Journal of Experimental Psychology: General
2013, Vol. 142, No. 1, 256–281


0096-3445/13/$12.00 DOI: 10.1037/a0028860

# The Versatility of SpAM: A Fast, Efficient, Spatial Method of Data Collection for Multidimensional Scaling

Michael C. Hout, Stephen D. Goldinger, and Ryan W. Ferguson
Arizona State University

Although traditional methods to collect similarity data (for multidimensional scaling [MDS]) are robust, they share a key shortcoming. Specifically, the possible pairwise comparisons in any set of objects grow rapidly as a function of set size. This leads to lengthy experimental protocols, or procedures that involve scaling stimulus subsets. We review existing methods of collecting similarity data, and critically examine the spatial arrangement method (SpAM) proposed by Goldstone (1994a), in which similarity ratings are obtained by presenting many stimuli at once. The participant moves stimuli around the computer screen, placing them at distances from one another that are proportional to subjective similarity. This provides a fast, efficient, and user-friendly method for obtaining MDS spaces. Participants gave similarity ratings to artificially constructed visual stimuli (comprising 2–3 perceptual dimensions) and nonvisual stimuli (animal names) with less-defined underlying dimensions. Ratings were obtained with 4 methods: pairwise comparisons, spatial arrangement, and 2 novel hybrid methods. We compared solutions from alternative methods to the pairwise method, finding that the SpAM produces high-quality MDS solutions. Monte Carlo simulations on degraded data suggest that the method is also robust to reductions in sample sizes and granularity. Moreover, coordinates derived from SpAM solutions accurately predicted discrimination among objects in same–different classification. We address the benefits of using a spatial medium to collect similarity measures.

*Keywords:* multidimensional scaling, similarity, spatial cognition

*Supplemental materials:* http://dx.doi.org/10.1037/a0028860.supp



Modern psychological theorizing often relies largely on a notion of similarity, or a sense of "sameness" among stimulus items (Goldstone & Medin, 1994; Medin, Goldstone, & Gentner, 1993). For example, predictions from memory theories (Gillund & Shiffrin, 1984; Hintzman, 1986, 1988; Hintzman & Ludlam, 1980; Hout & Goldinger, 2011), lexical access and production (Goldinger, 1998; Goldinger & Azuma, 2004), and categorization (Goldstone, 1994b; Goldstone & Steyvers, 2001; Nosofsky, 1986; Nosofsky & Palmeri, 1997) often hinge upon degrees of similarity between a stimulus and exemplars stored in memory. Shepard's universal law of generalization (Shepard, 1987, 2004) posits that the probability of generalizing from one item to the next decays (exponentially) as a function of their decreasing similarity. Proximity in psychological space can also generate stimulus confusions. For example, the well-documented "other-race effect" in face perception may arise from a psychological space that is more densely clustered for other-race faces, causing them to appear excessively similar to one another (Byatt & Rhodes, 2004; Goldinger, He, & Papesh, 2009; Levin, 1996; Papesh & Goldinger, 2010; Valentine, 1991).

Although similarity is a ubiquitous theoretical construct, it is both labile and challenging to quantify. How similar are the colors blue and green? To what degree do you resemble your mother rather than your father? Such questions are difficult to answer with direct, quantitative measures. Moreover, similarity estimates are highly context sensitive; the perceived similarity between items can change dramatically given different "backdrops" for comparison. For example, a pole vaulter and a boxer are not particularly similar, but if they were both members of the Norwegian Olympic team, parading in the opening ceremonies with teams from all other nations, their perceived similarity would doubtless increase. To faithfully estimate peoples' impressions of similarity, psychologists often rely on subjective similarity ratings, which are analyzed with multidimensional scaling (MDS) or a related approach (see Shepard, 1980). By analyzing overt ratings of perceived similarity, the frequencies of interitem confusions, or the latencies of correct discriminations between items, we can obtain a quantitative approximation regarding the similarity of items.

## MDS

To provide context for the present investigation, we begin with a brief review of MDS. Our goal is not to provide a comprehensive

This article was published Online First July 2, 2012.

Michael C. Hout, Stephen D. Goldinger, and Ryan W. Ferguson, Department of Psychology, Arizona State University.

Support was provided by National Institutes of Health Grant R01 DC 004535-11 to Stephen D. Goldinger. We thank Anthony Barnhart and Donald Homa for helpful suggestions. We are particularly grateful to Megan Papesh for help in stimulus creation and to Robert Goldstone for invaluable suggestions for improvement. We also thank Jessica Dibartolomeo, Evan Landtroop, Holly Sansom, Kyle Brady, Monica Poore, Geoff McKinley, Ciara Francis, and Alexi Rentzis for assistance in data collection.

Correspondence concerning this article should be addressed to Michael C. Hout or Stephen D. Goldinger, Department of Psychology, Arizona State University, Box 871104, Tempe, AZ 85287-1104. E-mail: michael.hout@asu.edu or goldinger@asu.edu

background (see Borg & Groenen, 1997; Kruskal & Wish, 1978; Rabinowitz, 1975); rather we aim to contextualize the present research, emphasizing a few challenging problems in MDS. As Shepard (1980) noted, since Isaac Newton's (1704) treatise on optics, it has been suggested that psychological (or perceptual) similarity is best approximated with spatial configurations, wherein the proximity of any two items reflects their perceived similarity. For instance, following Newton's suggestion, spectral hues can be represented on a "color wheel," with red proximal to orange, but distal from green, etc.

MDS is an exploratory data analysis technique that satisfies Newton's desire to represent similarity spatially; it uses various forms of data (matrices of item-to-item similarities or dissimilarities) to create spatial maps, intended to convey the relationships among items (Attneave, 1950; Mugavin, 2008; Richardson, 1938). More technically, MDS is a set of statistical techniques (e.g., Kruskal, 1964a, 1964b; Torgerson, 1958, 1965; Shepard, 1962a, 1962b, 1964) that generate geometric representations of the stimuli, with one point representing each item and the interitem distances representing the similarity (or "psychological distances") between them. There are many instantiations of MDS algorithms (e.g., PROXSCAL, Busing, Commandeur, & Heiser, 1997; ALSCAL, Young, Takane, & Lewyckyj, 1978; INDSCAL, Carroll & Chang, 1970; PREFSCAL, Busing, Groenen, & Heiser, 2005). Each performs MDS in slightly different ways, or for different purposes, and most are implemented in data analysis software. Generally, when the algorithms are executed, a random starting configuration is generated (in *k*-dimensional space, as specified by the analyst), and the proximities among points are calculated. Ideally, these proximities will respect the similarity ratings obtained from the data. A stress function (e.g., S-Stress, Stress I, Stress II; the choice depends on the particular MDS algorithm) is then calculated, quantifying the fit between the distances in space and the input proximities, with lower values indicating closer fits. MDS algorithms seek to minimize the stress function by iteratively moving the items in space, attempting to increase fidelity to the input data (Rabinowitz, 1975). This process is repeated (sometimes hundreds or thousands of times) until the configuration is optimal.[1]

The outcome of MDS (i.e., the spatial map) provides a visual representation of the underlying dimensions of a data set (Nosofsky, 1992), reflecting the important relationships within the data (Ding, 2006). By subjectively examining the MDS solution, one tries to identify which dimensions may have been used for object comparisons. For instance, when providing similarity ratings between animals, a person may (implicitly or explicitly) appreciate their respective sizes, ferocity, colors, habitats, etc. As such, a potential MDS solution may reflect the primary dimensions of *size* and *ferocity*: Small, docile animals (e.g., a mouse) may be located far from small, aggressive animals (e.g., a piranha), and farther still from large, aggressive animals (e.g., a lion). Examination of MDS solutions can reveal such key dimensions, or confirm prior hypotheses about their importance (Giguère, 2006).

To appreciate how MDS works, it is useful to imagine examining a map. One could easily use a map to generate a table of distances between all pairs of cities; a far harder task would be to do the reverse, creating a map from a set of distances (for a full treatment of this often-cited example, see Jaworska & Chupetlovska-Anastasova, 2009; Kruskal & Wish, 1978). This is what MDS achieves: Using proximity data (e.g., geographic distances), it generates a configuration of points that respects these pairwise ratings. In this case, the outcome would be a map with cities configured in a manner that respects their geographic locations; the dimensions would correspond to north–south and west–east directions. Although this example is psychologically uninteresting, it illustrates two important characteristics of MDS: (a) that it reduces an overwhelming data set (e.g., a large matrix of city-to-city proximities) into a manageable form and (b) that it provides spatial representations that allow simultaneous appreciation of many interrelations among data points.

With respect to the present research, a key issue is that psychological measurements are rarely as precise as measuring distances between cities. Two further aspects of MDS therefore merit brief consideration: choice of dimensionality and interpretation of solutions. First, the researcher must decide how many dimensions the algorithm should use. Increasing the dimensionality (i.e., the number of coordinate values used to locate points in space) adds degrees of freedom to the movement of individual items, thereby increasing the information represented by the solution (and decreasing its stress). In the animal example, one could plot the items *mouse*, *piranha*, and *lion* along a single dimension of *size*, collapsing over the dimension *ferocity*. From this configuration of points on a line, we would glean that a mouse is similar to a piranha and both are dissimilar to a lion. Only by adding the *ferocity* dimension can we appreciate the dissimilarity of "mouse–piranha" and the similarity of "piranha–lion." To choose the right number of dimensions, researchers will create scree plots, displaying stress as a function of dimensionality. Stress always decreases with added dimensions, but a useful heuristic is to look for the "elbow" in the plot, the value at which added dimensions cease to substantially improve fit (Jaworska & Chupetlovska-Anastasova, 2009; see also Lee, 2001, for a Bayesian approach to dimensionality determination). This conservatism is applied because increasing the dimensionality of an MDS solution is not always beneficial. As Rabinowitz (1975) noted, a common goal of MDS is to yield solutions in sufficiently low dimensionality to permit visual examination. Therefore, choosing the correct dimensionality will depend on stress, but also on interpretability (Kruskal & Wish, 1978). In essence, one must strike a balance between finding a good solution and finding one that is interpretable.

Second, MDS solutions vary, even when the algorithms are implemented on the same data set multiple times. No single solution will provide the best fit (unless one is using a single data matrix; Giguère, 2006).[2] For instance, one solution from our map example might display eastern cities on the right and western cities on the left. A second attempt may reverse these dimensions, or

[1] There are many different stopping rules; some algorithms allow for a maximum number of iterations to be input by the user. Others will stop once the change in stress value from solution to solution has dropped below a certain threshold, indicating that the fit is no longer improving with subsequent iterations.

[2] The exception to this rule is that when using only a single matrix of similarities, the MDS technique is the same as eigenvector or singular value decomposition in linear algebra, wherein there is a "perfect" solution (Giguère, 2006). We focus on the case of multiple matrices because it is more likely that psychologists collect data from multiple participants.

invert the solution. The interpretation of these solutions is the same, however, as the relations among points will remain stable. But with psychological data (that typically uses multiple participants, and noisy measurements), these interitem relationships may change across scaling attempts. A good solution is stable, such that it closely matches configurations across attempts, irrespective of its orientation along the dimensions. More important to note, MDS algorithms are blind to the "truth" of their solutions. The analyst must determine the coherence and utility of the solution, with dimensions that are subject to interpretation (see Green, Camone, & Smith, 1989; Schiffman, Reynolds, & Young, 1981).

## Methods for Collecting Similarity Data

Similarity is inherently a dynamic (and sometimes slippery) notion (Goldstone, Medin, & Gentner, 1991; Goldstone, Medin, & Halberstadt, 1997; Spencer-Smith & Goldstone, 1997). For any two objects, there are potentially infinite features shared between them (Tversky, 1977). Measuring the subjective similarity among objects can therefore be difficult, and there are different techniques for collecting such data. Jaworska and Chupetlovska-Anastasova (2009) distinguished between *direct* and *indirect* methods. In direct methods, participants knowingly rate or classify items, such as sorting stimuli into categories (e.g., Faye et al., 2004, 2006; Rosenberg, Nelson, & Vivekananthan, 1968). A proximity data matrix would be derived by counting how often stimuli are categorized together, across participants. By contrast, indirect methods typically involve data captured by secondary empirical measurements, such as stimulus confusability. For example, participants might briefly see pairs of stimuli for *same–different* judgments (e.g., Shepard, 1963). Proximities would be estimated by the percentage of trials wherein different items are mistakenly identified as the same (e.g., Wish & Carroll, 1974), or by speed of accurate responses (e.g., Papesh & Goldinger, 2010).

Perhaps the most commonly used direct method is simply to ask people to numerically rate object pairs (typically via Likert scales), collecting ratings for every possible pairwise combination of stimuli (hereafter denoted the *pairwise method*). For example, participants may respond "1" when the items are very similar, "9" when they are very different, and use intermediate numbers to represent varying levels of similarity.[3] Typically, participants are encouraged not to overthink their responses, but rather to make swift, "gut-feeling," similarity estimates. Such instructions are designed to discourage feature listing, explicit decisions about underlying dimensions, or strategy changes over the course of a session. Undoubtedly, the pairwise technique is useful and simple to implement. However, as Goldstone (1994a) noted, it confers several disadvantages. First, it is inefficient, as the number of required comparisons (to create a full matrix) increases as a quadratic function of set size: For a stimulus set of $n$ items, $n(n - 1)/2$ ratings must be made by each participant. Although it is possible to collect partial matrices from participants (see Spence & Domoney, 1974), researchers typically prefer to obtain complete matrices, because they provide more robust and precise solutions (Giguère, 2006). This inherent inefficiency creates lengthy experimental protocols. To preview our experiments, it took participants approximately 20–30 min to rate only 25 stimuli (300 comparisons) using the pairwise method. Second, using such lengthy protocols may cause participants to change strategies over time, become fatigued, or simply disengage and rate arbitrarily (Johnson, Lehmann, & Horne, 1990). Third, people are not particularly adept at using discrete rating systems. Likert scales limit the responses that people can make, thereby limiting resolution. Fourth, people often remember their previous responses and may be influenced by them (Parducci, 1965; Wedell, 1995). For instance, when presented with a pairing that strikes a participant as a "4," the participant may consider how often that number was used recently and shift the current response to a "5" (see Helson, 1964; Helson, Michels, & Sturgeon, 1954).

In response to these concerns, Goldstone (1994a) proposed a novel method for collecting similarity data. He suggested that researchers could benefit from utilizing the spatial nature in which people tend to conceptualize similarity (Casasanto, 2008; Lakoff & Johnson, 1980). This method (hereafter denoted the *spatial arrangement method*, or SpAM) involves presenting multiple stimuli (e.g., images) to the participant at once, randomly arranged on a computer screen. The participant's task is to arrange the items on the screen (using the computer mouse), such that their interitem distances reflect their perceived similarity. When the participant is finished organizing the space, a proximity matrix is derived from item-to-item Euclidean distances (i.e., dissimilarities). In essence, SpAM allows people to create their own MDS maps in two-dimensional planes.

SpAM has intuitive appeal, as participants can use space to their advantage, and it provides an extremely fast way to collect similarity ratings. The same stimuli that require a 20- to 30-min pairwise protocol can be scaled in 4–5 min with SpAM. It is also very efficient: Each movement simultaneously changes the relationships of the moved object to all other stimuli present on-screen. Of greater importance, SpAM allows quick appreciation of the entire stimulus set, such that all judgments can be made without variations in context. Finally, SpAM allows graded, high-resolution responding, limited only by the resolution of the computer monitor. Although the method has been occasionally applied (Busey & Tunnicliff, 1999; Levine, Halberstadt, & Goldstone, 1996; Perry, Samuelson, Malloy, & Shiffer, 2010), we find it surprising that SpAM has not been widely used. Perhaps researchers are more comfortable with the tried-and-true pairwise method, or are unable to implement SpAM. As such, this investigation had two primary goals: (a) to critically examine the quality of solutions derived by SpAM, relative to pairwise methods, and (b) to assess two new methods for collecting similarity data that combine aspects of pairwise and SpAM techniques.

## The Current Investigation

To compare methods for collecting similarity data, we first constructed stimuli with well-controlled perceptual dimensions. We then collected data using various methods and created comparable MDS solutions to assess how faithfully each method reproduced the original sets. As there exists no method for reveal-

[3] There are several variants of this method, such as *magnitude estimation* (Stevens, 1971), wherein one pair is chosen as a standard for other pairs to be judged against, and the *anchor stimulus method* (Borg & Groenen, 1997), which involves iteratively choosing items that are most similar to the "anchor" and removing them from the stimulus set until all items have been selected.

ing the "true" underlying structure of a psychological space (Goldstone & Medin, 1994), nor any analysis that perfectly reveals the quality of an MDS solution, our strategy was to amass converging evidence using several analytical techniques. Following Goldstone (1994a), we correlated interitem distances across methods to assess levels of agreement across MDS solutions. We also used deviational analyses and cluster analyses (described in detail later) to assess the quality of our solutions, relative to "ideal" organizations of the stimulus spaces.

In Experiment 1, we constructed two sets of stimuli: *wheels*, which were based on stimuli used by Shepard (1964), and *bugs*. We expected SpAM to perform well for two-dimensional stimuli, as it involves arranging objects on a two-dimensional plane, but we were unsure whether it would recover more than two underlying dimensions. As such, the wheels and bugs each consisted of two stimulus subsets (rated by different participants), including both two-dimensional and three-dimensional structures. Beyond evaluating SpAM, our second goal was to evaluate two new methods for collecting similarity data, as described in Experiment 1. In Experiment 2, we examined scaling for conceptual stimuli, consisting of two sets of animals (presented as text, not images). The first set (*categorical animals*, from Hornberger, Bell, Graham, & Rogers, 2009) consisted of animals that are easily categorized along two primary dimensions: an *avian* dimension (animals were either birds or not) and a *habitat* dimension (animals that live primarily on land or in/on water). The second set (*continuous animals*, from Henley, 1969; Howard & Howard, 1977) was chosen to compare techniques on stimuli with no salient dimensions. Finally, in Experiment 3, we assessed how well the solutions derived from SpAM and the pairwise method would predict stimulus discrimination. Participants rated the similarity of our bugs and novel, computer-generated faces; we then used the distances derived from the MDS spaces to predict speed and accuracy in two same–different discrimination tasks.

## Experiment 1

In Experiment 1, we collected similarity ratings on four sets of stimuli (two- and three-dimensional wheels and bugs). Following Shepard (1964), we constructed stimuli to vary along a small number of perceptually distinct and salient dimensions (see also Garner, 1974; Shepard, 1991). Our goal was to evaluate how well each of four methods (pairwise, SpAM, *total-set pairwise*, and *triad*) would discover these dimensions. In the total-set pairwise method, we modified the pairwise technique, endowing it with one of the advantages from SpAM. Specifically, by presenting all stimuli at once, participants are "instantly calibrated" to dimensional ranges of the stimuli. This approach places each decision in the greater context of the entire stimulus set. As Goldstone (1994a) noted, in pairwise ratings, the values assigned to the first few object pairs are arbitrary because the entire context only emerges with continued experience. The total-set pairwise method alleviates that concern (see also the *conditional rank-ordering task*, or the *free sorting method*; Ahn & Medin, 1992; Schiffman et al., 1981).

Our second new technique, the triad method, follows Chan, Butters, and Salmon (1997), who showed participants three items at once and asked them to choose which two were most similar. Proximity matrices were derived by counting how often each incorporated pair was chosen as most similar. In Experiment 1, we added the SpAM interface to the method from Chan et al. Participants were shown triads of objects, and created small-scale MDS maps of the three items by moving them around the screen. Thus, people were free to pair items together, but could also place them equidistant to one another (if they deemed no pairing as having higher similarity), or could apply any asymmetric organization that seemed correct.

### Method

**Participants.** Experiment 1 included 183 Arizona State University students who participated for partial course credit. All participants had normal or corrected-to-normal vision.

**Design.** Each participant provided similarity ratings for three stimulus sets. Because SpAM takes very little time, participants performed it twice, and also completed one of the lengthier procedures (pairwise, total-set, or triad). They first performed SpAM on a randomly selected set of stimuli, then rated a different set of items using another technique, then performed SpAM on a third set of stimuli. Short breaks were provided between sessions. Although we collected data for all three stimulus types (wheels, bugs, and animals) simultaneously, we consider the animal stimuli in Experiment 2, for clarity. Selection of methods (pairwise, total-set, triad), stimulus type (wheels, bugs, animals), and subset (two-dimensional vs. three-dimensional, categorical vs. continuous) was random, with the constraint that no individual participant scale the same stimulus type more than once.

**Stimuli.** Stimuli were line drawings: schematic one-spoked wheels and rudimentary bugs, as shown in Figure 1.

***Wheels.*** The two dimensions of variation were the thickness of the lines composing the drawing and the angle of the spoke. For the three-dimensional stimuli, we added a dimension of hue, filling the wheels with varying shades of red.

***Bugs.*** The two dimensions of variation were the number of legs and the shading of the back and head. For the three-dimensional stimuli, we added variation in the curvature of the antennae. Two-dimensional sets included 25 items; three-dimensional sets had 27 objects.

**Apparatus.** Data were collected with up to eight computers simultaneously; each was equipped with identical software and hardware (Gateway E4610 PC, 1.8 GHz, 2 GB RAM). Dividing walls separated subject stations on either side to reduce distraction. Each display was a 17-in. (43.18-cm) NEC (16.0 in. [40.64 cm] viewable) CRT monitor, with resolution set to 1280 × 1024 and refresh rate of 60 Hz. Display was controlled by an NVIDIA GE Force 7300 GS video card (527 MB). E-Prime (Version 1.2; Schneider, Eschman, & Zuccolotto, 2002) was used to control stimulus presentation and collect responses.

**Procedure.**

***Pairwise method.*** Participants were shown two items at a time and provided similarity ratings using a Likert scale (1 = *most similar*, 9 = *most dissimilar*). Each possible pairwise combination was presented in random order, for a total of 300 trials for two-dimensional stimuli and 351 trials for three-dimensional stimuli. Placement of items on the left or right of center was also randomized.

***SpAM.*** Participants were shown all the stimuli simultaneously, organized in discrete rows, with randomized item place-

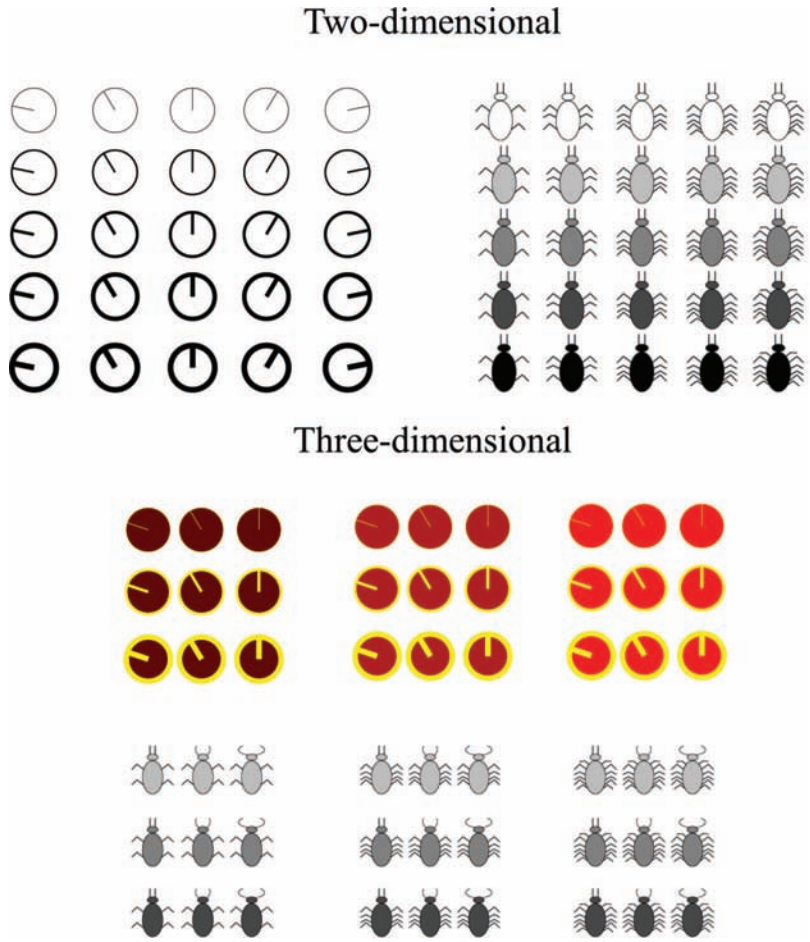


*Figure 1.* All stimuli used in Experiment 1. The top portion of the figure shows the two-dimensional wheels and bugs, and the bottom portion shows three-dimensional items.

ment. They were instructed to drag and drop objects, organizing the space such that the distance among items was proportional to their perceived similarity, with closer denoting greater similarity. Once participants finished arranging the items, a right mouse-button press completed the trial. To avoid accidental termination, participants were asked if the space was satisfactory, indicating responses via the keyboard, and were given more time as needed. Only a single trial was administered.

***Total-set pairwise method.*** The total-set pairwise method followed the same general procedure as the pairwise method. However, rather than present two items at a time, we presented all the stimuli simultaneously (organized in discrete rows, with randomized item placement). Participants gave similarity ratings to a single pair of items at a time, which were indicated by highlighting a black border around the to-be-scaled objects. Therefore, the number of trials matched the pairwise technique.

***Triad method.*** The triad method followed the same general procedure as SpAM. However, rather than present all stimuli simultaneously, we showed three items per trial, presented in an equilateral triangle at the center of the monitor. Trials were randomized with the constraint that each item could appear with any other item only once, as determined with a Steiner system (see Rumov, 2001): In a three-item Steiner system, the number of triads is equal to $n(n - 1)/6$, where $n$ is the total number of items. Thus, participants completed 100 and 117 trials for the two- and three-dimensional stimuli, respectively.

## Results

As noted earlier, our strategy was to provide converging measures regarding the quality of the MDS solutions. We first present results for two-dimensional stimuli, followed by three-dimensional stimuli. In each section, we show MDS spaces derived from each method of data collection, followed by the results of correlational and deviational analyses.[4]

[4] Although stress is a useful quantification of agreement between the MDS solution and its input proximities, we chose not to report stress values for two reasons. First, stress varies according to many different factors, such as the number of stimulus pairs or data matrices (Giguère, 2006). Accordingly, our stress values would not be directly comparable across methods or stimuli. Second, more informative analyses derive from a focus on the solutions themselves, rather than a blind measure of congruence with the data.

**MDS algorithm and choice of dimensionality.** All MDS solutions were derived with the PROXSCAL algorithm (Busing et al., 1997) with 1,000 random starts, via SPSS 15.0 (SPSS, 2006). This algorithm uses a least squares method of representation and can accommodate multiple data sources. As Davidson (1983) noted, selecting the number of dimensions for scaling depends largely on substantive knowledge that the analyst brings to bear. Because our stimuli were created with specific dimensions in mind (without supplementary visual characteristics), we did not rely on scree plots to choose dimensionality, but simply plotted solutions according to the input dimensions of the stimuli.

**Two-dimensional stimuli.** Figure 2 shows the MDS spaces generated by each method. The *x*-axis of each plot is the primary dimension, with the secondary dimension plotted along the *y*-axis. Note that, across methods, there was not always agreement about the primary dimension; sometimes, for instance, participants deemed the "thickness" dimension as most salient for the wheels, whereas others found "inclination" most important.

***Correlations.*** For each solution, we calculated the item-to-item distances for each stimulus pair, measured in Euclidean space, with 300 and 351 distances for two- and three-dimensional stimuli, respectively. These values were correlated across methods to measure the consistency of the solutions. Higher positive correlations indicate that two solutions have comparable layouts, regardless of rotation around the coordinate axes. Table 1 shows the Pearson product–moment correlation coefficients of each method to one another, for each stimulus set separately. All correlations were positive and significant ($p < .01$), and by Cohen's (1988) norms, all were moderate to large in effect size. The total-set method (.77) produced the largest average correlation, followed in order by SpAM (.75), pairwise (.71), and triad (.61). Correlations for bug stimuli (.91) were, on average, higher than those for wheel stimuli (.50). Thus, as suggested by the regularities across panels in Figure 2, all the methods generated roughly similar solutions.

***Deviations.*** Because our stimuli were constructed with specific dimensions, it was possible to derive "ideal spaces" for comparison to the solutions derived from each method. The ideal spaces had perfect, orderly arrangements of stimulus items, with equal intervals between the levels of each dimension; in essence, they were perfect squares or cubes. To assess the quality of the solutions, we derived ideal spaces that matched the height and width (and depth, for three-dimensional stimuli) of the solutions generated by each method, separately, and placed the coordinates at equal intervals along each dimension.[5] They were also rotated to match the orientation of each solution. We then calculated a deviation score for each stimulus item, measuring the Euclidean distance from the PROXSCAL coordinates to its "ideal location" (see Figure 3). Deviation values are arbitrary because no basic unit of measurement is present in MDS (Rabinowitz, 1975); however, because PROXSCAL generates solutions of approximately equal size across methods, the deviation scores are directly comparable. These values were entered into a 4 × 2 (Method × Stimuli) analysis of variance (ANOVA), with each stimulus item treated as a participant. Method and Stimuli were between- and within-subjects factors, respectively.

The ANOVA revealed a main effect of Method, $F(3, 96) = 30.80$, $\eta_p^2 = .49$, $p < .001$, with the smallest average deviations for the SpAM (.06), followed by total-set (.15), pairwise (.41), and triad (.43) methods. There was also a main effect of Stimuli, $F(1, 96) = 35.83$, $\eta_p^2 = .27$, $p < .001$, with smaller deviations for bugs (.18), relative to wheels (.35). The interaction of Method × Stimuli was also reliable, $F(3, 96) = 8.59$, $\eta_p^2 = .21$, $p < .001$. Although this deviation analysis does not perfectly measure the quality of the observed solutions, it does comport with subjective impressions regarding the organization of the spaces. For instance, our impression is that the two-dimensional bug solution from the total-set method is more orderly, relative to that of the triad method: This intuition is confirmed by the deviation analysis.

**Three-dimensional stimuli.** In the supplemental materials, Figures A2–A5 show the three-dimensional MDS spaces derived from each method. The solutions are shown in two panels: The left panels show the primary dimension along the *x*-axis and secondary along the *y*-axis. The right panels show the tertiary dimensions along the *y*-axis.

***Correlations.*** The observed correlations (see Table 1) were again all significantly positive ($p < .01$) and ranged from small or moderate to large in effect size. The total-set method again produced the largest average correlation (.44), followed by SpAM (.42), pairwise (.41), and triad (.28). Correlations for bug stimuli (.41) were higher, relative to wheels (.36).

***Deviations.*** The deviation analysis (see Figure 4) revealed a main effect of Method, $F(3, 104) = 4.00$, $\eta_p^2 = .10$, $p < .01$, with the smallest average deviations for the pairwise method (.60), followed by SpAM (.62), total-set (.65), and triad (.77). There was a main effect of Stimuli, $F(1, 104) = 10.38$, $\eta_p^2 = .09$, $p < .01$, with smaller deviations to bugs (.60), relative to wheels (.72). The interaction of Method × Stimuli was reliable, $F(3, 104) = 4.03$, $\eta_p^2 = .10$, $p < .01$. Inspection of individual solutions shows that it is not always clear which dimensions of the solutions most closely correspond to each stimulus characteristic. To give each solution the best chance of obtaining small deviation scores, we calculated scores for every possible combination of rotations and selected the combination that minimized the deviations scores for each solution.

**Monte Carlo simulations.** In Experiment 1, SpAM produced orderly solutions that were comparable in organization to the traditional pairwise technique. However, because each MDS solution is unique, our results may have been fortuitous. To address this possibility, we ran Monte Carlo simulations wherein scaling algorithms were applied to the pairwise and SpAM data 25 times each, per stimulus set.

We also attempted to isolate the characteristics of SpAM that elicit its high-quality solutions by performing Monte Carlo simulations on modified SpAM data. We considered two major aspects of SpAM that differ from the pairwise method: its granularity and sheer data mass. The data from SpAM have high granularity because the resolution of individual responses is greatly increased, relative to Likert scales. That is, one method allows nine scale

[5] It is likely that these ideal spaces are, to some degree, overly constrained. Specifically, the linearity assumption is likely too strict, and a more appropriate space may be one wherein there is unequal spacing between levels of each dimension, or skewed (e.g., curved) edges. However, because we used this analysis simply to complement subjective inspection of the MDS spaces, we deemed that square or cube ideal spaces provided the simplest, assumption-free metric.

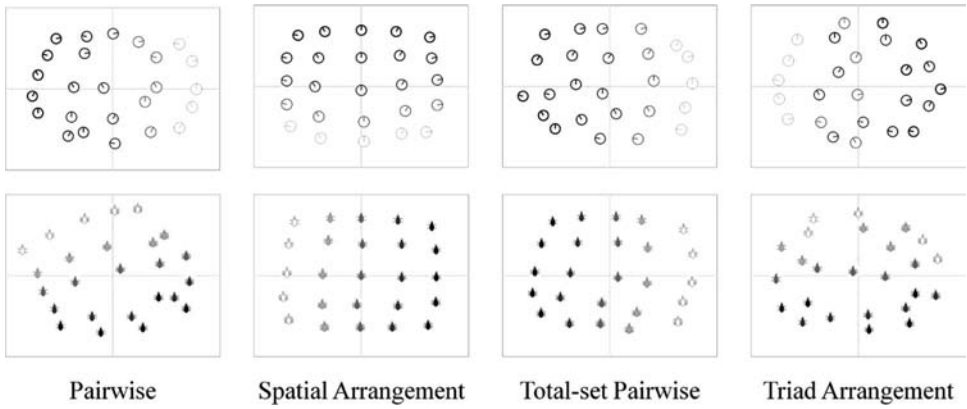


*Figure 2.* Two-dimensional multidimensional scaling spaces generated by each method, from Experiment 1. The top row of solutions presents the wheels; the bottom row presents the bugs.

values per trial, whereas the other allows hundreds of pixels. It is possible that having more continuous values promotes more accurate proximity matrices. In the *reduced granularity* simulations, we transformed the SpAM data into measures akin to a Likert scale by rounding each value to the nearest hundred (e.g., a distance of 430 pixels was reduced to a score of 4). Next, because SpAM takes so little time, the foregoing solutions represent large sample sizes (between 80 and 90 participants), relative to the pairwise method (between 10 and 20 participants). Having more representative samples could clearly contribute to the quality of the SpAM solutions. In the *reduced subjects* simulations, we reduced sample sizes to levels that matched the pairwise technique by randomly selecting subsets of participants for analysis. Finally, in the *both reduced* simulations, we reduced both granularity and sample sizes to ascertain whether the spatial interface alone was sufficient to engender accurate solutions.[6]

Table 1
*Pearson Product–Moment Correlation Coefficients for Interitem Distance Vectors, From Experiment 1*

| Method | SpAM | Total-set | Triad |
|---|---|---|---|
| | Wheels | | |
| Two-dimensional | | | |
| Pairwise | .47 | .55 | .45 |
| SpAM | | .90 | .31 |
| Total-set | | | .34 |
| Three-dimensional | | | |
| Pairwise | .44 | .44 | .21 |
| SpAM | | .52 | .25 |
| Total-set | | | .32 |
| | Bugs | | |
| Two-dimensional | | | |
| Pairwise | .96 | .96 | .86 |
| SpAM | | .99 | .86 |
| Total-set | | | .85 |
| Three-dimensional | | | |
| Pairwise | .53 | .53 | .31 |
| SpAM | | .51 | .24 |
| Total-set | | | .32 |

*Note.* All correlations are significant at $p < .01$. SpAM = spatial arrangement method.

**Within-method correlations.** We first calculated the interitem distances from each solution and correlated them within methodologies. Essentially, we tested to what degree the solutions generated by a single method were consistent across iterations. High positive correlations indicate stability within a data set. Supplement Figures A7 and A8 show histograms of correlation coefficients for two- and three-dimensional wheels and spokes, for each of our five simulations (refer to supplement Table A2 for the values used to generate these histograms, and for the percentages of correlations that were reliable). The highest stability was shown for the pairwise and SpAM simulations (average correlation coefficients of .70 for both), followed by reduced subjects (.65) and then reduced granularity and both reduced (.57 for both). Two-dimensional stimuli (.90) produced more stable solutions, relative to three-dimensional (.38), and bugs (.70) were more stable, relative to wheels (.57).

**Cross-method correlations.** We next correlated the interitem distances from each simulation with those of the pairwise method, using it as a baseline for comparison. Our questions were twofold: How well does SpAM correlate with the pairwise method across multiple iterations, and how does degradation of the SpAM data affect the agreement of the solutions? The most agreement was shown by the SpAM and reduced subjects simulations (.59 for both), followed by both reduced (.56) and reduced granularity (.51). Consistent with the within-method correlations, agreement was higher for two-dimensional stimuli (.75), relative to three-dimensional (.37), and for bugs (.67), relative to wheels (.46). (See supplemental materials for further details and analyses.)

[6] It should be noted that our procedures were biased against the reduced subjects and both reduced simulations. This is because for each of these simulations, we sampled random sets of data for SpAM solutions; our only constraint was that each participant's data be used at least once. By contrast, in each of the other simulations (pairwise, SpAM, reduced granularity), the same data were used in each simulation, providing total consistency in the similarity ratings provided.

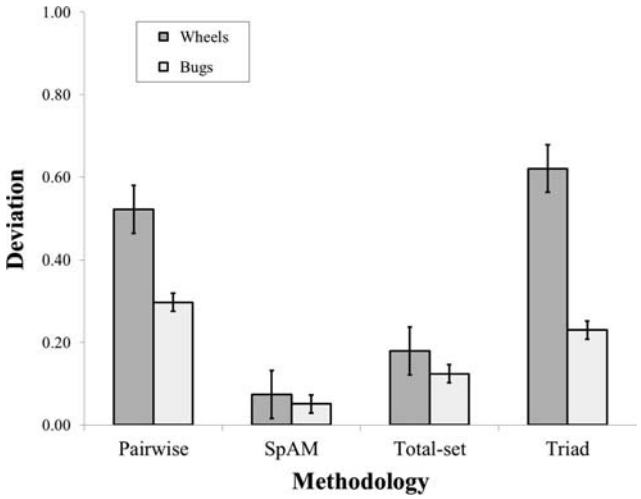


*Figure 3.* Results of the deviation analysis from Experiment 1, two-dimensional stimuli. Error bars represent ±1 standard error of the mean. SpAM = spatial arrangement method.

**Deviations.** As before, we calculated deviation scores for each solution, measuring the distance from each stimulus item to its ideal location (see Figure 5). These values were entered into three-way mixed-model ANOVAs (for two- and three-dimensional stimuli, separately): Simulation (pairwise, SpAM, reduced granularity, reduced subjects, both reduced) × Stimuli (wheels, spokes) × Iteration (1–25). Simulation was a between-subjects factor, whereas Stimuli and Iteration were within-subjects factors.

***Two-dimensional stimuli.*** The deviation analysis revealed a main effect of Simulation, $F(4, 120) = 73.36$, $\eta_p^2 = .71$, $p < .001$, with smallest average deviations for SpAM (.06), followed in order by reduced granularity (.07), reduced subjects and both reduced (both .21), and pairwise (.33). There was also a main effect of Stimuli, $F(1, 120) = 548.67$, $\eta_p^2 = .82$, $p < .001$, with smaller deviations to bugs (.13), relative to wheels (.22). There was a main effect of Iteration, $F(24, 97) = 39.50$, $\eta_p^2 = .91$, $p < .001$, and all the interactions were significant ($F$s $> 14$, $p$s $< .001$). (For brevity, we do not discuss these effects, but the full data set and histograms are found in the supplemental materials, Table A4 and Figure A11.)

***Three-dimensional stimuli.*** The analysis showed a main effect of Simulation, $F(4, 130) = 11.72$, $\eta_p^2 = .27$, $p < .001$, with the smallest deviations in pairwise and reduced subjects simulations (both .60), followed by SpAM (.64), both reduced (.68), and reduced granularity (.77). There was a main effect of Stimuli, $F(1, 130) = 45.99$, $\eta_p^2 = .26$, $p < .001$, with smaller deviations to bugs (.63), relative to wheels (.69). The main effect of Iteration was significant, $F(24, 107) = 6.19$, $\eta_p^2 = .58$, $p < .001$, as were each of the interactions (all $F$s $> 3$, $p$s $< .001$).

**Individual differences analysis.** As Goldstone (1994a) noted, a potential shortcoming of SpAM is that instructions about using the space may be interpreted differently across individuals. Indeed, subjective inspection of the solutions suggests that people "solved" the scaling problem in various ways, due to either differing interpretations of instructions or strategies used to construct arrangements. Consider Figure 6: Some participants (e.g., the top-left panel) produced spaces that appear highly structured and tend to correlate strongly with others. Other spaces (e.g., the top-right panel) appeared less well structured; such spaces reflect some appreciation for the stimulus dimensions, but correlate with others more weakly (or less often). Finally, there were participants (e.g., the bottom panels) whose spaces appeared unstructured, or exhibited "clustering" along one dimension without appreciation for another. How should we reconcile these individual differences, and what is the best way to integrate such participants' data into aggregate solutions? In this section, we show that these potential outliers are not particularly problematic for SpAM, and suggest a way to identify participants who produce irregular solutions.

Our general strategy was to identify outliers by analyzing the extent to which each participant's MDS space correlated with all others, for the SpAM and pairwise methods. This entailed several steps: (a) We created individual MDS spaces for each participant and derived vectors of interitem distances from those spaces. (b) Next, we correlated the distance vectors across all participants (for each stimulus set and methodology, separately). (c) For each participant, we then calculated two scores: their average correlation coefficient and the proportion of correlations that were reliable. (d) Finally, we rank-ordered the participants and (in separate analyses) identified those with the lowest average correlations or proportions of reliable correlations. The bottom 25% were identified as outliers; this is likely an overly conservative estimate of "outlying" data, but we chose this strict criterion for illustrative purposes. Many participants were considered outliers based on both measures, but some were outliers according to one measure and not another.

Once we identified these irregular participants, we created two MDS spaces, one that excluded the outliers and another for the outliers themselves. To gauge the extent to which these participants skewed the aggregate results, we then correlated

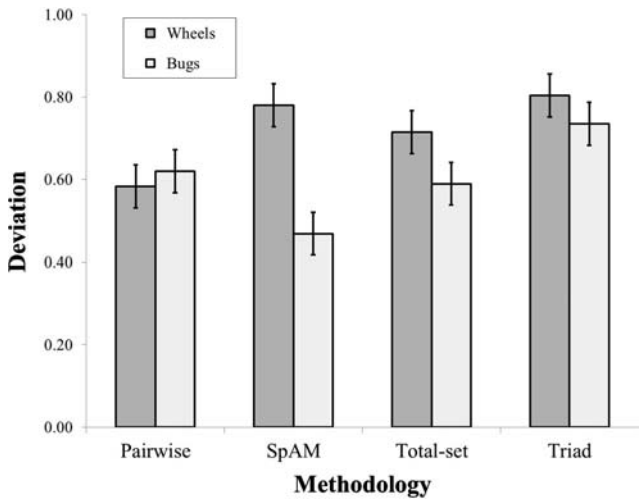


*Figure 4.* Results of the deviation analysis from Experiment 1, three-dimensional stimuli. Error bars represent ±1 standard error of the mean. SpAM = spatial arrangement method.

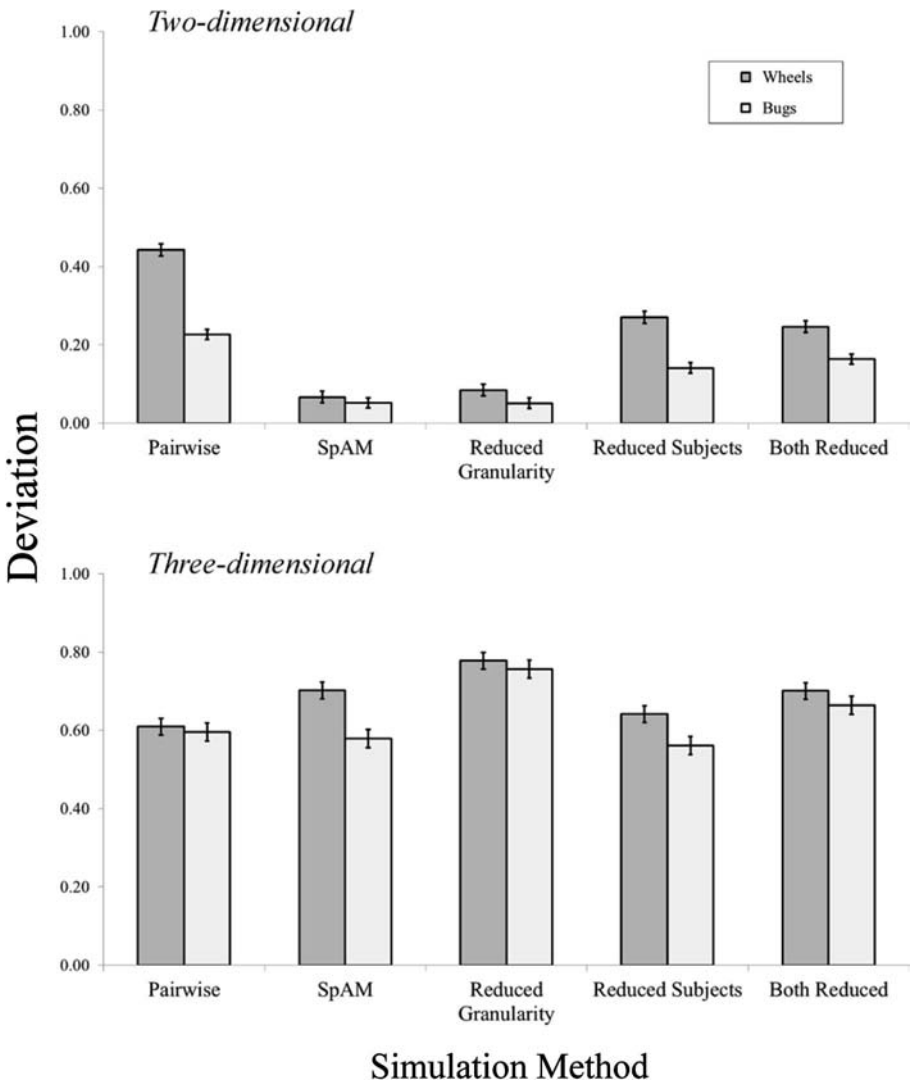


*Figure 5.* Results of the deviation analysis from Experiment 1, Monte Carlo simulations. Two-dimensional stimuli are shown in the top panel; three-dimensional stimuli are shown on the bottom. Error bars represent ±1 standard error of the mean. SpAM = spatial arrangement method.

the interitem distances from these exclusionary solutions to the entire data set. Figures 7 and 8 show the results of this procedure for the two-dimensional bugs, obtained by SpAM and pairwise methods, respectively. For SpAM, removal of irregular participants had very little effect on the solutions: Relative to the aggregate data, the "regular" solutions were in high agreement ($r$ = .99 for both exclusion criteria). By contrast, the irregular solutions deviated more strongly from the organization of the aggregate solution ($r$s = −.04 and .18 for the mean $r$ and proportion-significant criteria, respectively). For the pairwise method, removal of irregular participants also had little effect on the solutions (relative to the aggregate data, $r$s = .97 and .99 for the same criteria, respectively). However, the irregular pairwise solutions also showed a moderate or high resemblance to the aggregate data ($r$s = .90 and .43 for the same criteria, respectively). See supplemental materials (Table A6) for analyses concerning three-dimensional bugs.

## Discussion

In Experiment 1, we critically examined a novel method of collecting similarity ratings proposed by Goldstone (1994a), in addition to evaluating two new, hybrid techniques. The results are easily summarized: (a) The correlations of interitem distances across methods show that each method provides solutions with roughly comparable organizations. To the extent that the pairwise method is an appropriate baseline for comparison, SpAM provides solutions that closely agree with well-established procedures. The total-set method also produced comparable solutions, and to a lesser extent, the triad method did as well. (b) In comparison to ideal MDS spaces, SpAM produced solutions that were most orderly for two-dimensional stimuli (owing, no doubt, to its use of a two-dimensional plane). When three-dimensional stimuli were rated, SpAM no longer produced superior solutions, but nevertheless generated solutions that

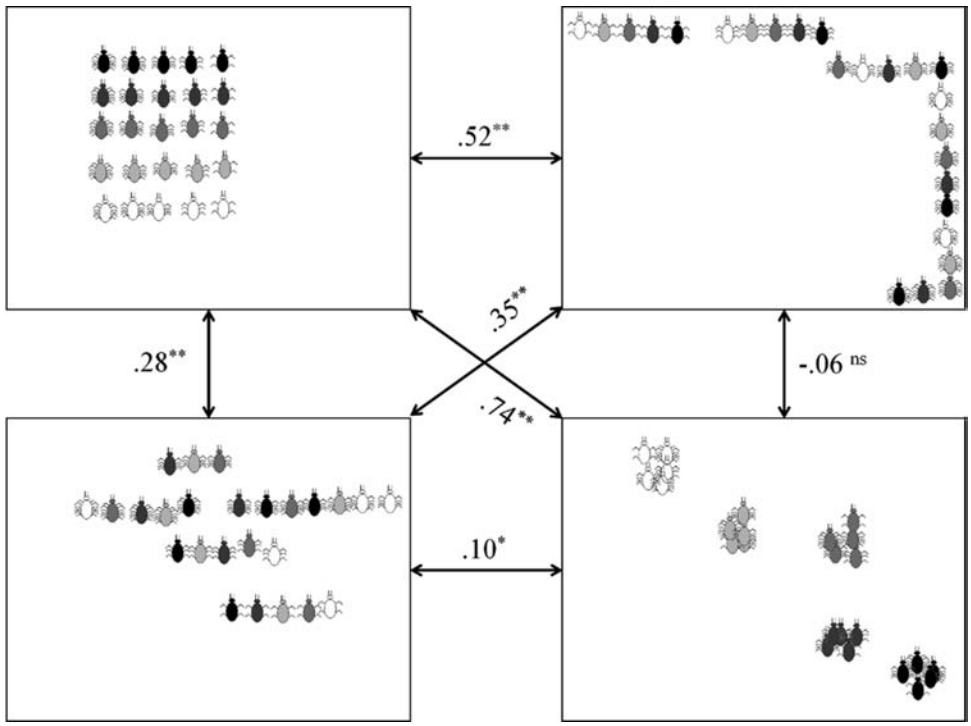


*Figure 6.* Spatial arrangement method spaces (for two-dimensional bugs) created by four participants, from Experiment 1. The numbers represent Pearson product–moment correlation coefficients ($^{*}p < .05$; $^{**}p < .01$) between the item-to-item distance vectors from each solution.



were comparable to the other techniques. (c) The Monte Carlo simulations revealed high levels of stability (across iterations of the scaling algorithms) for pairwise and SpAM methods, and show that the stability of SpAM was reduced slightly by reducing sample size or granularity. (d) The simulations also showed that SpAM consistently correlates highly with solutions from the pairwise technique; these correlations are generally unaffected by reductions in sample size or granularity. Finally, (e) the individual differences analyses suggest that participants approach SpAM in different ways. However, removing even a full quarter of the least regular data did not drastically affect the overall solutions provided by SpAM. We revisit the issue of individual differences in the General Discussion.

## Experiment 2

In Experiment 2, we further examined the foregoing methods, now considering conceptual stimuli (animal names) with loosely established underlying dimensions. In short, we assessed how well the methods would perform on psychologically interesting materials. As in Experiment 1, we present cross-method correlations of the interitem distances. We also added data derived by latent semantic analysis (LSA; Landauer & Kintsch, 2003), as another baseline condition. For the categorical stimuli, we also added a cluster analysis, designed to measure the degrees of separation between semantic categories.

### Method

The participants, apparatus, design, and procedure were identical to those in Experiment 1, as all data were collected simultaneously. The only exception was that, in analysis, we included data derived from LSA as an additional baseline. LSA uses statistical computations on a large text corpus to extract the contextual-usage meaning of words. Its core assumption is that shared contexts of appearance can reflect the similarity among words (Landauer, Foltz, & Laham, 1998; Wolfe & Goldman, 2003). We obtained (for each stimulus set) an LSA term-to-term comparison matrix (using a topic space that included "general reading up to 1st year college," with 300 factors) and fed these matrices into the MDS algorithms, just like data derived from our actual participants.

**Stimuli.** We used two sets of animal names (see supplement Table A1). *Categorical animals* (from Hornberger et al., 2009) were easily categorized along two dimensions. Each animal was either a bird or a four-legged animal (*avian* dimension), and was either a land or water dweller (*habitat* dimension). The *continuous animals* (from Henley, 1969) were selected with no obvious categorical classification or any prespecified underlying structure. Both stimulus sets included 25 items. Thus, for the pairwise and total-set methods, 300 trials were necessary to acquire a complete data matrix from each participant; 100 trials were necessary for the triad method, and one trial was used for SpAM.

### Results

We first present the results from categorical animals, followed by continuous animals. All MDS solutions were again derived with PROXSCAL (Busing et al., 1997) with 1,000 random starts. We scaled the categorical animals in two dimensions because they

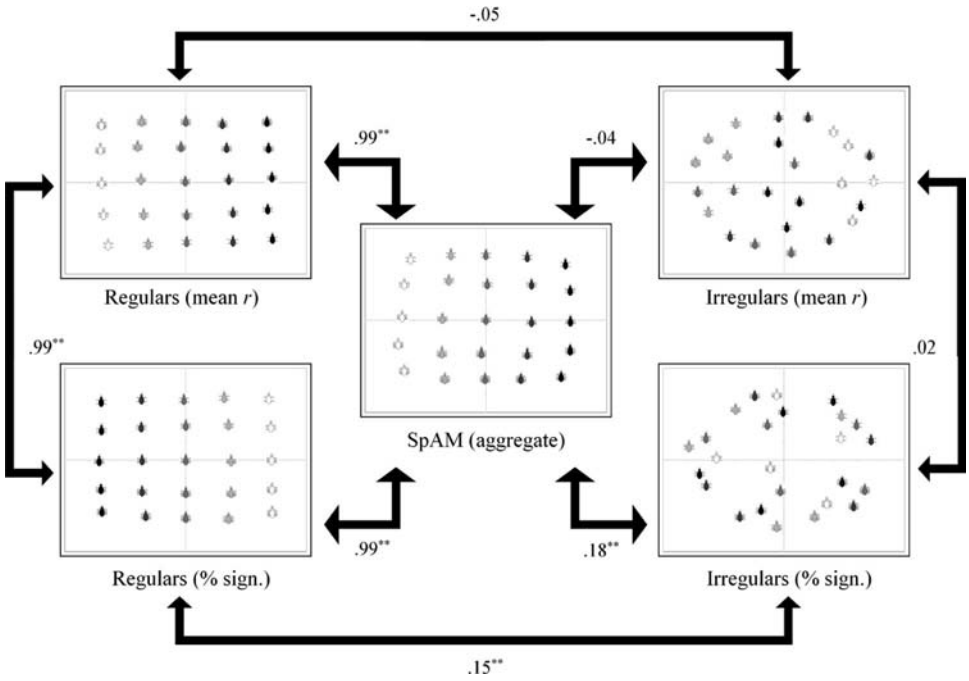


*Figure 7.* Multidimensional scaling spaces for two-dimensional bugs, derived by the spatial arrangement method (SpAM; Experiment 1). The left panels show solutions that exclude outlier participants; the right panels are solutions from only outliers. The numbers represent Pearson product–moment correlation coefficients (** $p <$ .01) between the item-to-item distance vectors from each solution.

were selected with two specific dimensions in mind. For the sake of consistency, we also scaled the continuous animals in two dimensions. Although solutions with higher dimensionality may have yielded additional information, we used two-dimensional solutions for ease of interpretation, and so both stimulus sets could be analyzed comparably.[7]

**Categorical animals.** Figure 9 shows the MDS spaces generated by each method and the LSA data. Again, the *x*-axis of each plot presents the primary dimension, and the secondary dimension is plotted along each *y*-axis. The categories are shown with different symbols: Birds are displayed with diamonds, and nonbirds with circles; land dwellers are shown with filled symbols, and water dwellers with unfilled symbols. From this, it is easy to rapidly identify each hypothesized dimension, for example, by comparing the locations of diamonds and circles.

***Correlations.*** Table 2 shows the Pearson product–moment correlation coefficients (for item-to-item distances generated by each MDS space) of all methods. All correlations were significantly positive ($p < .01$) and were moderate to large effects. The highest average correlation was produced by the total-set method (.71), followed by pairwise and SpAM (both .69), triad (.60), and LSA (.44).

***Cluster analyses.*** To estimate how well each MDS solution discovered the hypothesized underlying category structures, we calculated the average item-to-item distance from each stimulus item to (a) members of its own specific category (e.g., duck to goose); (b) items that matched on the habitat dimension, but not the avian dimension (e.g., duck to turtle); (c) items that matched on the avian dimension, but not the habitat dimension (e.g., duck to chicken); and (d) items that were opposites on both dimensions (e.g., duck to squirrel). A solution with consistent categorization should have small within-category distances, large distances to items that are opposites on both dimensions, and intermediate values for items that share singular features (see Figure 10). These values were tested in a two-way mixed-model ANOVA (again, treating each stimulus item as a subject): Method (pairwise, SpAM, total-set, triad, LSA) × Cluster (within-category, off-habitat, off-avian, off-both). Method and Cluster were between- and within-subjects factors, respectively.

The ANOVA revealed no effect of Method, $F(4, 120) < 1$, $p = .90$, reflecting the fact that PROXSCAL generates solutions of approximately equal size. We observed a main effect of Cluster, $F(3, 118) = 372.99$, $\eta_p^2 = .90$, $p < .001$, with the shortest average distances to within-category members (0.46), followed by off-habitat (0.80), off-avian (0.97), and off-both (1.16). The Method × Cluster interaction was reliable, $F(12, 312) = 4.42$, $\eta_p^2 = .13$, $p <$

[7] Indeed, the annulus structure of the solutions suggests that perhaps the stimuli should be scaled in a higher dimensionality. As such, for each method, we also scaled the data in one to five dimensions and assessed the stress values at each level. Scree plots showed an elbow that consistently appeared at Dimension 2. Although adding a third dimension reduced stress, the largest reductions occurred from one dimension to two dimensions (an average of 62% and 59% reduction in overall stress, for categorical and continuous animals, respectively). The reduction from two to three dimensions was modest (15% and 17%), and the reduction from three to four dimensions was minor (7% and 8%). This indicates that two-dimensional analyses were most likely appropriate.

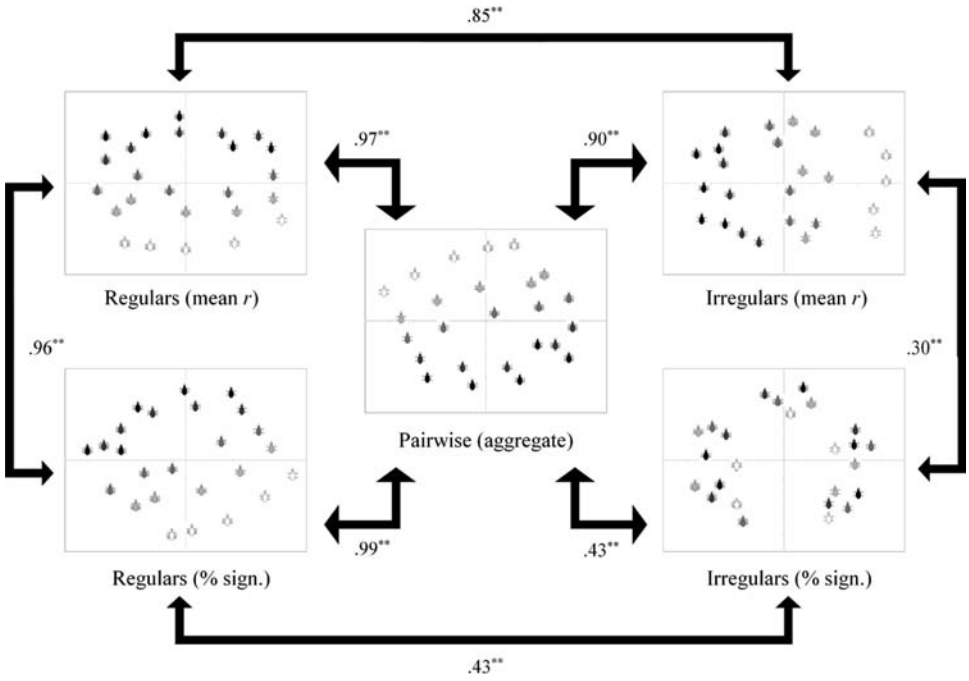


*Figure 8.* Multidimensional scaling spaces for two-dimensional bugs, derived by the pairwise method (Experiment 1). The left panels show solutions that exclude outlier participants; the right panels are solutions from only outliers. The numbers represent Pearson product–moment correlation coefficients (** $p < .01$) between the item-to-item distance vectors from each solution.



.001, driven largely by the comparatively poor performance of LSA. The effect of Cluster shows that our analysis reliably quantified the classifications drawn out by each solution, with distance in space increasing as a function of featural dissimilarity. Moreover, it affirms subjective inspection of the solutions, for example, showing that SpAM created the tightest categorical clusters (i.e., the smallest within-category distances and largest off-both distances).

**Continuous animals.** The MDS spaces generated from each data set are shown in Figure 11, with primary and secondary axes shown on the *x*- and *y*-axes, respectively. The correlations between each method were significantly positive, ranging from small to large effect sizes (see Table 2). The total-set method (.54) produced the highest average correlations, followed by pairwise (.51), SpAM (.48), triad (.42), and LSA (.23).

**Monte Carlo simulations.** Experiment 2 again showed that SpAM produced MDS spaces that were comparably organized, relative to more time-intensive methods. Moreover, this congruence was not limited to perceptual similarity but extended to conceptual similarity. To verify again that the findings were not simply a fortunate outcome, we performed Monte Carlo simulations wherein scaling algorithms were repeatedly applied (25 iterations each) to the pairwise and SpAM data, and to modified SpAM data (with reduced granularity, sample size, or both).

We first calculated the interitem distances from each solution and correlated them, within methods, to estimate internal stability. The pairwise method (.68) showed the highest average correlation, followed by SpAM (.65), reduced subjects (.60), reduced granularity (.54), and both reduced (.51). Categorical animals (.77) produced more stable solutions, relative to the continuous animals (.41). We next correlated the interitem distances from each simulation with those of the pairwise method to see how consistently SpAM correlates across iterations, and to examine how the systematic removal of its potential advantages might affect its performance. The highest agreement was shown by the full SpAM (.59), followed by reduced subjects (.55) and then reduced granularity and both reduced (both .52). The supplemental materials (Table A3 and Figures A9 and A10) show histograms of the correlation coefficients for each simulation, along with values used to generate each histogram.

**Cluster analysis.** As before, we calculated scores from each solution, measuring the average distance from an animal to members of its own category, to members sharing one feature, and to those sharing no features (for the categorical animals only). These values (see Figure 12) were entered into a three-way mixed-model ANOVA: Simulation (pairwise, SpAM, reduced granularity, reduced subjects, both reduced) × Cluster (within-category, off-habitat, off-avian, off-both) × Iteration (1–25). Simulation was a between-subjects factor, and Cluster and Iteration were within-subjects factors. There was no main effect of Simulation, $F(4, 120) < 1$, $p = .99$. There was a main effect of Cluster, $F(3, 118) = 1260.49$, $\eta_p^2 = .97$, $p < .001$, with the smallest distance to within-category items (0.41), followed by off-habitat (0.80), off-avian (0.98), and off-both (1.19). There was no main effect of Iteration, $F(24, 97) < 1$, $p = .99$. The Cluster × Method, Cluster × Iteration, and Simulation × Cluster × Iteration interactions were all significant (all $F$s $> 1.8$, $p$s $< .05$). For brevity, we do not discuss these effects (see supplemental materials, Table A5 and

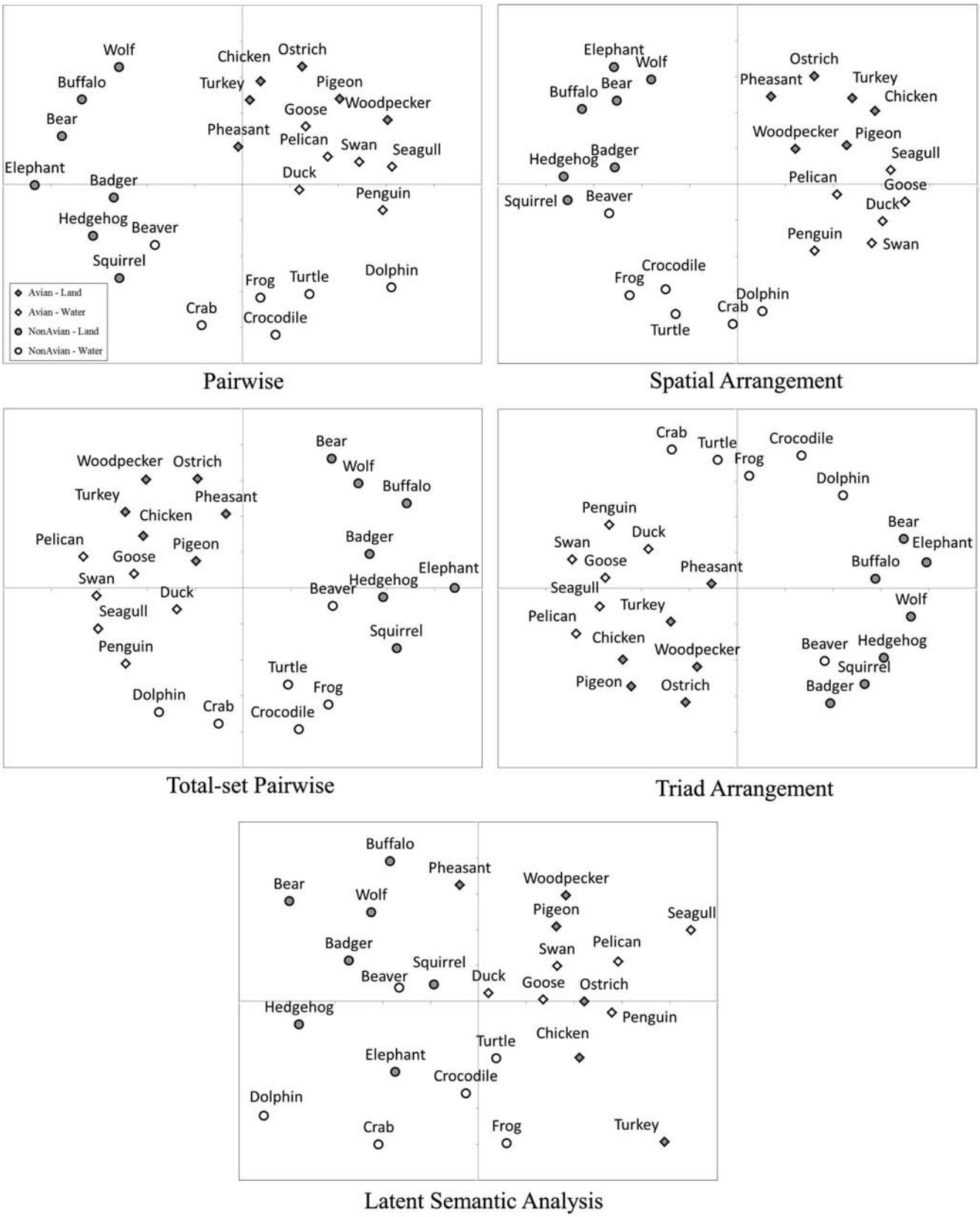


*Figure 9.* Categorical animal multidimensional scaling spaces generated by each of the four methodologies, and by latent semantic analysis, from Experiment 2.

Table 2
*Pearson Product–Moment Correlation Coefficients for Interitem Distance Vectors, From Experiment 2*

| Method | SpAM | Total-set | Triad | LSA |
|---|---|---|---|---|
| Categorical animals | | | | |
| Pairwise | .81 | .85 | .64 | .47 |
| SpAM | | .83 | .70 | .40 |
| Total-set | | | .66 | .49 |
| Triad | | | | .39 |
| Continuous animals | | | | |
| Pairwise | .61 | .75 | .48 | .18 |
| SpAM | | .58 | .45 | .29 |
| Total-set | | | .56 | .25 |
| Triad | | | | .20 |

*Note.* All correlations are significant at $p < .01$. SpAM = spatial arrangement method; LSA = latent semantic analysis.

Figure A12). Two important points can be gleaned from this analysis. First, we replicated the findings from Experiment 2, showing that the pairwise and SpAM techniques produce satisfactory categorical discrimination, with distance in space increasing as a function of featural dissimilarity. Second, neither reducing the sample size nor reducing the granularity of the SpAM data greatly hindered its ability to generate conceptually organized MDS solutions.

**Individual differences analysis.** As in Experiment 1, we identified outlier participants by examining how well each person's MDS space correlated with the others (for the SpAM and pairwise methods). Then we created solutions for the 75% most regular and 25% least regular participants, separately. In Experiment 1, we included solutions derived by two criteria: the average correlation coefficient per participant and the percentage of significant correlations. In Experiment 2, the outlying participants were the same people, for both methods, according to both criteria. Figure 13 shows the results (for brevity, we limit our analysis to the categorical animals; see supplemental materials, Table A6).

For SpAM, removing irregular data again had a minimal effect on the solutions. The regular solution was in high agreement with aggregate data ($r = .91$), whereas the irregular solution weakly correlated with the aggregate ($r = .12$). For the pairwise method, removal of irregular participants also had a minor effect; the regular solution was highly correlated with the aggregate ($r = .85$). In contrast to the SpAM irregular data, the pairwise irregular solution was more highly correlated with the aggregate solution ($r = .42$).

## Discussion

In Experiment 2, we examined conceptual, rather than perceptual, similarity. The results can be summarized as follows: (a) Correlating the interitem distances across methods showed that the solutions were comparably organized. Notably, both the SpAM and total-set methods reliably produced strong correlations (the triad method performed less well). LSA provided the least consistent data; this is not altogether surprising, however, as large corpora may lack the precision necessary to adequately mimic human performance. (b) With respect to categorical animals, each method clustered the stimuli such that interitem distances tended to increase as a function of featural dissimilarity. (c) The Monte

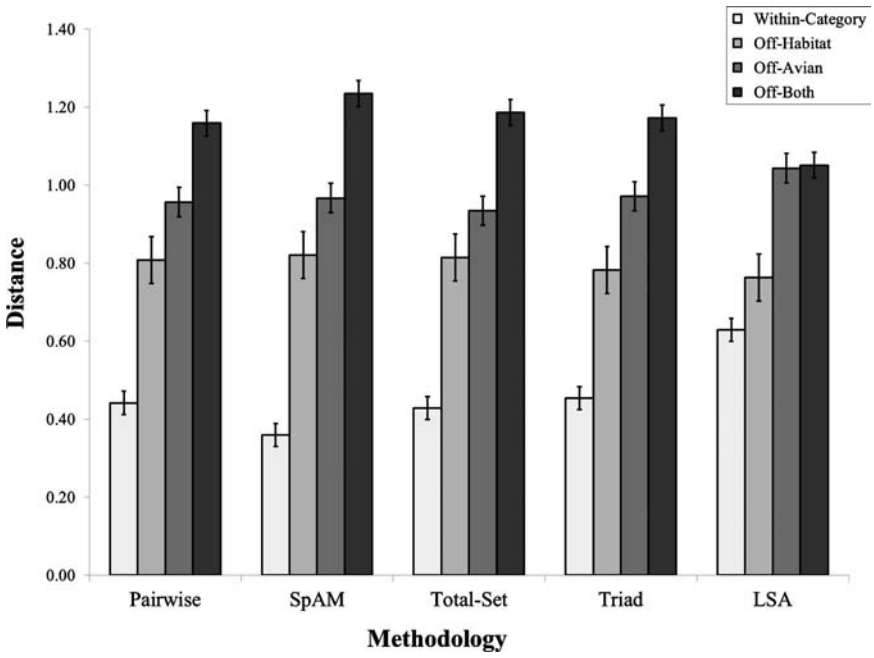


*Figure 10.* Cluster analyses showing the average item-to-item distance for stimuli that shared two features (within-category), one feature (off-habitat, off-avian), or no features (off-both), from Experiment 2 (categorical animals). Error bars represent ±1 standard error of the mean. SpAM = spatial arrangement method; LSA = latent semantic analysis.

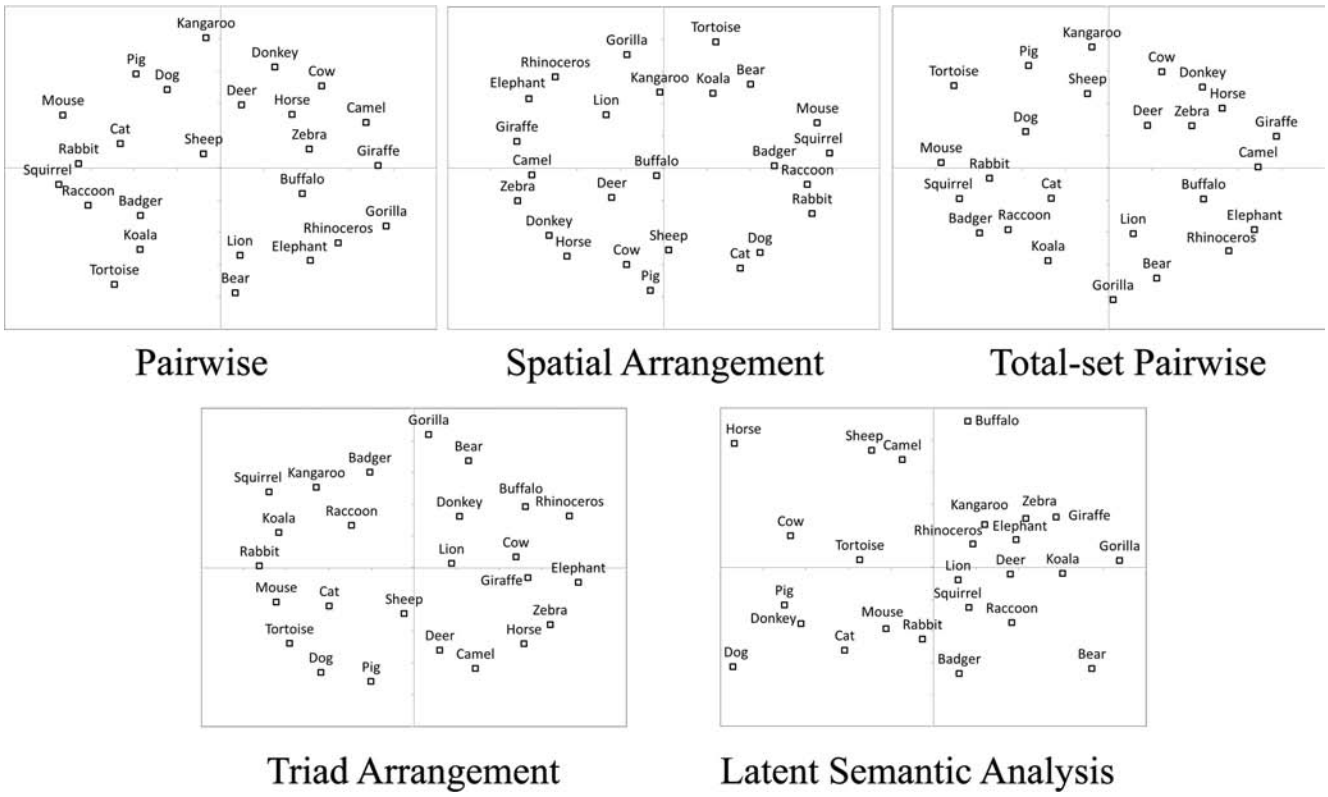


*Figure 11.* Continuous animal multidimensional scaling spaces generated by each of the four methodologies, and by latent semantic analysis, from Experiment 2.



Carlo simulations show that, across iterations, the pairwise method and SpAM demonstrate high stability. Stability of the SpAM solutions were again reduced only slightly by reducing the granularity of the data or the sample size. (d) In the same vein, SpAM consistently correlated with the pairwise method; this congruence was minimally affected by reduction of data mass and granularity. (e) The cluster analyses were replicated in the Monte Carlo simulations and showed no discernible effects from degrading the SpAM data. Lastly, (f) the individual differences analyses suggest that removing a full quarter of the least regular data from SpAM did not dramatically affect the overall solutions (i.e., the method is robust to outliers).

Taken together, the findings from Experiment 2 suggest that the utility of SpAM is not limited to stimuli with obvious perceptual similarity. It is also clear that the total-set technique offers a strategic advantage, relative to pairwise method; namely, instant appreciation of the entire stimulus set. Although we are cautious about subjective interpretation of the continuous animal spaces, the findings from these stimuli were also informative. Although the stimuli were not selected with any dimensions in mind, across methods and simulations, the solutions showed high agreement.

## Experiment 3

Each day people are faced with stimuli and situations that are nearly identical to those they have encountered previously. It is imperative that an organism be able to adequately generalize and discriminate; a successful creature is one that can detect when two situations (or stimuli) are similar enough to be acted upon as the same and when they are dissimilar enough to require different actions. Early theorists (e.g., Hull, 1943) recognized that no learning theory was complete without addressing how learning in one situation generalized to (or discriminated from) another. For example, a lifetime of experience drinking from coffee mugs affords a person the knowledge that a previously unseen ceramic cup is a more appropriate conduit for a hot beverage than a disposable, plastic cup. Shepard (1987) argued that generalization is an abstract cognitive act; we generalize not because we cannot tell the difference between situations, but because we reason that they belong to a larger set of situations (or "consequential regions") that share a common outcome. Importantly, Shepard and others (Russell, 1988; Shepard, 1957, 1958, 2004; Shepard & Chang, 1963) have shown that generalization gradients (i.e., the function relating the probability that two items will be acted upon as the same, against their distance in psychological space) follow a mathematical law, falling off exponentially as the disparity between stimuli increases (see also Henmon, 1906). This work has firmly established that points (representing objects or situations) lying closer together in psychological space will more often (and/or more quickly) lead to generalization and, conversely, that points lying farther apart in psychological space

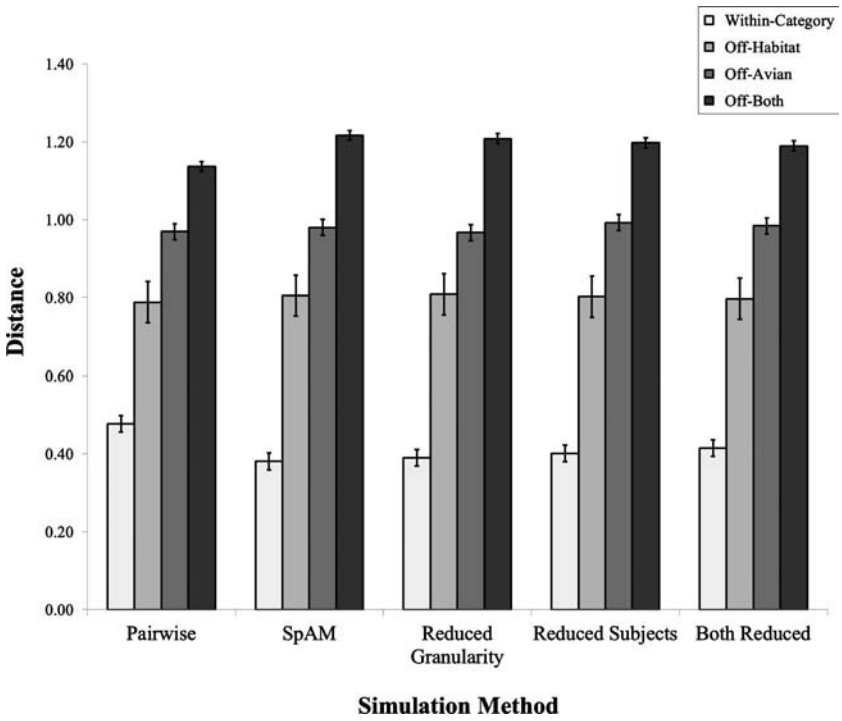


*Figure 12.* Cluster analyses showing (for each simulation type) the average item-to-item distance for stimuli that shared two features (within-category), one feature (off-habitat, off-avian), or no features (off-both), from Experiment 2, Monte Carlo simulations (categorical animals). Error bars represent ±1 standard error of the mean. SpAM = spatial arrangement method.

tend to elicit discrimination behavior (more frequently and/or more quickly).

Germane to the current investigation, MDS-derived perceptual spaces have been used to predict behavior in independent tasks, such as "same–different" classification (e.g., Gilmore, Hersh, Caramazza, & Griffin, 1979; Podgorny & Garner, 1979; Townsend, 1971). Accordingly, as a final test, we assessed the extent to which solutions derived from SpAM and pairwise methods predicted perceptual discrimination, using two variants of a same–different task. By assessing reaction times (RTs) and error rates on "different" trials, we gauged how well each method predicted discrimination across objects. Stimuli that are distant from one another in psychological space should elicit shorter RTs and fewer errors, relative to points that are closer together, as discrimination should be easier in these cases. Perceptual discriminations thus provide a robust metric to assess the quality of MDS solutions.

## Method

**Participants.** Experiment 3 included 48 new students from Arizona State University who participated for partial course credit. All participants had normal or corrected-to-normal vision.

**Design.** In the first part of the experiment, each participant provided similarity ratings for two stimulus sets, once using SpAM and once using the pairwise method; task order and method-to-stimuli pairing were counterbalanced across participants. Following these similarity ratings, participants completed two blocks of same–different classification; one block was speeded and one was nonspeeded (task order and stimuli pairing were again counterbalanced).

**Stimuli.**

***Bugs.*** We selected 16 of the two-dimensional bugs, crossing four levels of body color (light gray to black) with four levels of legs (three to six legs per side).

***Faces.*** Novel faces were generated with FaceGen Modeller software (Singular Inversions, 2004). Faces were created by generating a prototype (a racially ambiguous, male face) and then systematically distorting the prototype along two dimensions: skin shade and separation of the eyes (varying in equal steps from −1 to 2 and −3 to 3 for skin shade and eye separation, respectively; see supplement Figure A1).

**Procedure.**

***Similarity ratings.*** The pairwise and SpAM methods for collecting similarity ratings were identical to those of the previous experiments. There were 120 trials for the pairwise method and one trial of SpAM.

***Speeded classification.*** Participants made 152 judgments; each of the 120 "different" stimulus pairs was shown, with an additional 32 "same" trials, all in random order. Pairs were presented side by side, and participants quickly pressed buttons on the keyboard indicating "same" or "different." Feedback was given only for incorrect responses, and a 500-ms intertrial interval separated each pairing.

***Nonspeeded classification.*** The procedure was identical to speeded classification, except stimuli were presented sequen-

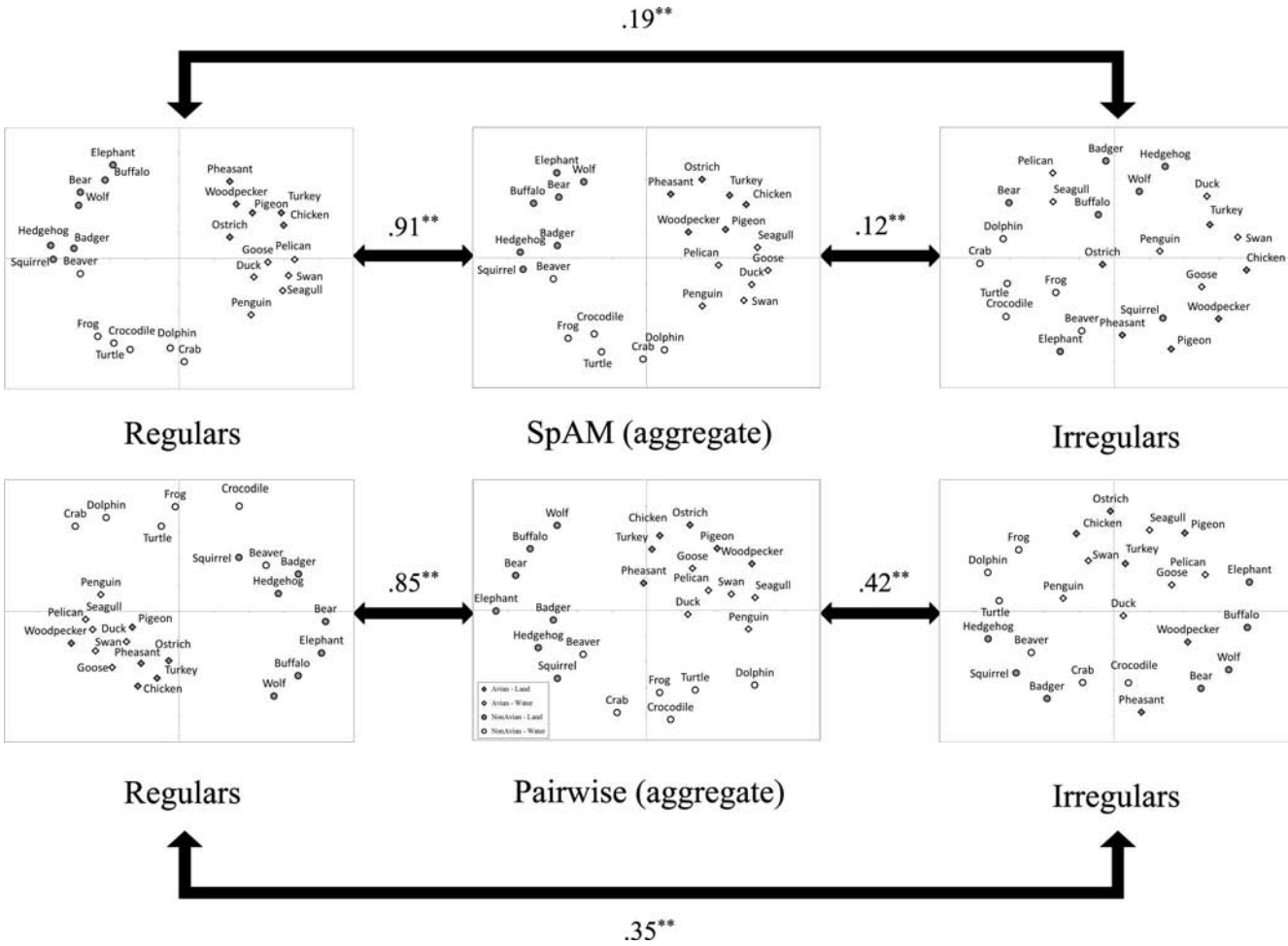


*Figure 13.* Multidimensional scaling spaces for categorical animals, derived by the spatial arrangement method (SpAM) and the pairwise method (Experiment 2). The left panels show solutions that exclude outlier participants; the right panels are solutions from only outliers. The numbers represent Pearson product–moment correlation coefficients (** $p < .01$) between the item-to-item distance vectors from each solution.

tially. Each trial began with a fixation cross (250 ms), followed by a noisy forward mask (250 ms) and then the first item of the pair (250 ms). This image was replaced by a 500-ms mask and then the second stimulus item (250 ms). The second stimulus was offset slightly to the left or right of the first (randomly), such that items could not be matched as templates of one another. Finally, a backward mask (250 ms) was presented, after which participants indicated responses using the keyboard.

## Results

All MDS solutions were derived with the same techniques used in Experiments 1 and 2; both the bug and face stimuli were scaled in two dimensions. The aggregate solutions are shown in supplement Figure A6.

**Discrimination gradients.** From each MDS solution, we acquired 120 values, representing the Euclidean distances in psychological space between all pairs of items. We then plotted these distances against the mean "different" RT and error rate for each pair (from the speeded and nonspeeded classification tasks, respectively). Exponential fit lines could not be applied to the raw error rates, because several pairs elicited no errors; therefore, we adjusted the error rates by adding .001 to each value. Next, we plotted the best fitting functions (logarithmic and exponential) relating discrimination to distance in psychological space. The results, as shown in Figure 14, were uniformly concave upward, with more efficient discrimination (i.e., faster RTs, fewer errors) as distance in psychological space increased.

Logarithmic trend lines produced uniformly better fits, relative to exponential trends. For the following results, logarithmic and exponential fit values are shown outside and inside brackets, respectively. For the bug stimuli, SpAM ($R^2_{adjusted}$ = .45 [.41] and .61 [.43] for RTs and errors, respectively) provided MDS coordinates that fit the same–different data better, relative to the pairwise method ($R^2_{adjusted}$ = .35 [.34] and .47 [.37] for RTs and errors, respectively). For the face stimuli, the pairwise method ($R^2_{adjusted}$ = .65 [.64] and .74 [.55] for RTs and errors, respectively) provided a better fit, relative to SpAM ($R^2_{adjusted}$ =

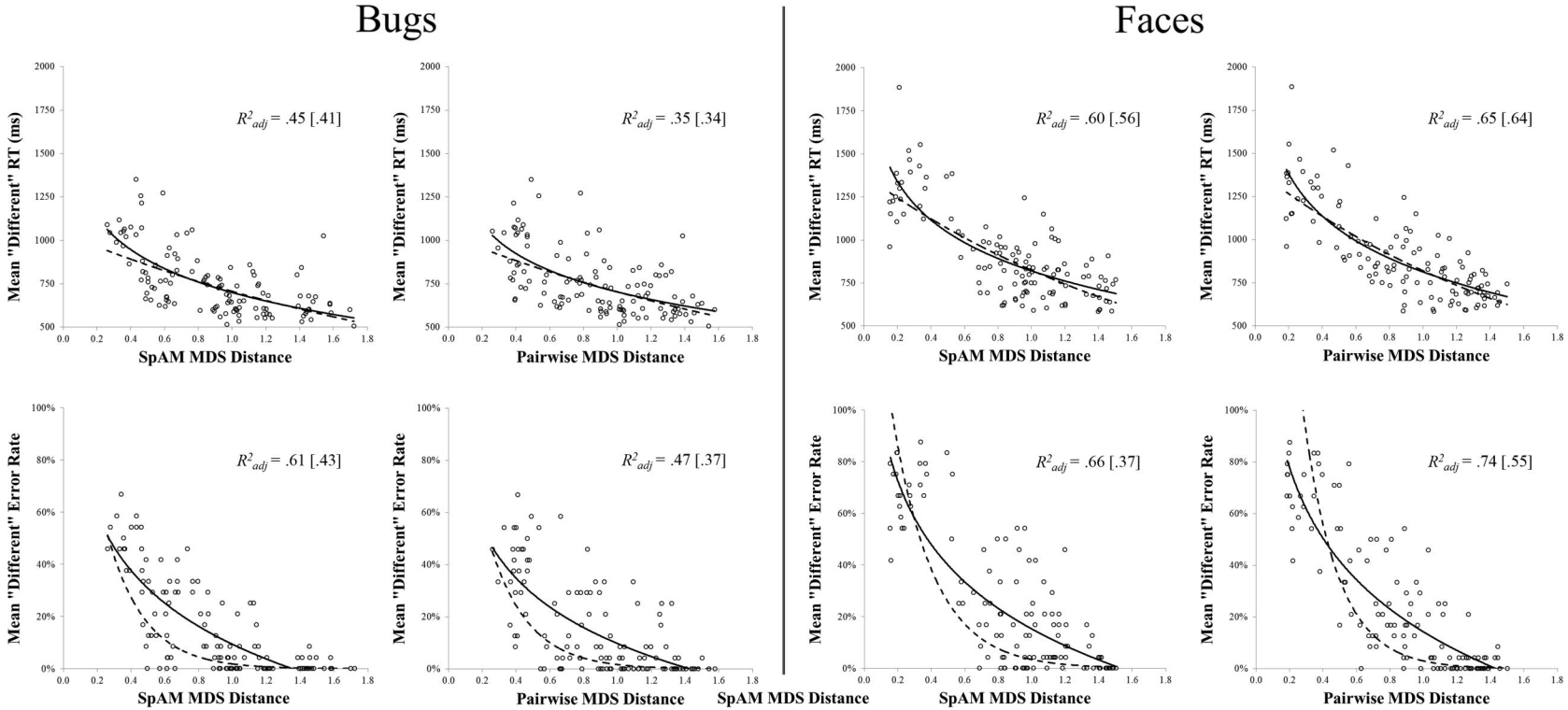


*Figure 14.* Discrimination gradients for bug and face stimuli, from Experiment 3. The *x*-axes show distance in psychological space, derived from the spatial arrangement method (SpAM; left subpanels) and the pairwise method (right subpanels). The *y*-axes show discrimination behavior: reaction times (RTs) and corrected-error rates from "different" trials (top and bottom panels, respectively). Solid trend lines represent the best fitting logarithmic function; dotted trends are exponential functions. Fit values are shown outside and inside brackets for logarithmic and exponential functions, respectively. MDS = multidimensional scaling; adj = adjusted.

.60 [.56] and .66 [.37] for RTs and errors, respectively). Each trend line produced a reliable fit (all $ps < .01$).[8]

### Discussion

Experiment 3 provided additional evidence that SpAM generates MDS solutions that are comparably organized, relative to the pairwise method. Moreover, the MDS coordinates produced by SpAM predicted stimulus generalization with approximately the same precision as those derived by the pairwise methodology. Thus, we suggest that SpAM's utility is not limited to producing solutions with reasonable or subjectively pleasing organizations. Rather, the spaces provided by SpAM are precise enough to predict psychological data from a task unrelated to scaling (cf. Shepard, 1987).

## General Discussion

In this investigation, we examined various methods used to collect similarity data for MDS. We systematically evaluated a relatively new, spatial arrangement method proposed by Goldstone (1994a), and we also evaluated two new, hybrid methods that each borrow aspects of the pairwise and SpAM techniques.

### Assessment of New Techniques: SpAM, Total-Set Pairwise, and Triad Methodologies

SpAM exhibits four methodological advantages, relative to the pairwise procedure, for collecting similarity data. First, it is fast and efficient. Each time a participant moves an object on the screen, the action simultaneously changes the relationship of the moved object to all other stimuli present. Thus, with a few movements, organization of the entire space can be modified: Our participants scaled 25–27 stimuli in roughly 5 min, compared with 25–30 min necessary for the pairwise method. This disparity, it should be noted, will grow as the stimulus set grows. SpAM allows a researcher to collect full data matrices from many participants with fewer concerns about fatigue or inconsistencies across trials (see Bijmolt & Wedel, 1995). Second, SpAM produces data with high resolution. Pairwise responses are often limited to points along a Likert scale, thereby limiting individual responses to approximately 10 units. By contrast, SpAM generates ratings that are only limited by the resolution of the computer monitor, as the ratings are Euclidean distances, measured in pixels.

Third, because all (or many) of the stimuli are presented simultaneously, participants are instantly calibrated to the full ranges of the important stimulus dimensions. This lies in stark contrast to the pairwise method, wherein the first several ratings may be arbitrary, and will likely conflict with later, better informed decisions. For example, if presented with the pairing *sparrow–goose*, one may be inclined to indicate a high degree of similarity, as both are birds. However, if the entire stimulus set was composed of numerous small birds and other large birds, this initial rating would prove relatively inaccurate, given full context. In the full set, *sparrow* should be rated similar to other small birds (e.g., *robin*) and less similar to larger birds such as a *goose*. One may assert that such a participant was not "zoomed in" enough on the initial rating, failing to appreciate relevant aspects of dissimilarity. If every possible pair is presented only one time, such uncalibrated responses can have deleterious effects on the overall coherence of the data matrix. SpAM provides a stable context in which to make similarity decisions, because the presented stimuli remain constant as each decision (i.e., movement of items in the space) is made. Fourth, SpAM is intuitive and user-friendly. Given a large stimulus set, pairwise methods can be quite tiresome. SpAM provides a more engaging technique for collecting data, and it exploits peoples' natural tendency to think of similarity in spatial terms, by giving them a spatial medium to indicate their perceptions.

Earlier, we speculated that researchers may be hesitant to use SpAM because pairwise procedures are better established and have been used across myriad domains. Interestingly, however, SpAM bears a striking resemblance to several other procedures that are currently in use. For instance, sensory analysts use a technique called "projective mapping" to collect consumer research data: People place products (e.g., samples of chocolate) on sheets of paper that are marked with coordinate axes in locations that respect the perceived similarity of each pair of items (King, Cliff, & Hall, 1998; Risvik, McEwan, Colwill, Rogers, & Lyon, 1994; Risvik, McEwan, & Rødbotten, 1997). "Napping" is a nearly identical technique (Nestrud & Lawless, 2011; Pagès, 2005; Perrin et al., 2008) wherein people place food or drinks on a sheet of paper or tablecloth. Importantly, the data acquired from these procedures have been analyzed with methods including MDS, such as INDSCAL (e.g., Qannari, Wakeling, & MacFie, 1995), three-way MDS (Abdi, Dunlop, & Williams, 2009; Abdi, Valentin, Chollet, & Chrea, 2007), and others. What brings these methods together is not a common analysis, but their shared use of space as the medium by which to acquire similarity estimates.

To assess the quality of SpAM data, relative to the well-established pairwise procedure, we sought converging evidence from several analytical techniques. With perceptual stimuli, we found that the method was adept at "discovering" the dimensions by which the stimuli were constructed. Using conceptual stimuli, we corroborated this finding, as SpAM uncovered the categorical (i.e., binary) dimensions by which our stimuli were selected. Generally, SpAM provided MDS solutions that were (a) comparably organized to those derived by pairwise procedures, (b) stable across multiple iterations of MDS, and (c) relatively robust to reductions in data mass or granularity. This final point suggests that although these aspects of SpAM contribute to its high quality solutions, they are not necessary elements. Moreover, SpAM-derived MDS solutions accurately predicted stimulus generalization in a same–different task. Giguère (2006) noted that there are no "convincing" statistical techniques for verifying the interpretation of an MDS space. But taken together, the present results suggest that SpAM provides consistent, stable, coherent, and, perhaps most importantly, useful MDS solutions.

[8] For each data set, we compared the fits (indexed by $R^2_{\text{adjusted}}$ values) for discrimination behavior, relative to MDS solutions with one, two, and three dimensions. The best fits were always provided by 2D solutions. We also plotted fit lines for distances derived from non-MDS data. That is, we used average aggregate proximities, without subjecting the data to a scaling algorithm. Relative to MDS-derived proximities, the "raw" proximities provided some improvements in fit, suggesting that raw values can also predict discrimination behavior.

We also evaluated two hybrid techniques for collecting similarity data: the total-set pairwise and triad methods. The total-set method follows the traditional pairwise procedure, but the entire stimulus set is shown at once, and participants are cued about which pair to rate in each trial. This simple modification of the pairwise procedure allows people to instantly calibrate themselves to the important dimensions of the stimuli. Across experiments, it consistently produced MDS solutions that were comparable to those of the pairwise method. Thus, the total-set method is an attractive technique for collecting similarity data, particularly when researchers are concerned that participants may not be zoomed in (or out) appropriately, given the context of the stimuli. However, this method suffers some of the same drawbacks (i.e., lengthy experimental protocols and low granularity) as the standard pairwise procedure.

The triad method can be envisioned as a series of miniature spatial arrangement procedures (SpAM lite!) wherein participants are shown three-item sets and arrange the objects at distances proportional to their similarity. This method uses the intuitive interface and high-resolution responding from SpAM. By limiting the number of stimuli presented per trial (rather than presenting large sets, as in SpAM), the triad method might encourage more thoughtful, accurate responding. Although we observed reliable congruence with the pairwise procedure, the triad method generally performed the worst among our tested methods (e.g., its correlations and deviation scores were often the lowest and highest, respectively). Note that although our approach in the present research was to display all stimuli simultaneously, Goldstone's (1994a) original method used multiple partial-set trials (20 items per trial). It therefore may not be necessary to display all items at once, but perhaps limiting each trial to three items is disadvantageous. Our triad procedure took approximately the same time to complete as the pairwise method, so the large number of trials (100–117) may have created fatigue. For now, our recommendation is to use this procedure with caution, and to consider presenting larger subsets of stimuli to diminish the length of the experimental protocol, but still encourage appreciation for the larger context of the set as a whole.

## Individual Differences

Goldstone (1994a) suggested that the SpAM is prone to individual differences in the interpretation of instructions. His participants sometimes treated distance as a continuous measure and at other times organized the stimuli into small clusters of items; our participants did this as well (see Figure 6). Judging by the robust solutions we observed, such individual differences appear as a minor concern. First, the Monte Carlo simulations suggested that SpAM produces stable results that are largely unaffected by separate iterations of the scaling algorithms; in short, individual differences do not appear to make SpAM solutions unstable. Second, our exclusionary individual differences analyses indicated that solutions derived from SpAM remain consistent even when a full quarter of the least regular data (as indexed by the extent to which individual data matrices correlate with others) is removed. Third, individual differences also arise in pairwise techniques, as some participants use the entire scale for responding, whereas others reduce their ratings to a few numbers (e.g., the lowest, middle, and highest scale values). Indeed, when we applied the same exclusion criteria to data obtained by pairwise methods, we found that the irregular pairwise solutions were substantially different from the aggregate data (as were SpAM's irregulars), indicating that SpAM is not uniquely susceptible to individual differences in task performance (see Hutchinson & Lockhead, 1977).

Sometimes individual differences in MDS solutions are informative rather than a nuisance. For example, Krumhansl and Shepard (1979) found that whereas the musically inclined tend to appreciate structural features of tone stimuli (e.g., tonal hierarchies), more naive participants attend to simpler stimulus characteristics (e.g., pitch height; see also Kessler, Hansen, & Shepard, 1984). Bimler, Kirkland, and Pichler (2004) observed "compressed" color spaces for individuals with different forms of color deficiency, relative to normal perceivers. Hollins, Bensmaïa, Karlof, and Young (2000) found that most people perceive two primary dimensions of tactile sensation (*rough*/*smooth* and *soft*/*hard*), but some appreciate a *sticky*/*slippery* dimension as well. And Schiffman, Reilly, and Clark (1979) observed wide variability in the perception of sweeteners, suggesting that sweetness perception may be mediated by many different properties (e.g., viscosity, aftertaste, bitterness).

Moreover, individual differences in the use of space may actually contribute to the quality of high-dimensional SpAM solutions. Although individual participants likely use two primary dimensions when arranging stimuli, aggregate data pooled across participants may yield satisfactory high-dimensional solutions because of the way different participants organize their spaces. To verify this notion, we created two hypothetical SpAM participants: Each appreciated only a single dimension of our two-dimensional bugs and ignored the other (e.g., bugs of varying colors were arranged along a line, with those sharing the same number of legs being stacked on top of one another). We then fed the coordinate values from these two participants into PROXSCAL and recovered the two-dimensional solution. The result was an aggregate solution that perfectly appreciated both dimensions of the stimuli. For three-dimensional stimuli (and beyond), this idea expands to a problem of representing multiple, overlapping dimensions on a two-dimensional plane. Again, we created hypothetical SpAM participants that each appreciated only a subset of the relevant stimulus dimensions for the three-dimensional bugs. We generated three participants; each appreciated two dimensions at a time and stacked items over the third dimension. When the data were scaled in three dimensions, all three stimulus characteristics were appreciated. Figure 15 shows the results: Dimension 1 organizes the items according to number of legs, Dimension 2 reflects antennae curvature, and Dimension 3 (most clearly visible in the lower right panel) reflects color.

Thus, even if participants produce solutions that only appreciate subsets of the full dimensionality, individual variations in the salience of these dimensions (i.e., which two dimensions any person deems important) can engender aggregate solutions that represent the full set in high-dimensional space. It should also be noted that more than two dimensions can be represented on a single plane. For instance, one could create three equidistant clusters of bugs grouped by color wherein each cluster is arranged according to legs and antennae. This strategy would create imperfect appreciation of some dimensions to the benefit of others, but again, individual differences in the salience of multiple dimensions will foster a high-quality group solution. As noted earlier, because

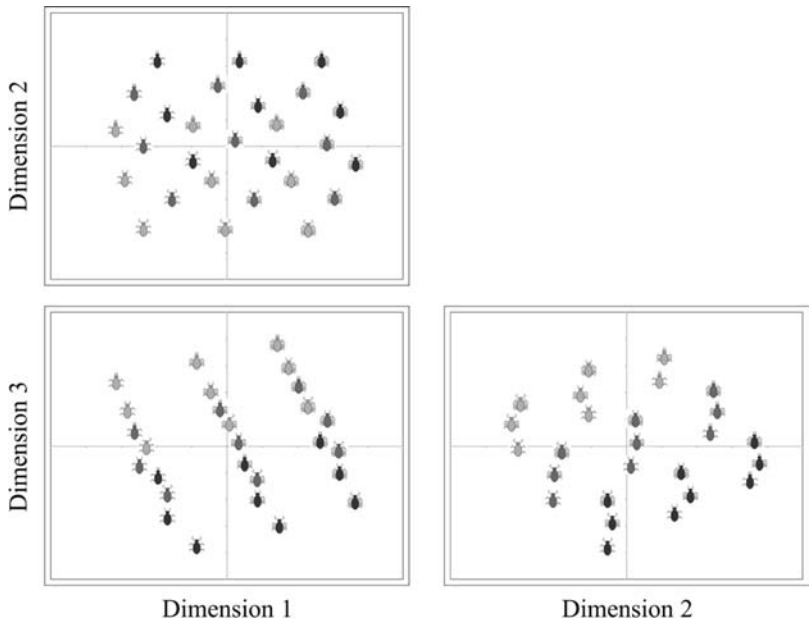


*Figure 15.* A three-dimensional bug solution derived from three hypothetical spatial arrangement method spaces, each of which appreciates only two dimensions at a time (e.g., legs and color, ignoring curvature of the antennae).



SpAM allows fast data collection, it is not particularly challenging to sample adequately.

Like other aspects of performing MDS (e.g., choosing the dimensionality of the solution, interpreting the dimensions), the approach to dealing with individual differences should ultimately be driven by the research question at hand. In some cases, individual differences are unimportant, and the analyst may choose to trust SpAM's relatively robust solutions. If the researcher deems individual differences a nuisance, outliers may be identified with the procedure suggested earlier, by testing the extent to which each participant's solution agrees with those of others. There are, of course, other criteria that could be used. For instance, if individual differences scaling (INDSCAL) is performed, the distribution of weights for individual dimensions could be used to identify unusual participants (e.g., MacKay, 1989). More generally, the rationale for concatenating data across participants is to reduce measurement error, but in some cases the averaging process may be fundamentally problematic (e.g., Estes, 1956). For instance, Ashby, Maddox, and Lee (1994; see also Lee & Pope, 2003) used simulated data matrices to show that an aggregate solution that respects the triangle inequality assumption of metric MDS (see Krantz & Tversky, 1975; Tversky & Krantz, 1970) may be comprised of individuals that violate that assumption. Lee and Webb (2005) advocate a Bayesian approach, wherein participants are partitioned into families, grouped by individual differences parameters; aggregation is applied within but not between families. This technique confers two advantages: It reduces noise by aggregating data within families, while simultaneously respecting individual differences between families. With respect to the current work, a researcher may apply this technique (broadly construed) by selecting families of individuals that correlate highly with one another and then generate separate MDS solutions for each group. (This is a less sophisticated approach than Lee and Webb described, but it is potentially useful.) Determining the optimal approach for handling individual differences is beyond the scope of this article. With respect to SpAM, however, there appear to be no inherent problems that do not also arise in the pairwise method, and its efficiency offers greater likelihood that outliers will not unduly affect the results.

## Why Use Space to Collect Similarity Estimates?

Although SpAM's speed and efficiency are appealing at the level of data acquisition, there are substantive, theoretical reasons to believe that space is an appropriate medium by which to acquire estimates of similarity. Lakoff and Johnson (1980) famously argued that metaphor plays a large role in conceptual representation and that space is a fundamental construct. Consider, for example, the so-called orientation metaphors such as *more is up* ("My income rose"), *less is down* ("Stocks fell"), *good is up* ("Things are looking up"), and *bad is down* ("I'm feeling really low"). Other examples include the *life as a journey* metaphor ("Look how far I've come") and the tendency to portray intimacy spatially ("We've drifted apart"; Lakoff, 1989). Metaphors, they reasoned, are the means by which unstructured domains of experience get organized on the basis of other highly structured domains, such as space. Shepard (2004) went a step farther, suggesting that spatial competence may underlie cognitive functions that do not, at a glance, seem spatial in nature (e.g., memory organization; Shepard, 1966). He gave a particularly compelling example regarding a game wherein two players select (without replacement) single digits from 1 to 9, with the aim of obtaining three digits that sum to 15:

> People are very slow to master this game. Yet, it is isomorphic to the trivial but spatially presented game of tic-tac-toe. This can be seen from the existence of a 3 × 3 magic square with the number 1 through 9 assigned to the nine cells in such a way that (for example) the top, middle, and bottom rows contain 8-1-6, 3-5-7, and 4-9-2, in that order. In this square, the three numbers in each row, each column, and each diagonal (and only these) sum to 15. (Shepard, 2004, p. 7)

Thus, people are capable of using space to their advantage, such that a challenging game of math and logic is reduced to the simple task of obtaining a straight, 3-point line. This heuristic is not an isolated trick, however. Mental rotation is a more cognitively demanding example: The time required to determine whether two objects (e.g., abstract shapes) are the same increases linearly as a function of the shortest rigid-axis rotation necessary to transform one shape into the other (Cooper, 1976; Shepard & Cooper, 1982). It is as if people literally rotated inner representations, as they would rotate objects in the physical world.

Furthermore, Landy and Goldstone (2007b, 2010) showed that spatial layouts can affect rule-based decision making, even when spatial relationships are irrelevant to task performance (see also Bassok, Chase, & Martin, 1998; Campbell, 1994). For instance, when mathematical expressions are widely spaced, people tend to give larger estimates, relative to more narrowly spaced problems (Landy & Goldstone, 2010). And when people write out expressions, multiplications tend to be grouped more closely than additions or equality signs, respecting the order of operations (Landy & Goldstone, 2007a). In another study, they found that the physical structure of algebraic expressions affects the reasoning of would-be problem solvers (Landy & Goldstone, 2007b). Their participants judged the validity of simple mathematical equations (e.g., "$a + b * c + d = b + a * d + c$?"); accuracy was highest when irrelevant grouping pressure (e.g., physical spacing) supported the correct order of operations.

Certainly, space is useful to ground potentially difficult constructs (as in spatial metaphors), and spatial relationships can be manipulated to help or hinder more abstract, cognitive processing. But what of the relationship between similarity and spatial proximity? Casasanto (2008) had participants give similarity ratings (using a Likert scale) to pairs of stimuli that varied as a function of how far apart they were placed on the computer screen. He found that ratings differed, depending on the distance between stimuli. For conceptual judgments (e.g., abstract nouns), stimuli presented close together were rated as more similar, relative to more distal stimuli. However, for perceptual stimuli (e.g., unfamiliar faces, object pictures), stimuli presented close together were rated as less similar. The latter finding would seem to contradict the former, but considering that one function of the perceptual system is stimulus discrimination, the finding is intuitive: It is hard to determine if a group of lines are the same length when they are far apart, but unique lengths "pop out" when the lines are placed close together. Thus, it appears that the relationship between physical and psychological proximity is not a one-way street.

We suggest that space is not just a convenient way to assess similarity relations; it is an appropriate one. Shepard (1984) made a compelling argument that internal representations are guided by the external constraints of the world. A key piece of evidence is the phenomenon of apparent motion (Carlton & Shepard, 1990a, 1990b; McBeath & Shepard, 1989)—the finding that alternately presenting two views of an object induces the experience of simple, rigid rotation of the object in three-dimensional space. Shepard contended that beyond perception, imagining, thinking, and dreaming also respect our lifetime experience with the physical world. If this is true, then it seems wholly appropriate to ask people to project their internal representations in a medium that respects both their internal and external constraints. We do not mean to suggest that SpAM spaces are veridical depictions of participants' mental representations. Our argument is only that space is appropriate to portray representations that are de facto easily conceptualized in spatial terms. A key benefit of using SpAM is that internal representations do not require conversion into an arbitrary rating system, such as a Likert scale. Rather, the computer monitor may serve as an extension of the rater's psychological space.

## Limitations

Although SpAM confers many advantages, it certainly has limitations. It is not apparent whether SpAM is equally appropriate for conceptual and perceptual similarity ratings, which answer different questions. Two things that are alike perceptually (e.g., a curtain and a blanket) may serve very different purposes, and thus be conceptually dissimilar (and vice versa; e.g., a curtain and window blinds). Goldstone (1994a) noted that SpAM may be more applicable to conceptual similarity and that confusion or discrimination measures may be more appropriate for perceptual similarity. Nevertheless, our findings suggest that SpAM is useful for collecting perceptual similarity data, especially considering that confusion measures often take as much time as pairwise procedures. Moreover, Experiment 3 suggests that SpAM solutions are, in fact, congruous with perceptual discrimination measures.

Clearly, SpAM's utility is constrained to the visual domain; pictures of objects or textual references to conceptual material. For nonvisual stimuli (e.g., olfactants, tastes), this method would seem to have limited direct utility. Nevertheless, cross-modal researchers may choose to rely on similar methods that involve physical manipulation of to-be-rated items (e.g., projective mapping and napping ). Alternatively, if a researcher wishes to rely on the convenient output from SpAM (i.e., the matrix of item-to-item distances), it would be possible to have SpAM display items on-screen that refer to stimuli in the physical world and have the rater manipulate the space accordingly.[9]

Of greater concern are arenas of similarity to which SpAM may not logically apply. Broadly, geometric models of similarity (Shepard, 1962a, 1962b) and contrast (or feature-matching) models (Gati & Tversky, 1982; Tversky & Gati, 1982) share the assumption of nonhierarchical representations; they focus on stimulus features, ignoring potential relational structures across stimuli. But as noted by Goldstone (1994c, 1996), estimating similarity is not simply a process of assessing the shared features between items. Consider the following terms: *Dog*, *puppy*, *cat*, *kitten*. Undoubtedly, dogs are more similar to puppies than they are to cats. But there is an aspect of the items that is not reflected by their isolated features, the parental relationships between a dog and puppy and between a cat and kitten. Moreover, as in analogical reasoning (see Gentner & Markman, 1997), in order for an accurate assessment to be given, aspects of one stimulus must be placed in correspondence with its comparison item.

It seems likely that given stimuli with complex or hierarchical relationships (e.g., a family tree), SpAM may not adequately capture these relational structures. Of course, one clear benefit of using SpAM is that the context of the stimuli is instantly revealed. Thus, for simple relational structures, SpAM may be more useful than pairwise methods. But when the relationships among to-be-rated items increase in complexity, the spatial medium may actually hinder the rater's ability to respect the relevant dimensions. It is tempting to thinking of similarity as a fixed, unwavering construct and to assume that collected data reflect it faithfully. But similarity is dynamic and changes with context. SpAM may not be universally applicable, but it has great utility for estimating psychological similarity.

## Software Availability

The software used in the present research is freely available from the first author's website (http://www.michaelhout.com). Resources are provided for conducting the SpAM, pairwise, and total-set MDS methods, along with Excel workbooks that include macros for data organization and concatenation. With these resources, any researcher with the appropriate software can create and analyze new MDS experiments (like its namesake, SpAM comes conveniently packaged and ready to use).

## Conclusion

The present research focused on evaluating a relatively new, spatial method for collecting similarity judgments for MDS (Goldstone, 1994a). Given the broad applicability of MDS to various areas of psychology, the availability of a robust and efficient method may have great impact. MDS has been used for test construction and validation (Napier, 1972), creation of personality profiles (Ding, 2006), organization of individual differences in counseling psychology (Dawis, 1992; Watson & Sinha, 1995), various forms of perceptual research (e.g., Bergmann Tiest & Kappers, 2006; Lawless, 1989), representation of emotions (Izmailov & Sokolov, 1991; Kroskaand & Goldstone, 1996), and thermal pain perception (Clark, Carroll, Yang, & Janal, 1986), among other examples. Similarity is, without question, a pivotal concept in the psychological sciences; our hope is that SpAM will help researchers measure similarity more easily and more accurately.

[9] It has been suggested to us that giving participants a three-dimensional layout might improve SpAM's ability to fit high-dimensional stimuli. Although this is almost certainly true, it is not feasible in the current platform (E-Prime). Moreover, a low-dimension solution is often more parsimonious than a high-dimensional one. For instance, Shepard (1980) noted that a two-dimensional representation of spectral colors is superior to Ekman's (1954) original five-dimensional solution.

## References


Abdi, H., Dunlop, J. P., & Williams, L. J. (2009). How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the bootstrap and 3-way multidimensional scaling (DISTATIS). *NeuroImage, 45,* 89–95. doi:10.1016/j.neuroimage.2008.11.008

Abdi, H., Valentin, D., Chollet, S., & Chrea, C. (2007). Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Quality and Preference, 18,* 627–640. doi:10.1016/j.foodqual.2006.09.003

Ahn, W.-K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science, 16,* 81–121. doi:10.1207/s15516709cog1601_3

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science, 5,* 144–151. doi:10.1111/j.1467-9280.1994.tb00651.x

Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology, 63,* 516–556. doi:10.2307/1418869

Bassok, M., Chase, V. M., & Martin, S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology, 35,* 99–134. doi:10.1006/cogp.1998.0675

Bergmann Tiest, W. M., & Kappers, A. M. L. (2006). Analysis of haptic perception of materials by multidimensional scaling and physical measurements of roughness and compressibility. *Acta Psychologica, 121,* 1–20. doi:10.1016/j.actpsy.2005.04.005

Bijmolt, T. H. A., & Wedel, M. (1995). The effects of alternative methods of collecting similarity data for multidimensional scaling. *International Journal of Research in Marketing, 12,* 363–371. doi:10.1016/0167-8116(95)00012-7

Bimler, D., Kirkland, J., & Pichler, S. (2004). Escher in color space: Individual-differences multidimensional scaling of color dissimilarities collected with a gestalt formation task. *Behavior Research Methods, Instruments, & Computers, 36,* 69–76. doi:10.3758/BF03195550

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications.* New York, NY: Springer-Verlag.

Busey, T. A., & Tunnicliff, J. L. (1999). Accounts of blending, distinctiveness, and typicality in the false recognition of faces. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1210–1235. doi:10.1037/0278-7393.25.5.1210

Busing, F. M. T. A., Commandeur, J. J. F., & Heiser, W. J. (1997). PROXSCAL: A multidimensional scaling program for individual differences scaling with constraints. In W. Bandilla & F. Faulbaum (Eds.), *Softstat '97: Advances in statistical software* (Vol. 6, pp. 67–74). Stuttgart, Germany: Lucius & Lucius.

Busing, F. M. T. A., Groenen, P. J. K., & Heiser, W. J. (2005). Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika, 70,* 71–98. doi:10.1007/s11336-001-0908-1

Byatt, G., & Rhodes, G. (2004). Identification of own-race and other-race faces: Implications for the representation of race in face space. *Psychonomic Bulletin & Review, 11,* 735–741. doi:10.3758/BF03196628

Campbell, J. I. D. (1994). Architectures for numerical cognition. *Cognition, 53,* 1–44. doi:10.1016/0010-0277(94)90075-2

Carlton, E., & Shepard, R. N. (1990a). Psychologically simple motions as geodesic paths I. Asymmetric objects. *Journal of Mathematical Psychology, 34,* 127–188. doi:10.1016/0022-2496(90)90001-P

Carlton, E., & Shepard, R. N. (1990b). Psychologically simple motions as geodesic paths II. Symmetric objects. *Journal of Mathematical Psychology, 34,* 189–228. doi:10.1016/0022-2496(90)90002-Q

Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an *N*-way generalization of "Eckart–Young" decomposition. *Psychometrika, 35,* 283–319. doi:10.1007/BF02310791

Casasanto, D. (2008). Similarity and proximity: When does close in space mean close in mind? *Memory & Cognition, 36,* 1047–1056. doi:10.3758/MC.36.6.1047

Chan, A. S., Butters, N., & Salmon, D. P. (1997). The deterioration of semantic networks in patients with Alzheimer's disease: A cross-

sectional study. *Neuropsychologia, 35,* 241–248. doi:10.1016/S0028-3932(96)00067-X

Clark, W. C., Carroll, J. D., Yang, J. C., & Janal, M. N. (1986). Multidimensional scaling reveals two dimensions of thermal pain. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 103–107. doi:10.1037/0096-1523.12.1.103

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooper, L. A. (1976). Demonstration of a mental analog of an external rotation. *Perception & Psychophysics, 19,* 296–302. doi:10.3758/BF03204234

Davidson, M. L. (1983). *Multidimensional scaling.* New York, NY: Wiley.

Dawis, R. V. (1992). The individual differences tradition in counseling psychology. *Journal of Counseling Psychology, 39,* 7–19. doi:10.1037/0022-0167.39.1.7

Ding, C. S. (2006). Multidimensional scaling modelling approach to latent profile analyses in psychological research. *International Journal of Psychology, 41,* 226–238. doi:10.1080/00207590500412219

Ekman, G. (1954). Dimensions of color vision. *Journal of Psychology: Interdisciplinary and Applied, 38,* 467–474. doi:10.1080/00223980.1954.9712953

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53,* 134–140. doi:10.1037/h0045156

Faye, P., Brémaud, D., Durand Daubin, M., Courcoux, P., Giboreau, A., & Nicod, H. (2004). Perceptive free sorting and verbalization tasks with naive subjects: An alternative to descriptive mappings. *Food Quality and Preference, 15,* 781–791. doi:10.1016/j.foodqual.2004.04.009

Faye, P., Brémaud, D., Teillet, E., Courcoux, P., Giboreau, A., & Nicod, H. (2006). An alternative to external preference mapping based on consumer perceptive mapping. *Food Quality and Preference, 17,* 604–614. doi:10.1016/j.foodqual.2006.05.006

Garner, W. R. (1974). *The processing of information and structure.* Potomac, MD: Erlbaum.

Gati, I., & Tversky, A. (1982). Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 325–340. doi:10.1037/0096-1523.8.2.325

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52,* 45–56. doi:10.1037/0003-066X.52.1.45

Giguère, G. (2006). Collecting and analyzing data in multidimensional scaling experiments: A guide for psychologists using SPSS. *Tutorials in Quantitative Methods for Psychology, 2,* 26–37.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67. doi:10.1037/0033-295X.91.1.1

Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics, 25,* 425–431. doi:10.3758/BF03199852

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279. doi:10.1037/0033-295X.105.2.251

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review, 11,* 716–722. doi:10.3758/BF03196625

Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: Insights from eye movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1105–1122. doi:10.1037/a0016548

Goldstone, R. (1994a). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers, 26,* 381–386. doi:10.3758/BF03204653

Goldstone, R. L. (1994b). The role of similarity in categorization: Providing a groundwork. *Cognition, 52,* 125–157. doi:10.1016/0010-0277(94)90065-5

Goldstone, R. L. (1994c). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 3–28. doi:10.1037/0278-7393.20.1.3

Goldstone, R. L. (1996). Alignment-based nonmonotonicities in similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 988–1001. doi:10.1037/0278-7393.22.4.988

Goldstone, R. L., & Medin, D. L. (1994). Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 29–50. doi:10.1037/0278-7393.20.1.29

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology, 23,* 222–262. doi:10.1016/0010-0285(91)90010-L

Goldstone, R. L., Medin, D. L., & Halberstadt, J. (1997). Similarity in context. *Memory & Cognition, 25,* 237–255. doi:10.3758/BF03201115

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General, 130,* 116–139. doi:10.1037/0096-3445.130.1.116

Green, P. E., Camone, F. J., Jr., & Smith, S. M. (1989). *Multidimensional scaling: Concepts and applications.* Needham Heights, MA: Allyn and Bacon.

Helson, H. (1964). *Adaptation-level theory: An experimental and systematic approach to behavior.* New York, NY: Harper & Row.

Helson, H., Michels, W. C., & Sturgeon, A. (1954). The use of comparative rating scales for the evaluation of psychophysical data. *American Journal of Psychology, 67,* 321–326. doi:10.2307/1418634

Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior, 8,* 176–184. doi:10.1016/S0022-5371(69)80058-7

Henmon, V. A. C. (1906). *The time of perception as a measure of differences in sensations.* New York, NY: Science Press.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428. doi:10.1037/0033-295X.93.4.411

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551. doi:10.1037/0033-295X.95.4.528

Hintzman, D. L., & Ludlam, G. (1980). Differential forgetting of prototypes and old instances: Simulation by an exemplar-based classification model. *Memory & Cognition, 8,* 378–382. doi:10.3758/BF03198278

Hollins, M., Bensmaïa, S., Karlof, K., & Young, F. (2000). Individual differences in perceptual space for tactile textures: Evidence from multidimensional scaling. *Perception & Psychophysics, 62,* 1534–1544. doi:10.3758/BF03212154

Hornberger, M., Bell, B., Graham, K. S., & Rogers, T. T. (2009). Are judgments of semantic relatedness systematically impaired in Alzheimer's disease? *Neuropsychologia, 47,* 3084–3094. doi:10.1016/j.neuropsychologia.2009.07.006

Hout, M. C., & Goldinger, S. D. (2011). Multiple-target search increases workload but enhances incidental learning: A computational modelling approach to a memory paradox [In *Object Perception, Attention, and Memory (OPAM) 2011 conference report*], *Visual Cognition, 19,* 1315–1318. doi:10.1080/13506285.2011.618773

Howard, D. V., & Howard, J. H., Jr. (1977). A multidimensional scaling analysis of the development of animal names. *Developmental Psychology, 13,* 108–113. doi:10.1037/0012-1649.13.2.108

Hull, C. L. (1943). *Principles of behavior.* New York, NY: Appleton-Century-Crofts.

Hutchinson, J. W., & Lockhead, G. R. (1977). Similarity as distance: A structural principle for semantic memory. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 660–678. doi:10.1037/0278-7393.3.6.660

Izmailov, C. A., & Sokolov, E. N. (1991). Spherical model of color and brightness discrimination. *Psychological Science, 2,* 249–259. doi:10.1111/j.1467-9280.1991.tb00143.x

Jaworska, N., & Chupetlovska-Anastasova, A. (2009). A review of multidimensional scaling (MDS) and its utility in various psychological domains. *Tutorials in Quantitative Methods for Psychology, 5,* 1–10.

Johnson, M. D., Lehmann, D. R., & Horne, D. R. (1990). The effects of fatigue on judgments of interproduct similarity. *International Journal of Research in Marketing, 7,* 35–43. doi:10.1016/0167-8116(90)90030-Q

Kessler, E. J., Hansen, C., & Shepard, R. N. (1984). Tonal schemata in the perception of music in Bali and the West. *Music Perception, 2,* 131–165.

King, M. C., Cliff, M. A., & Hall, J. W. (1998). Comparison of projective mapping and sorting data collection and multivariate methodologies for identification of similarity-of-use of snack bars. *Journal of Sensory Studies, 13,* 347–358. doi:10.1111/j.1745-459X.1998.tb00094.x

Krantz, D. H., & Tversky, A. (1975). Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology, 12,* 4–34. doi:10.1016/0022-2496(75)90047-4

Kroskaand, A., & Goldstone, R. L. (1996). Dissociations in the similarity and categorization of emotions. *Cognition and Emotion, 10,* 27–45. doi:10.1080/026999396380376

Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions with a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 579–594. doi:10.1037/0096-1523.5.4.579

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29,* 1–27. doi:10.1007/BF02289565

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29,* 115–129. doi:10.1007/BF02289694

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling.* Beverly Hills, CA: Sage.

Lakoff, G. (1989). A figure of thought. *Metaphor and Symbolic Activity, 1,* 215–225.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive Science, 4,* 195–208. doi:10.1207/s15516709cog0402_4

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284. doi:10.1080/01638539809545028

Landauer, T., & Kintsch, W. (2003). *Latent semantic analysis.* Retrieved from http://lsa.colorado.edu

Landy, D., & Goldstone, R. L. (2007a). Formal notations are diagrams: Evidence from a production task. *Memory & Cognition, 35,* 2033–2040. doi:10.3758/BF03192935

Landy, D., & Goldstone, R. L. (2007b). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 720–733. doi:10.1037/0278-7393.33.4.720

Landy, D., & Goldstone, R. L. (2010). Proximity and precedence in arithmetic. *Quarterly Journal of Experimental Psychology, 63,* 1953–1968. doi:10.1080/17470211003787619

Lawless, H. T. (1989). Exploration of fragrance categories and ambiguous odors using multidimensional scaling and cluster analysis. *Chemical Senses, 14,* 349–360. doi:10.1093/chemse/14.3.349

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology, 45,* 149–166. doi:10.1006/jmps.1999.1300

Lee, M. D., & Pope, K. J. (2003). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical Psychology, 47,* 32–46. doi:10.1016/S0022-2496(02)00019-6

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review, 12,* 605–621. doi:10.3758/BF03196751

Levin, D. T. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1364–1382. doi:10.1037/0278-7393.22.6.1364

Levine, G. M., Halberstadt, J. B., & Goldstone, R. L. (1996). Reasoning and the weighting of attributes in attitude judgments. *Journal of Personality and Social Psychology, 70,* 230–240. doi:10.1037/0022-3514.70.2.230

MacKay, D. B. (1989). Probabilistic multidimensional scaling: An anisotropic model for distance judgments. *Journal of Mathematical Psychology, 33,* 187–205. doi:10.1016/0022-2496(89)90030-8

McBeath, M. K., & Shepard, R. N. (1989). Apparent motion between shapes differing in location and orientation: A window technique for estimating path curvature. *Perception & Psychophysics, 46,* 333–337. doi:10.3758/BF03204986

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254–278. doi:10.1037/0033-295X.100.2.254

Mugavin, M. E. (2008). Multidimensional scaling: A brief overview. *Nursing Research, 57,* 64–68. doi:10.1097/01.NNR.0000280659.88760.7c

Napier, D. (1972). Nonmetric multidimensional techniques for summated ratings. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences: Vol. 1. Theory* (pp. 157–178), New York, NY: Seminar Press.

Nestrud, M. A., & Lawless, H. T. (2011). Recovery of subsampled dimensions and configurations from napping data by MFA and MDS. *Attention, Perception, & Psychophysics, 73,* 1266–1278. doi:10.3758/s13414-011-0091-0

Newton, I. (1704). *Opticks.* London, England: Smith and Walford.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57. doi:10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology, 43,* 25–53. doi:10.1146/annurev.ps.43.020192.000325

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104,* 266–300. doi:10.1037/0033-295X.104.2.266

Pagès, J. (2005). Collection and analysis of perceived product interdistances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley. *Food Quality and Preference, 16,* 642–649. doi:10.1016/j.foodqual.2005.01.006

Papesh, M. H., & Goldinger, S. D. (2010). A multidimensional scaling analysis of own- and cross-race face spaces. *Cognition, 116,* 283–288. doi:10.1016/j.cognition.2010.05.001

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72,* 407–418. doi:10.1037/h0022602

Perrin, L., Symoneaux, R., Maître, I., Asselin, C., Jourjon, F., & Pagès, J. (2008). Comparison of three sensory methods for use with the napping procedure: Case of ten wines from Loire Valley. *Food Quality and Preference, 19,* 1–11. doi:10.1016/j.foodqual.2007.06.005

Perry, L. K., Samuelson, L. K., Malloy, L. M., & Shiffer, R. N. (2010). Learn locally, think globally: Exemplar variability supports higher-order generalization and word learning. *Psychological Science, 21,* 1894–1902. doi:10.1177/0956797610389189

Podgorny, P., & Garner, W. R. (1979). Reaction time as a measure of inter- and intraobject visual similarity: Letters of the alphabet. *Perception & Psychophysics, 26,* 37–52. doi:10.3758/BF03199860

Qannari, E. M., Wakeling, I., & MacFie, H. J. H. (1995). A hierarchy of models for analysing sensory data. *Food Quality and Preference, 6,* 309–314. doi:10.1016/0950-3293(95)00033-X

Rabinowitz, G. B. (1975). An introduction to nonmetric multidimensional scaling. *American Journal of Political Science, 19,* 343–390. doi:10.2307/2110441

Richardson, M. W. (1938). Multidimensional psychophysics. *Psychological Bulletin, 35,* 659–660.

Risvik, E., McEwan, J. A., Colwill, J. S., Rogers, R., & Lyon, D. H. (1994). Projective mapping: A tool for sensory analysis and consumer research.

*Food Quality and Preference, 5,* 263–269. doi:10.1016/0950-3293(94)90051-5

Risvik, E., McEwan, J. A., & Rødbotten, M. (1997). Evaluation of sensory profiling and projective mapping data. *Food Quality and Preference, 8,* 63–71. doi:10.1016/S0950-3293(96)00016-X

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional scaling approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9,* 283–294. doi:10.1037/h0026086

Rumov, B. T. (2001). *Steiner system.* In M. Hazewinkel (Ed.), *Encyclopedia of mathematics.* Dordrecht, the Netherlands: Springer. Retrieved from http://www.encyclopediaofmath.org/index.php?title=Steiner_system&oldid=17791http://www.encyclopediaofmath.org/index.php?title=Steiner_system&oldid=17791

Russell, S. (1988). Analogy by similarity. In D. H. Helman (Ed.), *Analogical reasoning: Perspectives of artificial intelligence, cognitive science, and philosophy* (pp. 251–269). Boston, MA: Reidel.

Schiffman, S. S., Reilly, D. A., & Clark, T. B., III (1979). Qualitative differences among sweeteners. *Physiology & Behavior, 23,* 1–9. doi:10.1016/0031-9384(79)90113-6

Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications.* New York, NY: Academic Press.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide.* Pittsburgh, PA: Psychology Software Tools.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22,* 325–345. doi:10.1007/BF02288967

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology, 55,* 509–523. doi:10.1037/h0042354

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27,* 125–140. doi:10.1007/BF02289630

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika, 27,* 219–246. doi:10.1007/BF02289621

Shepard, R. N. (1963). Analysis of proximities as a technique for the study of information processing in man. *Human Factors, 5,* 33–48. doi:10.1177/001872086300500104

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1,* 54–87. doi:10.1016/0022-2496(64)90017-3

Shepard, R. N. (1966). Learning and recall as organization and search. *Journal of Verbal Learning and Verbal Behavior, 5,* 201–204. doi:10.1016/S0022-5371(66)80018-X

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210,* 390–398. doi:10.1126/science.210.4468.390

Shepard, R. N. (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review, 91,* 417–447. doi:10.1037/0033-295X.91.4.417

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323. doi:10.1126/science.3629243

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In G. R. Lockhead & J. R. Pomerantz (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 53–71). Washington, DC: American Psychological Association. doi:10.1037/10101-003

Shepard, R. N. (2004). How a cognitive psychologist came to seek universal laws. *Psychonomic Bulletin & Review, 11,* 1–23. doi:10.3758/BF03206455

Shepard, R. N., & Chang, J.-J. (1963). Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology, 65,* 94–102. doi:10.1037/h0043732

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations.* Cambridge, MA: MIT Press.

Singular Inversions. (2004). FaceGen Modeller (Version 3.1.4) [Computer software]. Retrieved from http://www.facegen.com

Spence, I., & Domoney, D. W. (1974). Single subject incomplete designs for nonmetric multidimensional scaling. *Psychometrika, 39,* 469–490. doi:10.1007/BF02291669

Spencer-Smith, J., & Goldstone, R. L. (1997). The dynamics of similarity. *Bulletin of the Japanese Cognitive Science Society, 4,* 38–56.

SPSS. (2006). *SPSS Base 15.0 user's guide.* Chicago, IL: Author.

Stevens, S. S. (1971). Issues in psychophysical measurement. *Psychological Review, 78,* 426–450. doi:10.1037/h0031324

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York, NY: Wiley.

Torgerson, W. S. (1965). Multidimensional scaling of similarity. *Psychometrika, 30,* 379–393. doi:10.1007/BF02289530

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics, 9,* 40–50. doi:10.3758/BF03213026

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352. doi:10.1037/0033-295X.84.4.327

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review, 89,* 123–154. doi:10.1037/0033-295X.89.2.123

Tversky, A., & Krantz, D. H. (1970). The dimensional representation and the metric structure of similarity data. *Journal of Mathematical Psychology, 7,* 572–596. doi:10.1016/0022-2496(70)90041-6

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 43*A, 161–204. doi:10.1080/14640749108400966

Watson, D. C., & Sinha, B. K. (1995). Dimensional structure of personality disorder inventories: A comparison of normal and clinical populations. *Personality Individual Differences, 6,* 817–826. doi:10.1016/S0191-8869(95)00130-1

Wedell, D. H. (1995). Contrast effects in paired comparisons: Evidence for both stimulus-based and response-based processes. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1158–1173. doi:10.1037/0096-1523.21.5.1158

Wish, M., & Carroll, J. D. (1974). Applications of individual differences scaling to studies of human perception and judgment. In E. C. Carterville & M. P. Friedman (Eds.), *Handbook of perception: Vol. 2. Psychophysical judgment and measurement* (pp. 449–491). New York, NY: Academic Press.

Wolfe, M. B. W., & Goldman, S. R. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods, Instruments,& Computers, 35,* 22–31. doi:10.3758/BF03195494

Young, F. W., Takane, Y., & Lewyckyj, R. (1978). ALSCAL: A nonmetric multidimensional scaling program with several individual-differences options. *Behavior Research Methods, 10,* 451–453. doi:10.3758/BF03205177

Received July 7, 2011
Revision received April 28, 2012
Accepted April 30, 2012 ■