learn the important relations in the world (e.g. that the sun appears in the east as the earth rotates during its orbit each day), they represent these structures symbolically (Gentner et al., 1995). Such representations would allow a mapping mechanism to operate reliably over structured symbols for consistent and effective similarity-based reasoning.

However, several empirical findings may present a challenge for this approach. For instance, similarity-based reasoning does not always follow a perceptual-to-relational shift. Adults are sometimes lured towards perceptual matching (Morrison et al., 2011) and even children flexibly generalize based on perceptual *or* relational similarity depending on the task context and their prior experiences (Bulloch & Opfer, 2009; Opfer & Bulloch, 2007).

Recently, we showed that these patterns of behavior spontaneously emerge over the course of development in a connectionist model that relies on domain-general statistical learning and fully distributed internal representations (Thibodeau, Tesny, & Flusberg, 2014). Indeed, context-sensitive, similarity-based reasoning is one of the key strengths of the connectionist framework (Flusberg & McClelland, 2014), and while critics have commonly

[1] With all due respect to our high school English teachers, who convincingly argued that Shakespeare's brilliance lies in the fact that his metaphors are intrinsically multifaceted, and thus no single interpretation is really "correct."

assumed that this only applies to perceptual or feature-based similarity (see, e.g., the accompanying commentary to Rogers & McClelland, 2008), we have argued – and demonstrated – that fully distributed neural networks can in fact reason and draw inferences based on shared relational structure (Flusberg et al., 2010; Thibodeau et al., 2013). A critical feature of these models is their ability to extract and represent the statistical structure of their inputs over the course of training. By treating (relational) language as part of the fabric of experience, we can see the emergence of sophisticated metaphorical and analogical abilities.

One unique prediction that follows from this type of model is that what counts as "similar" in comparisons among novel stimuli will dynamically change as a result of recent experience. This is because internal representations of environmental structure are updated relatively quickly during early epochs of training, which will cause the object features that are prominent in experience to have greater weight in object representations (and therefore in emergent similarity relationships). The current study represents a novel test of this prediction by investigating whether recent experience attending to particular feature dimension really does lead that feature to figure more prominently in analogical reasoning. Finding support for this possibility would present a challenge for models of analogical reasoning that do not naturally accommodate dynamic, experience-based updating of distributed object representations.

To bring it back to Shakespeare, if an audience member has a great deal of (recent) experience making similarity judgments based on color, we would expect her to interpret Romeo's remark as a comment on Juliet's outfit. If, on the other hand, she has been attending more to the attractiveness of those around her, we would expect her to interpret the metaphor as a comment on Juliet's radiant beauty.

In the present study, we tested the hypothesis that getting participants to attend to and learn about a particular facet of experience would influence their subsequent similarity-based reasoning. Participants first completed a category-learning task where they had to figure out the meaning of a novel category label that referred to one of several possible features (size, shape, or brightness). They then completed an ambiguous analogy task and a similarity-rating task. Our hypothesis was that experience in the training phase would influence subsequent performance in the similarity-based reasoning tasks.

Experiment

Methods

Participants We sampled 400 participants from Amazon's Mechanical Turk. Of these, eight submitted incorrect completion codes and were excluded from analysis, leaving data from 392 for analysis.

Materials and Design The experiment consisted of four between-subjects conditions and three tasks: a category label training task, an ambiguous analogy task, and a similarity rating task.

Training Task Participants were randomly assigned to one of four training groups. Three of them differed in how they provided feedback to participants learning about a novel category label. A fourth, control condition omitted the training phase altogether.

On each of the 16 trials of the training task, participants were presented with two shapes and asked to choose "Which is the {truffet/lodi}?" (see Figure 1 for an illustration of a trial of the training phase). For half of the participants, the target label was "truffet" and for the other half, the target label was "lodi." The two shapes always differed along three dimensions: *size*, *shape*, and *brightness*. There were two levels of each dimension (i.e. large vs. small, circle vs. square, and bright vs. dim). Participants were forced to choose between the shapes and were given accuracy feedback ("correct" or "incorrect"). One training group was given feedback indicating that the novel label referred to the dimension of size (either meaning "bigger" or "smaller", counterbalanced across participants), another group was given feedback indicating it referred to shape ("square" or "circle"), and the last training group was given feedback indicating it referred to brightness ("dimmer" or "brighter"). While the feedback participants received always consistently mapped on to one of these relations, they were never explicitly told what the target label meant.

The pairs of shapes shown in the 16 training trials were identical across conditions, though the order in which they were presented was randomized.

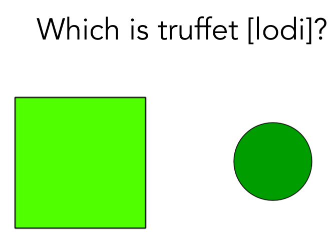


Figure 1. An example trial from the training phase.

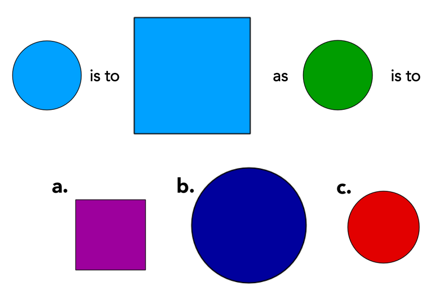


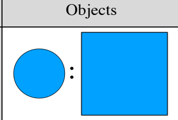
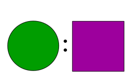
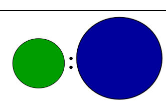
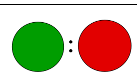
Figure 2. An example trial from the ambiguous analogy task.

Ambiguous Analogy Task Following the training phase, participants completed eight trials of an ambiguous analogy task. The instructions for this task read: "On the following screens, you will see a series of analogy questions, as in *A is* *to B as C is to what?* You will see three items that could potentially complete the analogy. Choose the one that you think best completes the analogy." Figure 2 shows an example trial of this task.

The analogy task is ambiguous because none of the forced choice responses are a perfect match to the sample relation. For instance, in Figure 2 the best answer for the analogy would be a large, green, square. A large green square would be a *different* size and shape than the small green circle but the *same* brightness; this would mirror the relationship between the large blue square and the small blue circle, which are *different* sizes and shapes but the *same* brightness.

This "perfect" analogy was not provided as an option to participants, though. Instead, the three forced choice options were designed to be relationally similar to the sample in one of three ways: by shape, size or brightness. For instance, in Figure 2 (see also Table 1), (*a*) was a relational match to the sample by shape (because, like the sample objects, these two were *different* shapes) but not by size or brightness (because, unlike the sample objects, these two were the *same* size and *different* in brightness); (*b*) was a relational match to the sample by size (because, like the sample objects, these two were *different* sizes) but not by shape or brightness (because the two objects were, unlike the sample, the *same* shape and *different* in brightness); (*c*) was a relational match to the sample by brightness (because, like the sample objects, these two were the *same* brightness) but not by size or shape (because, unlike the sample, were the *same* size and the *same* shape).

Table 1. Illustration of the relationship between the sample objects and potential analogical matches for one of the ambiguous analogy trials. In this case, option a is the best analogy on the basis of shape, option b is the best analogy on the basis of size, and option c is the best analogy on the basis of brightness.

Option	Objects	Shape	Size	Brightness
Sample		Different	Different	Same
a.		**Different**	Same	Different
b.		Same	**Different**	Different
c.		Same	Same	**Same**

Note, however, that although we have used the terms *same* and *different* to refer to the relations in the table, the relationships themselves are dimension-specific. For instance, objects that *differ* in size, *differ* because one is *larger* (*smaller*) than the other; objects that *differ* in brightness differ because one is *brighter* (or *dimmer*) than the other.

As with the training phase, the stimuli for the ambiguous analogy task were identical across conditions. No feedback was given on this task. The order of response options and trials were randomized between participants.

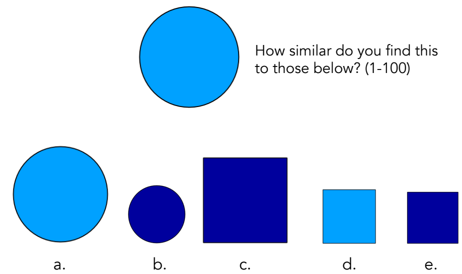


Figure 3. An illustration of the similarity-rating task.

Similarity Rating The final task consisted of four trials where participants made similarity ratings. On each trial, a sample object was shown at the top of the screen and five others were shown below (see Figure 3). Of these five items, one was identical to the original (e.g., option *a* in Figure 3) and one was maximally different, given the similarity space (i.e. option *e* in Figure 3, which is different in size, shape and brightness from the original). The other three items matched the original along one of the three target dimensions but not the other two. For instance, in Figure 3, *b* was a shape match, *c* was a size match and *d* was a brightness match.

Participants indicated how similar they viewed the original to each of the five items by using a 101-point slider bar (0 = lowest similarity, 100 = highest similarity). The order of the five comparison objects was randomized between participants, as was the order of the trials.

Results

Training The results of the training phase indicated that participants reliably learned the "meaning" of the target word they were assigned. A repeated-measures logistic regression with a predictor for trial (1-16) revealed that participants significantly improved (i.e. changed their response patterns to be more consistent with the feedback that they were given) over the course of training, $\chi^2(1)=80.053$, $p<0.001$ (AIC_1, a model with an intercept only, $=4442.1$; AIC_2, in which a predictor for trial number was added, $=4364.1$), $B=0.078$, $SE=0.008$, $p<0.001$.

A repeated-measures logistic regression with separate predictors for trial by condition (shape, size, brightness) revealed that participants in each of the three conditions significantly improved over the course of training, but it revealed that they did so at different rates, $\chi^2(2)=267.57$, $p<0.001$ ($AIC_1=4364.1$, $AIC_2=4100.5$). People improved the fastest in the "shape" condition, $B=0.438$, $SE=0.035$, $p<0.001$. People improved more gradually in the "size", $B=0.072$, $SE=0.012$, $p<0.001$, and "brightness" conditions, $B=0.028$, $SE=0.011$, $p=0.009$ (see Figure 5).

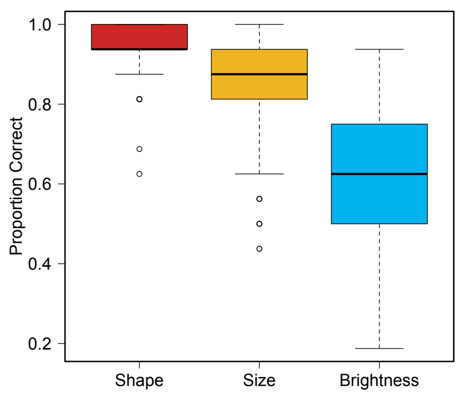


Figure 5. Boxplot illustrating differences in the degree to which participants were attuned to the feedback in the Shape, Size, and Brightness conditions.

By design, the training phase required participants to test hypotheses about the meaning of the target word. Does “truffet” (or “lodi”) mean *larger*, *smaller*, *square*, *circle*, *brighter*, or *dimmer*? The differences in the rates of change across the conditions suggest that participants were most likely to test the possibility that the novel label described the shape of the object. This is consistent with research showing that children typically have a shape bias in word learning (Landau, Smith, & Jones, 1988).

Data from the final trial of the training phase support this hypothesis: People answered the final trial in a way that was congruent with their condition 98.0% of the time in the shape condition, 83.2% of the time in the size condition, and 70.0% of the time in the brightness condition, $\chi^2(2)=29.333$, $p<0.001$.

Ambiguous Analogy Task As shown in Table 2, the basic pattern of data from the ambiguous analogy task are consistent with our prediction: After training that focused on shape, people were more likely to choose the shape-match (69.7% compared to 61.9% in the control condition); after training that focused on size, people were more likely to choose the size-match (54.8% compared to 32.7% in the control condition); after training that focused on brightness, people were more likely to choose the brightness-match (14.5% compared to 5.3% in the control condition).

Table 2. Percentages of responses by training condition.

	Shape	Size	Brightness
Control (*n* = 89)	62	33	5
Shape (*n* = 102)	**68**	29	3
Size (*n* = 101)	40	**55**	5
Brightness (*n* = 100)	50	34	**15**

To analyze these data, we fit a series of mixed-effect logistic regression models to predict participants’ choices. In these models, there were two fixed effects: response type (shape-match, size-match, and brightness-match), and training condition (shape, size, brightness, and control). Participant and trial number were included as random effects (c.f. Clark, 1973; Jaeger, 2008;).

We first tested whether there were significant differences in the likelihood of selecting the shape-, size-, and brightness-matches by comparing a model *without* a predictor for response type to a model *with* a predictor for response type, $\chi^2(2)=1890.7$, $p<0.001$ ($AIC_1=11983$, $AIC_2=10096$). This model revealed that people were more likely to choose the shape-match (54.8% overall) than the size-match (38.1% overall), $B=0.677$, $SE=0.051$, $p<0.001$, or the brightness-match (7.0% overall), $B=2.773$, $SE=0.0785$, $p<0.001$. People were also more likely to select the size-match than the brightness-match, $B=2.096$, $SE=0.079$, $p<0.001$.

We then tested whether there were significant differences in response patterns by training condition by adding interaction terms between condition and response type, $\chi^2(6)=369.43$, $p<0.001$ ($AIC_1=10096$, $AIC_2=9744.5$). The main effects of response in this model were consistent with the previous model (i.e. people were most likely to choose the shape-match and least likely to choose the brightness-match). In full model, the control condition served as a baseline (see Figure 5, which shows differences from the control condition). Critically, the model revealed the predicted congruence effects.

Compared to those in the control condition, participants in the shape condition were more likely to choose the shape-match ($B=0.262$, $SE=0.108$, $p=0.015$) and less likely to choose the brightness-match ($B=-0.538$, $SE=0.260$, $p=0.038$), though they were no less likely to choose the size-match ($B=-0.179$, $SE=0.111$, $p=0.108$).

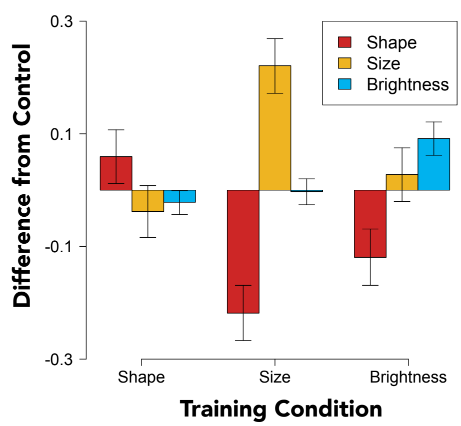


Figure 5. Results of the ambiguous analogy task. The difference in the proportion of shape-matching, size-matching, and brightness-matching choices from the control condition are shown for each condition. Error bars denote standard errors of the means.

Compared to participants in the control condition, those in the size condition were more likely to choose the size-match ($B=0.914$, $SE=0.107$, $p<0.001$) and less likely to choose the shape-match ($B=-0.888$, $SE=0.105$, $p<0.001$), though they were no less likely to choose the brightness-match ($B=-0.053$, $SE=0.231$, $p=0.818$).

Compared to those in the control condition, those in the brightness condition were more likely to choose the brightness-match (B=1.101, SE=0.195, p<0.001) and less likely to choose the shape-match (B=-0.487, SE=0.105, p<0.001, but were no less likely to choose the size-match (B=0.124, SE=0.109, p=0.256).

These analyses reveal that training systematically influenced behavior on the ambiguous analogy task. When the training phase encouraged people to attend to the shapes of the objects, they were subsequently more likely to use shape information to complete the analogy task. Parallel effects were found for the size and brightness conditions.

Individual Differences We can further investigate the effect of training by testing whether people who performed better on the training task showed more pronounced effects on the ambiguous analogy task. We found such a relationship for each of the three conditions in three separate repeated-measures logistic regressions. In the shape condition, people who selected more shape-matches in training chose more shape-matches on the ambiguous analogy task, B=7.578, SE=3.837, p=0.048. Similarly, people in the size and brightness conditions who chose more size- and brightness-matches in training chose more size- and brightness-matches in the ambiguous analogy task, B=7.399, SE=1.792, p<0.001 and B=3.533, SE=1.568, p=0.024, respectively. This shows that when we take individual differences in performance on the training task into account, we see even more nuanced and systematic effects on the ambiguous analogy task.

Similarity-Rating Results from the similarity-rating task mirrored those of the ambiguous analogy task, showing a systematic effect of training. A mixed-effect ANOVA with condition and comparison object as fixed effects and participant and trial number as random effects revealed a main effect of comparison object, F[4, 30952]=18159.528, p<0.001 and an interaction between comparison object and training condition, F[12, 30952]=33.627, p<0.001. There was a marginal main effect of condition, F[3, 388]=2.605, p=0.052.

Comparing the fit of nested models confirmed these effects: a model with a predictor for a main effect of comparison object provided a significantly better fit to the data than a model without such a predictor, $\chi^2(4)$=37127, p<0.001 (AIC_1= 310610, AIC_2=273524). Including interaction terms between comparison object and condition provided an even better fit to the data, $\chi^2(15)$=408.93, p<0.001 (AIC_1=273524, AIC_2=273270).

The full model revealed that people rated the identity match as the most similar to the comparison object (M=97.00, sd=7.85), followed by the shape-match (M= 51.56, sd=18.95), brightness-match (M=36.32, sd=18.38), size-match (M=32.34, sd=19.05), and, finally, the object that differed from the original on each of the three dimensions (M=14.04, sd=15.34). All pairwise comparisons were significant (ps<0.001).

To investigate the interaction between training condition and comparison object, we treated similarity ratings from the control condition as a baseline (see Figure 6). We found that people in the shape condition rated the shape-match as marginally more similar to the comparison object than people in the control condition, B=2.463, SE=1.422, p=0.083. These participants rated the size-match as significantly less similar, B=-3.721, SE=1.636, p=0.023. Similarity ratings of the anchors and brightness match did not differ. That is, people in the shape condition considered shape as a marginally more important dimension for determining the similarity of objects, and size as a less important dimension, after being forced to attend to shape in the training phase.

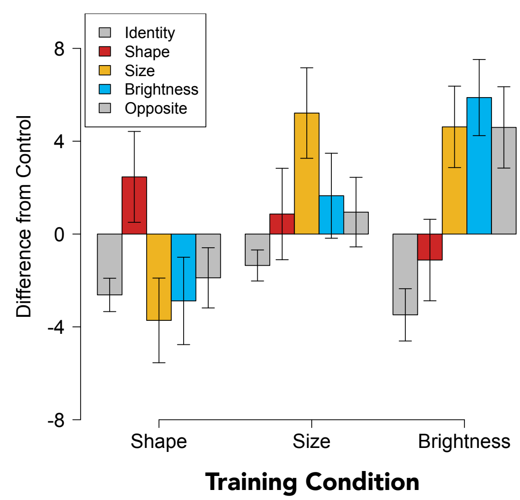


Figure 6. Results from the similarity-rating task. The difference in the mean similarity of the identity-match, shape-matching, size-matching, brightness-matching, and maximally different objects from the control condition are shown by shape, size, and brightness condition. Error bars reflect standard errors of the means.

We found that people in the size condition rated the size-match as significantly more similar to the comparison object than people in the control condition, B=5.213, SE=1.640, p= 0.001. Similarity ratings of the anchors, shape-match, and brightness-match did not differ. In other words, people in the size condition viewed size as a more important dimension for determining the similarity of objects after being forced to attend to size in the training phase.

Finally, we found that people in the brightness condition rated the brightness-match as significantly more similar to the comparison object than people in the control condition, B=5.881, SE=2.070, p=0.004. These participants also rated the size-match as more similar, B=4.618, SE=1.643, p=0.005, and the maximally different object as marginally more similar, B=4.595, SE=2.606, p=0.078. They rated the identity-match as less similar to the comparison object, B=-3.481, SE=1.517, p=0.022. Their ratings of the shape-match did not differ from the control condition. In this condition too, we see the predicted congruence effect, but we also see that people seem to shift their conception of similarity more globally as well.

General Discussion

Is Juliet the sun because she is rising in the east, because she is beautiful, or because she is yellow? The answer, of course, is "yes." However, our experiment has shown that the similarity match people make depends in no small part on their learning history and which aspects of experience they have been attending to. In our study, participants first completed a category training task where they had to determine the meaning of a novel category label by choosing which of two objects on a given trial fit the label (or they completed no training task at all). Feedback was given that was consistent with a label that mapped onto the size, shape, or brightness of objects. Participants then completed an ambiguous analogy task and a similarity-rating task.

A series of analyses demonstrated that responses to the analogy and similarity-rating task were systematically influenced by the particular category-training feedback participants received. Participants who learned that a novel category label referred to size, for example, were more likely to choose an object on the ambiguous analogy task where size was the critical dimension, even though options based on shape or brightness were equally valid. They also rated objects that matched in brightness as relatively more similar in the similarity-rating task.

These findings support a dynamic view of similarity. The salience of object features can change as a function of experience, and, in turn, affect similarity-based reasoning and inference. This finding has important implications for theories of analogical reasoning, metaphor, and similarity. In particular, popular approaches to modeling analogy and other types of similarity-based reasoning would have difficulty accommodating these results because they have no natural way of dynamically updating feature weights as a result of ongoing experience. Conversely, this is exactly the sort of finding that we would expect based on models of analogy based around principles of statistical learning and distributed representation (Thibodeau et al., 2013; 2014).

References

Bulloch, M.J., & Opfer, J.E. (2009). What makes relational reasoning smart? Revisiting the perceptual-to-relational shift in the development of generalization. *Developmental Science, 12,* 114-122.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335-359.

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

Flusberg, S. J. & McClelland, J. L. (2014). Connectionism and the emergence of mind. *Oxford University Press Handbook of Cognitive Science* (online edition)

Flusberg, S. J., Thibodeau, P. H., Sternberg, D. A., & Glick, J. J. (2010). A connectionist approach to embodied conceptual metaphor. *Frontiers in Psychology*, *1*(0), 1-11

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child development*, pages 47–59.

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *WIREs Cognitive Science, 2,* 266–276.

Gentner, D., Rattermann, M. J., Markman, A. B., & Kotovsky, L. (1995). Two forces in the development of relational similarity. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 263-313). Hillsdale, NJ: LEA.

Gentner, D., Simms, N., & Flusberg, S. (2009). Relational language helps kids reason analogically. *Proceedings of the 31st annual meeting of the Cognitive Science Society, Amsterdam.*

Holyoak, K. J., & Thagard, P. (1996). *Mental leaps: Analogy in creative thought.* Cambridge, MA: MIT Press.

Hummel, J. E. (2010). Symbolic versus associative learning. *Cognitive Science, 34,* 958-965.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220–264.

Hummel, J. E., & Holyoak, K. J. (2005). Relational reasoning in a neurally plausible cognitive architecture. *Current Directions in Psychological Science, 14*(3), 153–157.

Jaeger F. T. (2008) Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434-446.

Landau, B., Smith, L.B., & Jones, S.S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3,* 299-321.

Morrison, R.G., Doumas, L.A.A., & Richland, L.E. (2011). A computational account of children's analogical reasoning: blanacing inhibitory control in working memory and relational representation. *Developmental Science, 14,* 516-529.

Opfer, J. E., & Bulloch, M. J. (2007). Causal relations drive young children's induction, naming, and categorization. *Cognition*, *105,* 207-217.

Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences*, *31*, 689–749.

Thibodeau, P. H., Flusberg, S. J., Glick, J. J., & Sternberg, D. A. (2013). An emergent approach to analogical inference. *Connection Science, 25*(1), 27-53

Thibodeau, P. H. & Tesny, E. M., & Flusberg, S. J. (2014). Accounting for the relational shift and context sensitivity in the development of generalization. *Proceedings of the 36th annual meeting of the Cognitive Science Society, Quebec, Canada.*