Task-General Object Similarity Processes

Gavin W. Jenkins (gavin-jenkins@uiowa.edu)
Larissa K. Samuelson (larissa-samuelson@uiowa.edu)
John P. Spencer (john-spencer@uiowa.edu)
Department of Psychology, E11 Seashore Hall
Iowa City, IA 52245 USA

Abstract

The similarity between objects is judged in a wide variety of contexts from visual search to categorization to face recognition. There is a correspondingly rich history of similarity research and many known behavioral trends and models of similarity. Nevertheless, most similarity behaviors have been identified and tested only in a comparatively narrow set of unique contexts. This leaves open the question of the extent to which similarity judgments rely on common processes or resources and the specific nature of those processes if so. We tested three diverse yet well-established measures of object similarity using identical, psychometrically controlled stimuli and identical analyses across tasks. We found several consistent behavioral effects across tasks that provide clues as to the nature of task-general similarity processes and serve as diagnostic targets for computational models of similarity.

Keywords: similarity; psychology; concepts and categories; decision making; vision

Overview

Similarity judgments between objects occur across diverse contexts and tasks. Judging the similarity between perceived objects is necessary for following a map, identifying growth of a tumor between scans, noticing a defective product on an assembly line, or inventing new categories for novel objects.

The ubiquity of similarity judgments raises the question of whether they may derive from general, task-independent cognitive processes. If so, the specific nature of those processes and which similarity judgment behaviors they map to will be critical in better understanding the many tasks involving similarity judgments. One way to determine the nature of any core similarity processes is to test for task general behaviors. If tasks are diverse from one another, yet a set of behaviors is found to be common across them, this would suggest not only the existence of core processes, but that the behaviors in question derive from those core processes and offer clues about their nature.

Formal models in particular are well suited to investigating the nature of similarity judgment processes. Several formal models of similarity or that involve similarity exist (SIAM—Goldstone, 1994a; SUSTAIN—Love, Medin, & Gureckis, 2004; COVIS—Ashby, Paul, & Maddox, 2011; ALCOVE—Kruschke, 1992; the SME—Gentner & Markman, 1997), and although there are overlaps, few specific behaviors are captured by a wide variety of models. If any general processes of similarity judgments exist, however, then task-general behaviors likely associated with those general processes would serve as invaluable general target data for developing computational theories of those core processes.

Where do we begin, however, to search for evidence of general processes of similarity judgments? Many distinctive behaviors have been found in different similarity judgment contexts. Certainly, similarity judgments correlate with measurable differences in features between objects like color hue or size. This is qualitatively evident at face value, and feature comparisons have also been incorporated into formal, quantitative models since at least the early 20th century (Richardson, 1938). Details of these metrics have taken longer to establish, however. For example, there is evidence for both the use of Euclidean (Hout, Goldinger, & Ferguson, 2013) and of taxicab/city-block (Shepard, 1964) algorithms for determining the quantitative difference between two objects' feature values. Circular, wrap-around dimensions like angle or color hue present additional considerations. Analogous to city-block and Euclidean metrics, but within a single circular dimension, differences between objects could potentially follow a linear-type metric (Fig. 1) or a "chord length" metric (Shepard, 1962).

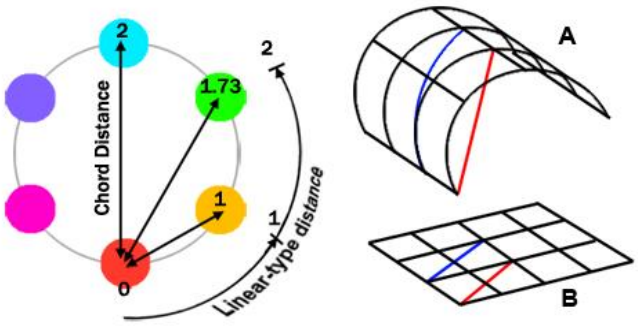

Figure 1: Left: differences along a circular dimensions can be measured or perceived in two ways: as if the dimension were linear around the circle, or as if along chords through the circle, yielding different ratios between pairs. Right: a two-dimensional feature space where one dimension is circular may be perceived as a curved (A) or flat (B) manifold when a subset of items are sampled, with red (chord) and blue (linear type) lines showing metric options.

Further complicating the study of similarity judgments, Tversky (1977) classically demonstrated that similarity judgments do not always follow pure metric assumptions at all. For example, China may be judged less similar to North

Korea then North Korea is to China. This "asymmetry" effect cannot be explained by static feature value differences, since the difference between two real numbers does not metrically change with order of presentation.

Since Tversky, a variety of other non-metric patterns have been observed in similarity judgments (in addition to metric patterns). A given magnitude of difference between two objects may become exaggerated in similarity judgments as the objects may become more similar along other dimensions or more "alignable" (Gentner & Markman, 1997). For example, differences between an atom and the solar system are easier to point out than differences between an atom and a toaster. A given difference between two objects can also be magnified in judgments if a person knows of many other objects near one or both in features (a high "neighborhood density," Krumhansl, 1978), such as differences between minor breeds of dogs. Similarities or differences between objects can also be perceived differently if attending to a certain feature dimension over others, like color over shape (Nosofsky, 1991).

A variety of tasks have been used to find and test similarity judgment behaviors. Pairwise tasks are particularly common, where pairs of two items from a larger set are judged at a time. Often, a ratings scale is used, or a two-alternative "same/different" choice. Alternatively, grouping, piling, and other arrangement methods allow participants to see a larger number of objects at once, then sort them into patterns to indicate similarity.

Establishing Task-General Similarity Processes

In order to test for the existence of task-general similarity processes among this wide array of tasks and behaviors, three steps must be taken. First, a set of candidate behaviors must be chosen that hold the potential to be consistently observed across tasks. We ruled out any behaviors already known to differ between tasks or ones that cannot be demonstrated in certain tasks. For example, Tversky's asymmetry effect, although seminal in the field, is difficult to observe in a multi-object arrangement task (Goldstone, 1994b), due to the geometric constraints of a workspace. We investigated the influence of basic feature value differences on similarity judgments, degree of feature dimensional bias, participants' sensitivity to circular dimensions, and the influence of neighborhood densities in feature space.

The second step in investigating possible core processes is to test all candidate behavioral effects redundantly across a diverse variety of available judgment tasks. Common behaviors despite diverse tasks suggests that those behaviors may hold clues to core processes shared across context. We chose three similarity tasks that are all widely used but differ from one another along key characteristics to cover a meaningful range of cognitive environments. Our first task used pairwise ratings, a task where participants judge object pairs by clicking on a 1-9 similarity scale. This task allowed for quick trials but was not time pressured.

The second task used binary "same/different" judgments. Compared to the ratings task, the same/different task was faster and less deliberative. It also included a time pressure element and had right and wrong answers.

The third task was the Spatial Arrangement Method (SpAM, Goldstone, 1994b; Hout, et al., 2013). SpAM involves arranging many objects at once into a pattern such that distances correspond to dissimilarities between any two objects. The task was the least time-pressured, allowed the highest response precision, and afforded the greatest ability to form intentional patterns of judgments, since the full context of all items was visible throughout the task.

The final step in testing for the existence of task-general similarity processes is to isolate the variable of task by utilizing a consistent environment of stimuli and analyses.

After describing the general task environment, we will describe the methodologies and findings of each task in detail, as well as the theoretical and modeling implications of task-general similarity.

Common Stimuli

In order to rule out stimulus-based confounds and to align tasks to allow for identical analysis, we used a single set of stimuli across tasks. The stimuli were shapes with two metric feature dimensions—color and shape. Fig. 2 depicts the full set. Both feature dimensions were psychometrically controlled in previous experiments and developed explicitly such that mathematical steps equal perceptual steps in these dimensions for average participants. The color dimension varied in hue according to the CIE l*a*b color space designed for a perceptually equal gradient, and the shape dimension consisted of circles modified by sine waves in a way that has been previously established to be perceived by participants as an equally spaced single, circular feature dimension of shape (Drucker & Aguirre, 2009).

Dimensions where mathematical steps equal perceptual steps means that the effect of any metric component of similarity can be quantitatively predicted, such as the influence of distances between objects along individual feature dimensions.

The full set of stimuli formed a 25 object grid across the two feature dimensions, as seen in Fig. 2. This is a commonly used pattern of stimuli for studies of object similarity due to its symmetry, uniformity, and predictability (Hout, Goldinger, & Ferguson, 2013; Kriegeskorte & Mur, 2012). However, one consistently observed factor in object similarity judgments is neighborhood density. Difference judgments of objects with many other objects near them in feature space are magnified compared to objects in sparse areas (Krumhansl, 1978; Love, Medin, & Gureckis, 2003).

To allow us to better test neighborhood density effects across tasks, we manipulated the subset of objects that participants worked with across two conditions. Half of all participants judged the similarities of objects in a basic grid pattern in feature space (a smaller 4x4 grid within the full 5x5 set). The other half of participants judged objects from a less symmetric, two-wide "L" shaped pattern consisting of the same number of objects as the grid pattern but with

overall less neighborhood density. Fig. 2 shows both patterns with colored overlays on the full stimulus set.

Stimuli were sampled from 180 degrees of each of their full circular dimensions, to make distances unambiguously unidirectional between pairs of objects. Both dimensions are still circular, however, and could be perceived as such.

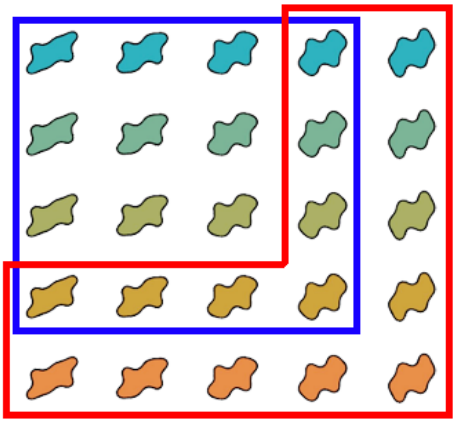

Figure 2: The full 5x5 set of stimuli, varying by sine wave-based shape and by l * a * b color space hue. The blue box indicates stimuli included in the grid subset, and the red box indicates stimuli in the "L" subset.

Common Analyses

Since our goal was to determine task-general similarity processes, we performed the same set of analyses on the data from each task.

Multidimensional Scaling

Multidimensional scaling (MDS) is an algorithm that takes as input a matrix of pairwise differences between all the pairs of items in a set.[1] It outputs positions for each item so that the distance between pairs of items are as proportional as possible to the input differences. Most notably, MDS provides visualization of the "shape" of a set of similarity judgments, such as the overall degree of metric uniformity and the compression, expansion, or warping of feature dimensions.

We performed MDS analyses of both group averages within experimental conditions and of individuals' data. Group analysis allowed us to visualize the strongest, most influential trends of judgments across tasks and between stimulus conditions (grid versus "L" subsets of items), while individual analysis indicated the range and variety of task "strategies." In particular, we were interested in how metrically organized group judgments were and whether participants showed dimensional modulation.

[1] Unlike our stimulus dimensions, the response scales in our tasks were not carefully psychometrically equalized. Therefore, non-metric, rank order MDS was appropriate for all tasks. The best fit from 50 random starts was used for MDS analyses.

Circular Dimension Sensitivity

We hypothesized that participants' similarity judgments might be consistently sensitive to the fact that our stimuli were sampled from circular dimensions.

If participants fail to notice a dimension's circularity, then they should judge each perceptually equal step linearly. If participants recognize curvature in the dimension, however, then they may judge pairs of objects according to chord distances "through the circle of the dimension." The right side of Figure 1 shows this distinction as applied to a grid of stimuli with one circular dimension. To test this, we analyzed raw similarity judgments by object pair, calculating root mean square errors to both linear and chord-based predictions to find the closer fit for each task.

Neighborhood Density Sensitivity

Previous studies (e.g., Krumhansl, 1978) suggest that in some cases, denser neighborhoods of objects in feature space can bias similarity judgments toward more "different" responses in that neighborhood than for the same number of feature steps of difference in a sparser neighborhood. We measured neighborhood density of each object pair as a count of both objects' immediately adjacent neighbors in feature space (up to 8 neighbors each in our stimulus set shown in Fig. 2).

We then correlated the neighborhood densities of each pair with the degree to which dissimilarity judgments differed from metric predictions. A positive correlation, therefore, would indicate inflated dissimilarities between objects in denser local neighborhoods in feature space.

Experiment 1 – Pairwise Ratings Task

The pairwise ratings task is the most common and straightforward of our three similarity judgment tasks. Most of the behaviors we tested originated from data using this task. The task is open-ended, unconstrained, without time pressure, and focused on pairs of objects at a time.

Methods

Twenty adult participants performed the pairwise ratings task. One participant was dropped due to MDS being unable to converge on a solution for his ratings.

Participants provided informed consent and were then seated in front of a computer terminal in an unadorned room. All instructions were on-screen. Participants were first exposed to the full set of 25 stimuli for context, one per second, and told to watch passively. Afterward, a ratings scale appeared and remained at the bottom of the screen throughout the rest of the experiment. The scale was labeled from 1-9, with 1 being labeled as least similar and 9 labeled most similar. Each trial consisted of a 500ms initial fixation cross, which was then replaced by two horizontally separated objects.[2] Participants were instructed to click the

[2] Separation was eight degrees of visual angle and stimuli subtended approximately five degrees of visual angle.

number on the ratings scale corresponding to how similar they thought each pair of objects was.

Each participant was grouped into one of the two stimulus conditions—the 4x4 grid subset of stimuli or the 2-wide "L" shaped subset of stimuli. Each participant received two trials each of every pair of objects within the subset of their condition, for a total of 272 trials per participant, randomly ordered per participant in one block.

After the experiment, all similarity ratings were inverted (10 – rating) to yield *dis*similarity ratings, and the set of common analyses described above was applied to the dissimilarity data from all participants.

Results

Before analyzing MDS results, we needed to determine the appropriate number of dimensions to use for fitting data in the MDS algorithm. We accomplished this by fitting multiple solutions at different numbers of dimensions and using "scree plots" shown in Fig. 3. Two-dimensional solutions were determined most valid across all conditions and experiments. An "elbow" is visible across conditions in the scree plots at two dimensions, indicating the point at which more dimensions begin to yield diminishing returns in fits that no longer justify the greater complexity of a higher dimensional model. Additionally, two dimensions is the simplest fit for all conditions that outperforms comparison results using random input data (black dots, Spence & Ogilvie, 1973).

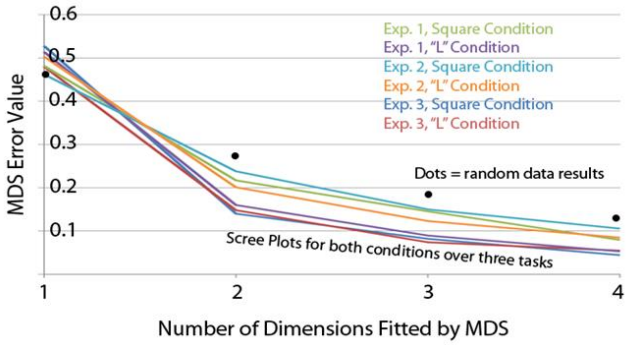

Figure 3: Scree plots of data from all conditions. All conditions outperform random data at two dimensions, and "elbows" are visible across conditions at two dimensions.

Group MDS solutions for the pairwise ratings task are shown in Fig. 4. Intersection points between lines represent object positions as placed by MDS. To aid visualization, green lines connect objects one step apart in color, and red lines connect objects one step apart in shape.

Both conditions show clear metric feature comparison influence: aside from a few items with swapped positions in the upper right, the grid conditions shows participants judging similarity roughly by a grid, and the "L" condition shows two unambiguous "arms" of objects, as expected for the "L" shaped subset of stimuli.

The "L" condition also shows non-linear warping of the predicted shape. The overall solution is "bent," meaning participants rated objects in the arms of the "L" differently than the feature values alone suggest. Both arms of the "L" also show exaggerated differences across shape (green lines are longer than red lines), suggesting dimensional modulation.

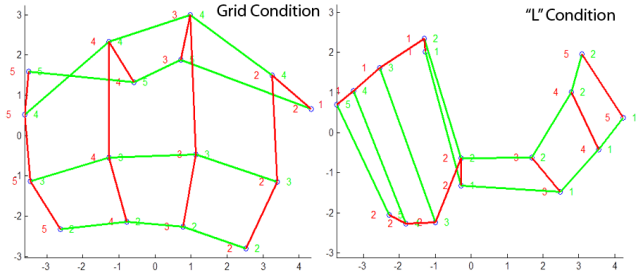

Figure 4: Group MDS solutions for Experiment 1.

Individual MDS solutions confirmed that the group patterns were not artifacts of averaging. Several individual participants showed grid like results in the grid condition, and several showed less organized but distinct "L" patterns.

Additionally, a number of individual results demonstrated dimensional modulation more dramatically than group results, yielding tightly clustered groups of objects along one dimension in MDS solutions. Fig. 5 contrasts a dimensionally even solution with a clustered solution. Overall, nine participants showed evenly spaced dimensional patterns, and eleven showed clustering patterns.

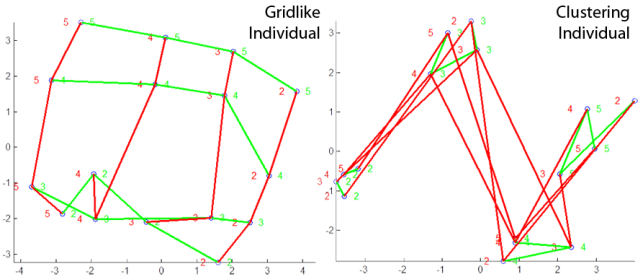

Figure 5: Individual MDS solutions of two individuals from the grid condition of Experiment 1, showing mostly evenly weighted dimensions (left) and clustering (right).

Circular dimension analysis showed that participants' similarity judgments matched chord-based predictions more closely than linear-type predictions (RMSE of 1.24 [chord] vs. 1.65 [linear] for the grid condition and 0.87 vs. 1.70 for the "L" condition). This suggests that participants likely perceived dimensions as more circular than linear.

Neighborhood density analysis showed moderate correlations between (observed – expected) dissimilarity ratings and the neighborhood densities of objects in a pair, r = 0.30 across conditions. This is consistent with predictions that high neighborhood density should magnify differences.

Experiment 2 – Pairwise Same/Different Task

Our second task was a binary judgment "same"/"different" ratings task. The task was speeded and designed to be

overall faster and with less opportunity for deliberative thought than the pairwise ratings task.

Methods

Twenty-two adult participants performed the pairwise same/different task. Two participants were dropped for not meeting a predetermined 70% accuracy cutoff (see below).

The procedure for the same/different task was identical to that of the ratings task, up until the point of test trials. Participants were still given a 500ms fixation cross followed by pairs of objects at a time, but instead of a ratings scale, participants were instructed to use keyboard keys "A" and "L" to indicate "same" or "different" (counterbalanced) for object pairs. Unlike in the ratings task, each trial had a correct answer. Participants were instructed that "'Different' pairs are different in EVERY way. 'Same' pairs are the same in ANY way." This particular rule was used, because it allowed a much more even distribution of "same" and "different" trials than if "same" were defined as identical, thus avoiding excessive repetition of identical pairs. Feedback was given at the end of each trial as a green check mark or a red "X" in the center of the screen for 500ms.

The task was "speeded" by the addition of a loud, annoying buzz that sounded whenever participants took longer than 1500ms from the onset of stimuli to respond. Despite the buzz, all trials continued until an answer was recorded, to avoid missing data.

Participants were again grouped into grid and "L" conditions. Each participant saw each pair of objects in their condition at least five times. Some randomly chosen "same" pairs appeared a sixth time, to equalize the number of "same" and "different" trials for each participant. Overall, participants in the grid condition completed 728 trials, and participants in the "L" condition completed 740 trials.

Dissimilarity ratings used in analyses were derived from the ratio of same:different responses across duplicate trials of each pair for a participant. If a given pair of objects was shown five times, for example, and a participant answered "different" to three of them, then the dissimilarity judgment for that object pair was interpreted as 3/5 = 0.6 out of 1.0.

Results

MDS group analysis is shown in Fig. 6. The results reflect those of Experiment 1, although with greater noise. In the grid condition, more green rows (objects sharing color) have swapped positions than in Experiment 1, but judgments are overall still dimensionally organized, with objects of shared feature following consistent orders and patterns across the stimulus set and dimensions being relatively perpendicular to one another. In the "L" condition, differences along shape are again exaggerated relative to differences along color, providing evidence of dimensional modulation.

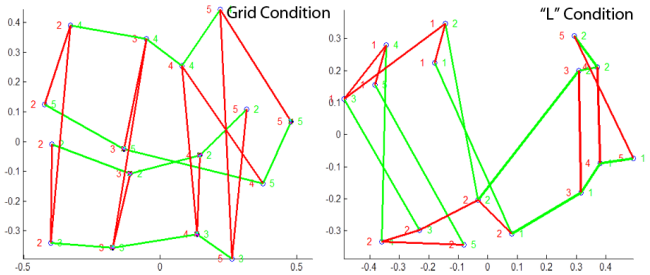

Figure 6: Group MDS solutions for Experiment 2.

Individual MDS analyses again confirmed that some individual results matched those of the group. Ten participants showed clustering patterns, and eleven showed even dimension ratios.

Circular analysis showed that participant behavior again fit more closely to chord-based predictions than linear-type predictions, indicating that judgments were sensitive to the circular feature dimensions used (RMSE of 0.2 vs. 0.23 for the grid condition and 0.16 vs. 0.23 for the "L" condition).

After subtracting out the contribution of feature values alone, neighborhood density analysis again showed moderate correlations between behavioral difference ratings and the neighborhood densities of objects in a pair, r = 0.21 across conditions, suggesting that judgments of differences were magnified in high density feature neighborhoods.

Experiment 3 – Spatial Arrangement Task

Our third task, SpAM, used distance relationships between many arranged objects to indicate similarity judgments. The task was slower, more contextual, and more allowing of thoughtful patterns of judgments than the previous tasks.

Methods

Twenty-three adult participants performed the SpAM task. One participant was dropped for not arranging any stimuli.

Participants used the same apparatus and were shown the same 25 item exposure phase as in the previous tasks. They were then presented with a single test trial. All 16 items in their condition (grid or "L") were displayed in columns along the sides of the screen, and a square workspace took up the center space. Participants were instructed to click and drag all items into the workspace, such that once all were placed, the distance between any pair of items would represent the dissimilarity between those items. Participants were allowed to move items after initial placement.

Dissimilarity ratings in the SpAM task were recorded as simply the pixel distances between each pair of item placements. These were then used to perform the common set of analyses.

Results

As seen in Fig. 7, group MDS results for the grid condition matched those of the pairwise experiments, taking the form of a noisy grid pattern with some swapped rows or columns. The "L" results showed a unique pattern. Feature

comparison is still apparent as a basis for judgments, but the two arms of the "L" were in this case laid out perpendicular to one another and with heavily swapped orders of feature values. Dimensional modulation is still suggested, but here, shape differences are exaggerated in only one arm, while color differences were exaggerated in the other.

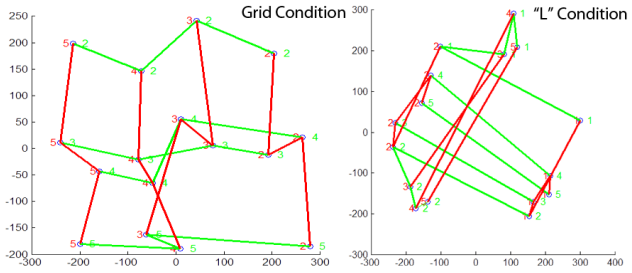

Figure 7: Group MDS solutions for Experiment 3.

Individual MDS analyses again confirmed the validity of group patterns in some participants, and revealed another strong split between even dimension patterns (fifteen) and clustering patterns (six).

Participants showed mixed sensitivity to circular dimensions, with better fitting RMSEs to linear distances in the square condition (520 vs. 624) and better fits to chord-based distances in the "L" condition (100 vs. 107).

Participants also again demonstrated a weak to moderate sensitivity to neighborhood densities, with neighborhood density measures correlating with feature-controlled dissimilarity ratings at r = 0.17 across conditions.

General Discussion

All three of our tasks showed evidence of a feature comparison influencing similarity judgments, roughly accurate representation of grid versus "L" stimulus patterns, uneven dimensional modulation to the extreme of clustered judgments in some participants, sensitivity to circular dimensions, and sensitivity to neighborhood density.

This large number of behaviors consistent across diverse tasks presents a strong case for the existence of core similarity processes. Furthermore, the identity of these particular behaviors may offer important clues as to the nature of those processes, particularly when targeted by formal, computational models of similarity judgments. Any general theory of similarity will likely require a flexible memory space that allows for both linear and circular feature space metrics (unlike traditional, Cartesian frameworks), and should describe processes allowing for modulating feature dimensions in linear (e.g., clustering) and non-linear (e.g., neighborhood density) ways.

Our findings also serve as a convenient quantitative modeling target due to the quantitative nature of our analyses, the psychometrically controlled and evenly perceptually spaced stimuli, and our consistent testing environment.

Acknowledgments

This research was supported in part by award number R01HD045713 from the Eunice Kennedy Shriver NICHD but does not necessarily represent the views of the NIH.

References

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. Pothos, E. M. & Wills, A. J., eds. *Formal approaches in categorization.* Cambridge University Press.

Drucker, D. M. & Aguirre, G. K. (2009). Different spatial scales of shape similarity representation in lateral and ventral LOC. *Cereb. Cortex,19*, 2269–2280.

Gentner, D. & Markman, A. B. (1997). Structure mapping in analogy and similarity. *Am. Psychol., 52(1),* 45-56.

Goldstone, R. L. (1994a). Similarity, interactive activation, and mapping. *J. Exp. Psychol.-Learn. Mem. Cogn., 20(1),* 3-28.

Goldstone, R. L. (1994b). An efficient method for obtaining similarity data. *Behav. Res. Meth. Ins. C., 26(4),* 381-386.

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *J. Exp. Psychol.-Gen., 142(1),* 256-281.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2003). SUSTAIN: A network model of category learning. *Psychol. Rev., 111,* 309-332.

Kriegeskorte, N. & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Front. Psychol.: Perception Science, 3,* 1-13.

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial data. *Psychol. Rev., 85(5),* 445-463.

Kruschke, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychol. Rev., 1,* 22-44.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cogn. Psych., 23,* 94-140.

Richardson, M. W. (1938). Multidimensional psychophysics. *Psychol. Bull., 35,* 659-660.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function, part II. *Psychometrika, 27(3),* 219-246.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *J. Math. Psychol., 1,* 54-87.

Spence, I. & Ogilvie, J. C. (1973). A table of expected stress values for random rankings in nonmetric multidimensional scaling. *Multivar. Behav. Res., 8(4),* 511-517.

Tversky, A. 91977). Features of similarity. *Psychol. Rev., 84(4),* 327-352.