REPLY

SpAM Is Convenient but Also Satisfying: Reply to Verheyen et al. (2016)

Michael C. Hout
New Mexico State University

Stephen D. Goldinger
Arizona State University

Hout, Goldinger, and Ferguson (2013) critically examined the spatial arrangement method (SpAM), originally proposed by Goldstone (1994), as a fast and efficient way to collect similarity data for multidimensional scaling (MDS). When using SpAM, many stimuli are presented at once (on a computer screen). The participant moves the items around, arranging them so their relative distances are proportional to the person’s subjective sense of similarity. We argued that SpAM was fast, efficient, and user-friendly, especially when compared to the historically established “pairwise” method, wherein participants are shown two items per trial and provide similarity ratings (e.g., 1–7) for each, repeating this process many times. This pairwise method suffers a key shortcoming: The number of comparisons grows rapidly as the stimulus set increases, leading to laborious experimental protocols that may elicit participant boredom, fatigue, indifference, strategy shifting, and so forth. Because all stimuli are shown together in SpAM, participants can (potentially) instantly appreciate their various dimensions of similarity. In the pairwise method, such high-dimensional appreciation likely emerges only after many trials, after the participant has already recorded numerous judgments. Our analyses suggested that SpAM produces high-quality data, often equivalent to those from the pairwise method, and therefore may be adopted as a faster, more-intuitive method for collecting similarity data.

Motivated by a reanalysis of our data, Verheyen, Voorspoels, Vanpaemel, and Storms (2016) articulated three potential caveats for SpAM. They suggested that (1) SpAM biases participants toward spatial representations, at the cost of featural representations; (2) SpAM is a crude instrument for measuring more than two dimensions at once; and (3) there are burdens on the reliability of averaged SpAM data. Although we agree on some points, we offer brief rejoinders to each of their caveats, with special focus on their second and third. We also suggest that when the pairwise and SpAM methods are equated in terms of data-collection time, SpAM is clearly superior in terms of predicting classification data. We agree that caution is required when adopting a new method but suggest that fair assessment of SpAM requires a richer data set.

**Keywords:** multidimensional scaling, similarity, methodology

Hout, Goldinger, and Ferguson (2013) critically examined a spatial arrangement method (SpAM), originally proposed by Goldstone (1994) as a fast and efficient way to collect similarity data for multidimensional scaling (MDS). When using SpAM, many stimuli are presented at once (on a computer screen). The participant moves the items around, arranging them so their relative distances are proportional to the person’s subjective sense of similarity. We argued that SpAM was fast, efficient, and user-friendly, especially when compared to the historically established “pairwise” method, wherein participants are shown two items per trial and provide similarity ratings (e.g., 1–7) for each, repeating this process many times. This pairwise method suffers a key shortcoming: The number of comparisons grows rapidly as the stimulus set increases, leading to laborious experimental protocols that may elicit participant boredom, fatigue, indifference, strategy shifting, and so forth. Because all stimuli are shown together in SpAM, participants can (potentially) instantly appreciate their various dimensions of similarity. In the pairwise method, such high-dimensional appreciation likely emerges only after many trials, after the participant has already recorded numerous judgments. Our analyses suggested that SpAM produces high-quality data, often equivalent to those from the pairwise method, and therefore may be adopted as a faster, more-intuitive method for collecting similarity data.

Motivated by a reanalysis of our data, Verheyen, Voorspoels, Vanpaemel, and Storms (2016) articulated three potential caveats for SpAM. They suggested that (1) SpAM biases participants toward spatial representations, at the cost of featural representations; (2) SpAM is a crude instrument for measuring more than two dimensions at once; and (3) there are burdens on the reliability of averaged SpAM data. Although we agree on some points, we offer brief rejoinders to each of their caveats, with special focus on their second and third. We also suggest that when the pairwise and SpAM methods are equated in terms of data-collection time, the advantages of SpAM become more clear and compelling.

**Caveat 1 (SpAM Favors Spatial Over Featural Representations)**

The first drawback identified by Verheyen et al. (2016) came from an analysis of the distributional characteristics (the skew and centrality) of our SpAM and pairwise data. The results suggested that the spatial nature of SpAM can impose a structure on the proximities that may bias eventual solutions toward spatial representations (e.g., MDS), rather than featural representations (e.g., tree structures). We largely agree with this point and believe the data speak for themselves: Our pairwise data showed a representational distinction between the perceptual and conceptual stimuli (picture and word stimuli favored spatial and featural representations, respectively), whereas our SpAM data favored spatial representations for both types of stimuli. On this point, Hout et al. (2013, p. 277) wrote that “it is not apparent whether SpAM is equally appropriate for conceptual and perceptual similarity ratings.” Despite this point, we must emphasize that the “conceptual” MDS spaces produced by the pairwise and SpAM methods showed
remarkably similar organizations (see Hout et al., 2013, Figures 9 and 11). To quantify this similarity, we correlated the interpoint distance vectors from the different MDS spaces, estimating the agreement between the SpAM and pairwise solutions (see Hout et al., 2013, Table 2). This analysis quantifies the extent to which the arrangement of points is consistent across solutions (i.e., high correlations indicate that objects close together in one MDS space are close in the other MDS space). In doing so, we found stronger agreement between solutions for the featural stimuli \((r = .81)\) than for the spatial stimuli \((r = .61)\). This indicates that the MDS organization resulting from SpAM-derived data is highly consistent with that produced by the pairwise method (regardless of the featural or spatial nature of the stimuli). Such high congruity for featural stimuli also contradicts Caveat 1 from Verheyen et al., which argues that SpAM should poorly approximate the pairwise method for objects with clear featural distinctions.

**Caveat 2 (SpAM Is a Crude Instrument for Measuring Many Dimensions)**

The second caveat regarded the number of dimensions that arise in MDS solutions from SpAM. In many cases, researchers cannot predict how many dimensions may be appreciated in their stimuli, and higher dimensions may carry theoretical interest. Verheyen et al. (2016) suggested that the two-dimensional (computer screen) nature of SpAM induces participants to represent similarity according to two principal dimensions, or to organize objects in ways that poorly represent higher dimensionality. They supported this claim through a reanalysis of our data (using individual differences scaling; Carroll & Chang, 1970; Takane, Young, & De Leeuw, 1977). Specifically, for our three-dimensional perceptual stimuli, many participants attended to only two dimensions in their organizations, or represented three dimensions in suboptimal fashion. Thus, SpAM may lead researchers to underestimate the dimensions in their stimuli.

In the General Discussion of our original article (Hout et al., 2013), we specifically addressed the ability of SpAM to reveal more than two dimensions. We created hypothetical data wherein each of three “participants” appreciated only two out of three dimensions for our “bug” stimuli (respectively, Dimensions 1–2, 1–3, and 2–3). The average MDS space nicely reflected all three dimensions. We suggested that, when participants focus on different stimulus features, rich high-dimensional solutions may emerge. Nevertheless, Verheyen et al. (2016) found that many real participants appreciated only two dimensions at a time. Given this fact, if most participants focused on the same two salient dimensions, aggregate SpAM solutions may indeed underestimate the psychological importance of higher dimensions.

Despite this pattern in our reported data, it is important to appreciate the implicit assumption behind Caveat 2. Specifically, there is a large difference between the actual experiments conducted by Hout et al. (2013) and the universe of possible experiments that researchers might conduct using SpAM. Our article was exploratory: Although we presented three-dimensional stimuli, we never instructed participants to try to capture all three dimensions in their arrangements. We merely asked people to arrange all items to reflect their subjective impressions of similarity, which seemed appropriate for exploring a novel method. By changing the task instructions, we might have easily encouraged more-nuanced arrangements with higher dimensionality. For example, participants could be shown several examples wherein objects are arranged to reflect three simultaneous dimensions and would be asked to consider whether their target stimuli might require analogous arrangements to fully communicate how the objects relate to each other. Figure 1 shows a two-dimensional arrangement of animals that communicates four dimensions of stimulus similarity, although not perfectly. The space is divided into quadrants, with mammals above and birds below (Dimension 1). The left- and right-hand quadrants contain wild and domesticated animals, respectively (Dimension 2). Within each quadrant, animals are arranged vertically by size (Dimension 3), although some compromises are evident. Finally, animals most commonly eaten appear on the right-hand side (Dimension 4). We suggest that if people were directly encouraged to appreciate more than two dimensions, their SpAM arrangements would follow suit.

Although Figure 1 displays a hypothetical example, two important points deserve mention. First, despite Caveat 1 (regarding featural representations), the animal space clearly conveys features, such as mammal/bird. Second, the arrangement is “noisy,” as occurs in any similarity estimation procedure. For example, cats are smaller than dogs, but their vertical positions are reversed. This occurs because the lion and tiger “pull” the cat upward, so the space is challenging to make “perfect.” But consider the pairwise method: For the set of animals here, participants would have to produce 435 pairwise ratings just to view all possible pairs once. To recreate the dimensions shown in Figure 1, they would have to give systematically high ratings to all intermammal pairs and all interbird pairs, such that MDS would “discover” the vertical dimension. They would also have to give systematically higher ratings within the wild and domestic categories, respectively, so MDS could discover the second dimension. Participants’ ratings would also need to reflect size variations and the “food status” of different animals, all without degrading the prior dimensions.

This contrast between the SpAM and pairwise methods is central to our original article (Hout et al., 2013) and the present reply. In a pairwise procedure, assigning values to pairs of animals can be exceedingly challenging and noisy, especially given its serial nature. Imagine yourself as a participant, using a 10-point rating scale. The first pair you see is horse–zebra. After a moment’s thought, you give the pair a 9, given their clear similarity. Next, you see pig–parrot, which is likely a 1. Then you see monkey–rat, wolf–cardinal, rabbit–tiger, owl–ostrich, pigeon–goose, ostrich–bear, and chicken–tiger . . . . Even as a thought exercise, this procedure quickly descends into chaos: It is impossible to rationally and consistently choose values. The MDS algorithm may still recover these underlying dimensions, but it would require remarkable levels of nuance and consistency in the participant’s trial-by-trial ratings. As a result, high-dimensional pairwise solutions may potentially reflect noise, rather than participants’ truly appreciating subtle dimensions. Indeed, if the pairwise-derived MDS solution were well organized and meaningful, it would be remarkable. In all likelihood, data would be required from numerous participants, seeing pairs in many different orders, in order to stabilize the underlying ratings matrix.

Conversely, consider the SpAM procedure with these same items, which is exactly how Figure 1 was created. The second author (Stephen D. Goldinger) took approximately 8 min to arrange these 30 animals. He could see them all together and directly
communicate his impressions regarding mammals versus birds, domestication, size, and so forth. The recovery of multiple dimensions (and features) did not require luck: They were consciously chosen, albeit with compromises that reflect higher order relations. By designating the pairwise method as the “gold standard” for estimating similarity, Verheyen et al. (2016) essentially argued that a noisy process of generating disconnected ratings is more psychologically valid than allowing people to perfectly control their expressions of perceived similarity. Stated in these terms, the pairwise approach is challenging to accept as superior, despite its historical precedent.

Caveat 3 (There Are Burdens on the Reliability of Average SpAM Data)

The final caveat raised by Verheyen et al. (2016) was that the SpAM data from Hout et al. (2013) showed greater dispersion than did the pairwise data. Upon examining our data, they found that individual SpAM participants tended to correlate less strongly with their aggregate (group average) MDS solution, relative to participants from the pairwise method. This suggests that more noise existed in the SpAM data. In this regard, a clarification is necessary. Verheyen et al. (p. 379) noted that

SpAM’s proclaimed efficiency rests primarily on the fact that fixing a single stimulus’ position effectively establishes its proximity to all other stimuli. This should render the positioning of each stimulus a taxing problem, which seems at odds with the speed with which participants solve the task.

Although this description is reasonable, it slightly mischaracterizes our argument regarding the efficiency of SpAM. As noted, the placement of an object establishes its relationship with all others, such that moving any object changes many similarity estimates simultaneously. The method, however, does not require people to painstakingly track such relationships. Rather, because SpAM has an intuitive visual–spatial interface, people can easily appreciate the relationships between items without laboring over the position of every object. SpAM gains leverage because people naturally use space to express similarity relationships.

Verheyen et al. (2016) suggested that, to achieve consistency comparable to that in the pairwise method, roughly twice as many SpAM participants are needed. We have to ask: Is consistency among participants truly the proper benchmark for high-quality data? Similarity is a subjective concept, and assessing the “accuracy” of MDS data is tricky. We suggest high-quality MDS data should help explain other psychological data (Godwin, Hout, & Menneer, 2014; Hout & Goldinger, 2015). For example, in Experiment 3 of Hout et al. (2013), we collected similarity ratings for two sets of stimuli (bugs and faces), using the SpAM and pairwise methods. We then had participants perform speeded same–different classification to images shown side-by-side. We also examined nonspeeded classification trials wherein image pairs were shown very briefly, separated by a mask. Correct
reaction times (RTs) and error rates were examined in these tasks, respectively, and we used MDS distances to predict these classification data. Our results (see Hout et al., 2013, Figure 14) suggested that the methods were comparable. In light of Caveat 3, however, we suggest that our comparison was unfair: It compared the predictive value of the SpAM and pairwise methods without considering their respective “costs” in terms of data-collection time.

To provide a more-balanced comparison, we revisited our Experiment 3 data, sampling participants in a manner that equated the tasks for efficiency. Specifically, we sampled subsets of pairwise data, comparing them to SpAM data collected in equivalent time frames. We biased these analyses against the SpAM method, estimating that SpAM takes 4 min to scale 16 stimuli and the pairwise method requires 10 min (for 120 comparisons). With this in mind, we randomly sampled 10 pairwise participants, performed MDS, and used the results to predict discrimination RTs and errors. We repeated this five times with different random samples. The results (explained variance in generalization gradients for full SpAM and pairwise data, and the pairwise subsamplings) are shown in Figure 2. When data collection time is equated across methods, SpAM almost universally outperforms the pairwise method. For the bug stimuli, moving from the full pairwise data to the average subsample, we found a 16% reduction in explained variance, which is 25% less than that for the SpAM data. For the faces, the average subsample explained 20% less variance than did the full pairwise data and 10% less than did the SpAM data. Of course, this is only one experiment wherein the pairwise and SpAM approaches were directly compared, but the message is clear: When data collection time is equated, SpAM predicts behavior better than does the pairwise method.

SpAM Is Convenient and Satisfying

The core message from Verheyen et al. (2016) is entirely sensible: SpAM has limitations and may not be universally preferred for collecting similarity data. They suggested that researchers weigh the pros and cons of SpAM (e.g., efficiency, tendency for low-dimensional solutions) when considering whether to adopt it.

---

1 We are grateful to Rob Goldstone for suggesting this analysis.
Although this is sound advice, their core assumption (that the pairwise method is de facto superior) is not clearly supported by evidence. The pairwise method has a rich empirical history and is clearly an excellent approach. It also has obvious drawbacks, and there are doubtless many examples of pairwise experiments that did not “work” as researchers may have hoped. Given that few MDS solutions can clearly be deemed as “correct,” we suggest that both methods come with their own caveats.

References


