Caveats for the Spatial Arrangement Method: Supplementary Materials

Steven Verheyen, Wouter Voorspoels, Wolf Vanpaemel, & Gert Storms
University of Leuven, Leuven, Belgium

Data

Our comment is based on reanalyses of the dissimilarity data from Experiments 1 and 2 in Hout, Goldinger, and Ferguson (2013). In Experiment 1 four perceptual stimulus sets were used (Wheels 2D, Wheels 3D, Bugs 2D, Bugs 3D). Each set comprised of 25 (2D) or 27 (3D) artificially constructed visual stimuli varying along two or three perceptual dimensions. In Experiment 2 two conceptual stimulus sets were used (Categorical Animals, Continuous Animals). Each set comprised of 25 animal names varying along two salient (Categorical) or less-defined (Continuous) dimensions. The combination of six stimulus sets with two methods for obtaining data (SpAM vs. Pairwise) yields 12 sets of proximity data.

The proximity data resulting from SpAM and from the pairwise method differ in two notable respects. (i) More participants generated proximities using SpAM than using the pairwise method because the former takes less time and Hout et al. (2013) had participants provide several SpAM data sets in a single session. (ii) The pairwise proximities are coarser than the SpAM proximities because the resolution of the former is limited to the number of scale points of the Likert scale, while the latter are expressed as the number of pixels that separate two stimuli on a computer screen. In some analyses we equated the data sets in terms of the number of participants or in terms of their granularity (by rounding off the number of pixels). We will refer to these equated data sets as SpAM Reduced Subjects (SpAM RS) and SpAM Reduced Granularity (SpAM RG), respectively, akin to the similarly named simulations in Hout et al. (2013).

The caveats for SpAM were identified using the very data Hout et al. (2013) used to make a case for SpAM. Although this exempts us from accusations of selection bias against SpAM, we admit the employed materials/data are not ideally suited for a systematic comparison of proximity data collection methods, either. In such an investigation it would have to be ensured that stimulus nature, number of stimuli, dimensionality, and number of participants are not confounded. We urge readers to take these considerations into account when reviewing the analyses presented in Hout et al. (2013) and in these *Supplemental Materials*.

Caveat 1: SpAM favors spatial over featural representations

Background

The distributional characteristics of proximity data reveal the type of representation that is appropriate of a particular set of stimuli. The most widely used characteristics are skewness and elongation (Sattath & Tversky, 1977) and centrality and reciprocity (Tversky & Hutchinson, 1986). We restrict our discussion to skewness and centrality, because unlike elongation and reciprocity, the results of these diagnostics are not affected by differences in the granularity of proximity data, a characteristic on which SpAM and pairwise data differ. Positively skewed dissimilarity data accord well with spatial representations, while negatively skewed dissimilarity data accord better with featural representations (Sattath & Tversky, 1977). Centrality values higher than 2 are taken to indicate that the data are better represented by featural models than by spatial ones (Tversky & Hutchinson, 1986).

Evidence

First, we computed the skewness of each participant's dissimilarity data and averaged the resulting values across all participants within each combination of method and stimulus set. The average skewness values are listed in Table 1 (average skewness) along with the results of independent samples t-tests for method differences. Depending on the outcome of Levene's test for equality of variances, the variances were assumed to be equal or not. The SpAM dissimilarity data tend to be more positively skewed than the pairwise data. This finding is in line with known distributional characteristics of distances obtained from spatial representations such as the one used in SpAM (Sattath & Tversky, 1977). For both the conceptual stimulus sets (Categorical and Continuous Animals) the average skewness values differ significantly between methods. While the dissimilarity data obtained with the pairwise method tend to be negatively skewed, the SpAM dissimilarities have a positive skew.

Table 1: Skewness of the dissimilarity data.

	average skewness						skewness average	
set	Pairwise	SpAM	variances	t	df	p	Pairwise	SpAM
Wheels 2D	.09	.16	assumed equal	$-.74$	98	.23	$-.12$	$-.74$
Wheels 3D	.12	.26	no assumption	-1.33	16.07	.10	$-.05$	$-.51$
Bugs 2D	.26	.28	no assumption	$-.11$	9.50	.46	.03	.00
Bugs 3D	$-.17$	.27	assumed equal	-4.94	109	$< .001$	$-.35$	$-.34$
Categorical Animals	$-.99$	.13	no assumption	-5.76	12.48	$< .001$	$-.91$	$-.83$
Continuous Animals	$-.71$	.21	no assumption	-6.22	17.65	$< .001$	$-.80$	-1.32

Note: All tests, hypothesis is Pairwise < SpAM. Average skewness is average of individual dissimilarities' skewness. Skewness average is average dissimilarities' skewness.

Second, we computed the centrality of each participant's dissimilarity data using formula (1) from Tversky and Hutchinson (1986) where $S = \{0, 1, \dots, n\}$ is the set of stimuli and N_i reflects the focality of i with $N_i = 0$ if there is no element in S whose nearest neighbor is i and $N_i = n$ if i is the nearest neighbor of all other stimuli. Because of the occurrence of multiple ties in the pairwise dissimilarity data and its potential influence on the results, the computation was repeated 100 times, each time breaking ties at random. Averages across participants and replications are presented in Table 2 (average centrality). Centrality is significantly higher for pairwise than for SpAM dissimilarities, with the highest discrepancy again arising for the conceptual stimulus sets (Categorical and Continuous Animals). This finding is in line with known distributional characteristics of distances obtained from spatial representations as well. Low-dimensional spatial representations, such as the ones used in SpAM, tend to have centrality values under 2 (Tversky & Hutchinson, 1986).

$$C = \frac{1}{n+1} \sum_{i=0}^{n} N_i^2 \tag{1}$$

Conceptual and perceptual stimuli have been taken to be best represented by featural and spatial representations, respectively, based on the negative skew of the former, but not the latter (Dry & Storms, 2009; Pruzansky, Tversky, & Carroll, 1982) and on centrality values higher than 2 for the former, but not the latter (Tversky & Hutchinson, 1986). While the distributional characteristics of the pairwise proximity data from Hout et al. (2013) support this distinction, the SpAM proximity data do not. SpAM consistently yields data with properties that are characteristic of low-dimensional spatial representations. The pairwise method proofs able to yield data that do not necessarily demonstrate these characteristics. This indicates that the data patterns that can be obtained with SpAM are restricted.

Table 2: Centrality of the dissimilarity data.

	average centrality						centrality average	
set	Pairwise	SpAM	variances	t	df	p	Pairwise	SpAM
Wheels 2D	1.82	1.61	no assumption	7.04	45.91	$<$ .001	1.87	1.48
Wheels 3D	1.82	1.63	assumed equal	4.18	100	$<$ .001	1.48	1.52
Bugs 2D	1.88	1.52	assumed equal	6.20	89	$<$ .001	1.64	1.32
Bugs 3D	1.80	1.60	assumed equal	3.86	109	$<$ .001	1.62	1.67
Categorical Animals	2.06	1.50	assumed equal	8.87	105	$<$ .001	2.36	1.40
Continuous Animals	2.01	1.52	assumed equal	9.23	102	$<$ .001	2.12	1.48

Note: All tests, hypothesis is Pairwise > SpAM. Average centrality is average of individual dissimilarities' centrality. Centrality average is average dissimilarities' centrality.

Tables 1 and 2 also list the skewness (skewness average) and centrality (centrality average) of the averaged dissimilarity data. The averaging does not impact the centrality results much, but it does change the skewness results considerably. The averaged data are less positively skewed for both the pairwise method and SpAM, but the difference with the individual data is most pronounced for SpAM. This points toward a discrepancy between the average and the individual SpAM data. It appears that the average across individuals in the case of SpAM might not necessarily be representative of all of the individual data. We address this finding in further detail in Caveats 2 and 3.

Caveat 2: SpAM is a crude instrument for measuring many dimensions

Background

Individual differences scaling (INDSCAL) structurally incorporates individual differences (Carroll & Chang, 1970; Takane, Young, & De Leeuw, 1977). It assumes a stimulus configuration that is common to all individuals (the group stimulus space), but allows that the dimensions are differently weighted by the individuals to accommodate different proximity judgments. In that sense, the group stimulus space is a mere theoretical construct (it does not match any data), used to arrive at representations of individuals by multiplying the coordinates of the stimuli in the group stimulus space with (positive) individual weights. This has the effect of shrinking or stretching the shared stimulus space to yield individual configurations.

The INDSCAL model is most commonly used to study individual or group differences (for an overview of the range of applications, see Takane, 2007). Here it can also be used to see how the proximity data collection methods affect the proximity data that are generated. In the case of the Wheels 3D and Bugs 3D sets, where there are three salient dimensions, the INDSCAL model can be employed to establish whether SpAM participants will convey all three dimensions in their organization of stimuli on the two-dimensional computer screen or will restrict themselves to only conveying two. In the former case, we would expect the individual weights to be estimated around 1 for all three dimensions. The individuals' representations will then correspond to the three-dimensional group stimulus space. In the latter case, we would expect the individual weights to be around 1 for two of the dimensions, but much smaller than 1 for the remaining dimension. If we assume that all three dimensions are apparent to the participants, they would then effectively be dropping a dimension in conveying their proximity data.

Evidence

For both the Wheels 3D set and the Bugs 3D set from Hout et al. (2013), we conducted a simultaneous individual differences scaling of the proximity data obtained with the pairwise method and with SpAM. To this end we reduced the granularity of the SpAM proximity data (akin to the SpAM Reduced Granularity simulations in Hout et al., 2013). Differences between the two methods can easily be conveyed in this manner. We used INDSCAL with the non-metric and stress 1 options to obtain a group stimulus and a weight space in three dimensions. The group stimulus space is a configuration of points, one for each of the 27 stimuli in a set, which represents the Euclidean distances between the stimuli. The weight space contains a point for each individual who contributed proximity data, indicating the weights s/he attributed to the dimensions of the group stimulus space. There are 102 such weight vectors for the Wheels 3D set (15 pairwise participants + 87 SpAM participants). There are 111 weight vectors for the Bugs 3D set (13 pairwise participants + 98 SpAM participants).

For both sets the three-dimensional group space reflects the three-dimensional nature of the stimuli. For Wheels 3D Dimension 1 corresponds to the thickness of the stimuli, Dimension 2 corresponds to their hue, and Dimension 3 corresponds to the angle of the spoke in the wheels. For Bugs 3D Dimension 1 corresponds to the number of legs, Dimension 2 corresponds to the shading of the back and head, and Dimension 3 corresponds to the curvature of the antennae. The corresponding dimension weights are depicted in Figure 1 for Wheels 3D and in Figure 2 for Bugs 3D. The dimension weights for pairwise participants are depicted in red and are found to be centered around coordinate (1,1,1) indicating that the three-dimensional group representation is a good reflection of their individual representations. That is, the pairwise participants appeared to employ all three stimulus dimensions when providing judgments of inter-stimulus dissimilarity. There is much more variability among the dimension weights for the SpAM participants. Notably, there is a considerable number of SpAM participants found in the corners of the weight space indicating that two of the three stimulus dimensions were emphasized during the arrangement of stimuli in terms of dissimilarity. The two-dimensional nature of SpAM appears to bring a large number of participants to focus on only two dimensions of variation when judging the stimuli (assuming that all three stimulus dimensions are apparent to the participants, which is tenable given the manner in which the stimuli were constructed and the observation that the pairwise participants recognize all three dimensions). These observations are confirmed by the results of Levene's tests for equality of variances of the SpAM and pairwise dimension weights. For Wheels 3D the null hypothesis that there is no difference in the variances of the dimension weights for the two methods was rejected for both Dimension 1 ($F(1, 100) = 4.37, p = .04$, ratio= 5.26), Dimension 2 ($F(1, 100) = 5.77, p = .02$, ratio= 3.43) and Dimension 3 ($F(1, 100) = 5.03, p = .03$, ratio= 5.33). For Bugs 3D, the null hypothesis of equal variance was rejected for Dimension 1 ($F(1, 109) = 4.71, p = .03$, ratio= 4.82) and Dimension 3 ($F(1, 109) = 8.29, p = .005$, ratio= 9.77), but not for Dimension 2 ($F(1, 109) = 3.57, p = .06$, ratio= 4.60).

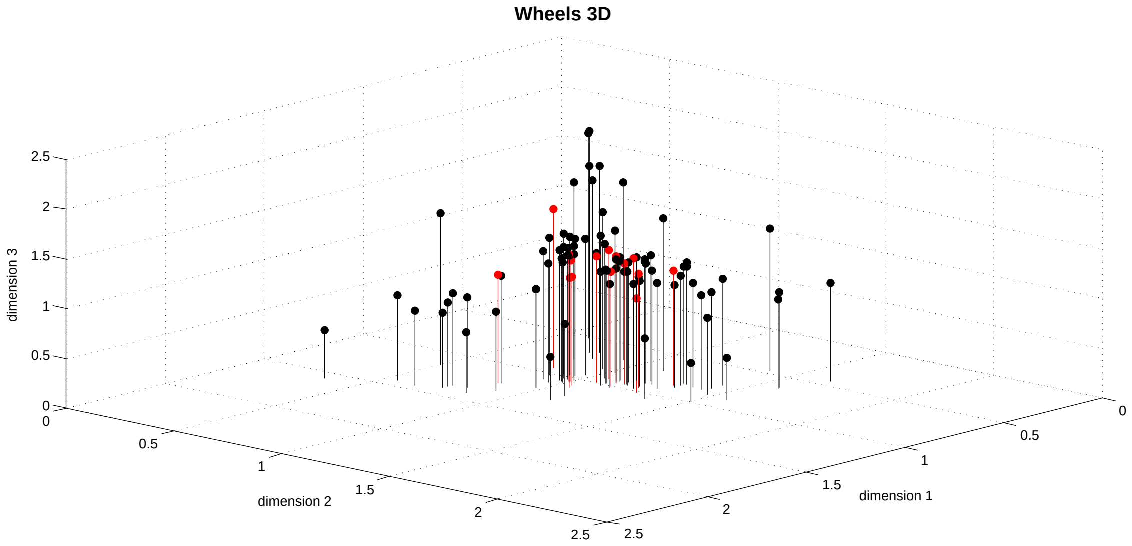

Figure 1. Dimension weights of the SpAM Reduced Granularity individuals (black) and the pairwise individuals (red) for Wheels 3D.

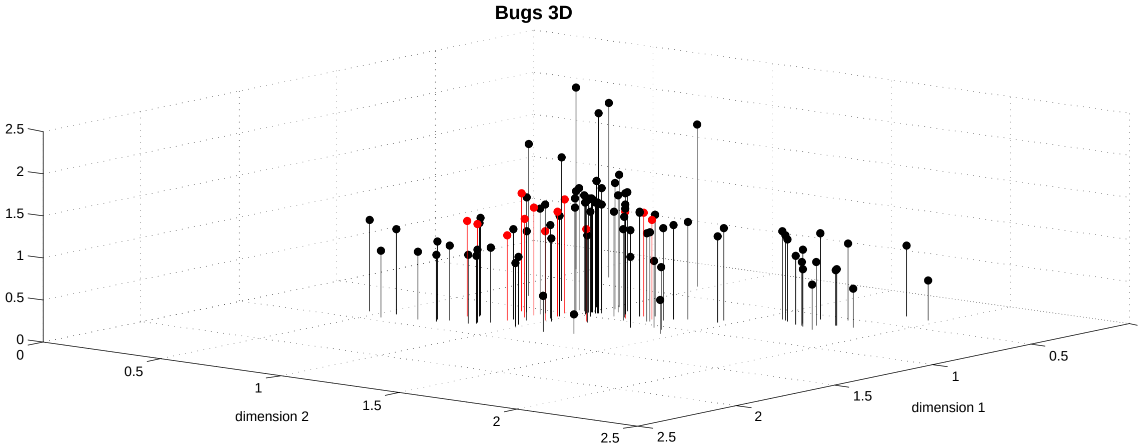

Figure 2. Dimension weights of the SpAM Reduced Granularity individuals (black) and the pairwise individuals (red) for Bugs 3D.

In addition, those SpAM participants with dimension weights around (1,1,1) tend to have a higher stress per subject (an indication of an participant's badness of fit) than those located more in the corners of the space. The Spearman correlation between a SpAM individual's stress per subject and the distance of the individual's weights from (1,1,1) is -.69 for Wheels 3D and -.70 for Bugs 3D (both $p < .0001$). These correlations indicate that participants struggled to convey the three stimulus dimensions in a two-dimensional arrangement. If we restrict our investigation of stress per subject to those SpAM individuals whose weights are not further removed from (1,1,1) than the furthest pairwise individual is, the stress per subject of the SpAM individuals is significantly greater than the stress per subject of the pairwise individuals (Wheels 3D: $t(83) = 1.77, p = .04$, one-tailed; Bugs 3D: $t(65) = 2.78, p < .01$, one-tailed). SpAM participants who wanted to communicate all three dimensions thus could not fully accomplish this in the two-dimensional arrangement they had available. Note that the pairwise individuals are in the minority in both INDSCAL analyses. Therefore there is no reason to suspect that the representations are biased toward them.

Caveat 3: Burdens on the reliability of average SpAM data

Background

While in some MDS applications individual differences constitute the topic of interest (as dealt with in Caveat 2), other applications favor an analysis of the average data, aiming to uncover the structure that is shared among participants. The differences between individuals are then considered random errors and the purpose of the averaging is to reduce this error. The reliability of the average reflects how well the differences between individuals have been cancelled out. It has as its basis a correlation between the averages across two halves of the participants. If the individual differences have been averaged out sufficiently, the two halves should correlate strongly, indicating that the average is reproducible with a sample of equal size and thus is a good reflection of the structure participants share. A low split-half correlation indicates that the averages are considerably influenced by deviating individuals and calls for the recruitment of additional participants to cancel out these influences and make for a more representative estimate of the group average (Lord & Novick, 1968). We show that there are more individual differences among SpAM than among pairwise participants and that the negative impact this has on the reliability of the average proximity data needs to be overcome by increasing the sample size.

Evidence

First, we calculated the correlation of individual MDS distances with the group MDS distances. The latter was obtained by averaging the individual dissimilarity data prior to scaling. For both the individual and group scalings we used non-metric MDS with stress 1 as an objective function to obtain Euclidean distances in two-dimensional spaces, except for the Wheels 3D and Bugs 3D stimulus sets where we employed three-dimensional spaces. The correlations were then z-transformed (Fisher's z transformation) and subjected to independent samples t-tests of the hypothesis that the correlations resulting from the pairwise method are greater than the correlations resulting from SpAM. The results listed in Table 3 show that the individual solutions tend to differ more from the group solution for SpAM than for the pairwise method for all stimulus sets, except Wheels 2D and Bugs 2D. For Wheels 2D ($t(47.48) = -2.34, p = .01$, one-tailed) and Bugs 2D ($t(25.91) = -1.89, p = .035$, one-tailed) the reverse holds.

Table 3: Mean correlations of individual MDS distances with group MDS distances

set	Pairwise	SpAM	variances	t	df	p
Wheels 2D	.44	.60	no assumption	−2.34	47.48	.99
Wheels 3D	.61	.41	assumed equal	4.33	100	< .001
Bugs 2D	.72	.83	no assumption	−1.89	25.91	.97
Bugs 3D	.62	.42	no assumption	2.50	13.29	.01
Categorical Animals	.50	.03	assumed equal	1.97	105	.03
Continuous Animals	.29	−.02	no assumption	3.14	18.61	.003

Note: All tests, hypothesis is Pairwise > SpAM.

The extent to which the average across individuals is reliable, is quantified by the split-half correlation r (the correlation between the average across one half of the participants and the average across the other half of the participants) corrected with the Spearman-Brown formula (Lord & Novick, 1968), averaged across 10,000 random splits of the dissimilarity data:

$$\rho = \frac{2r}{1+r} \tag{2}$$

The resulting reliability estimates are always higher for SpAM than for the pairwise method (Table 4), but more participants generated proximities using SpAM than using the pairwise method. The direction of the reliability difference changes when we equate the number of participants for both data collection methods. From the SpAM dissimilarities we selected a random subset of the data corresponding to the number of participants who provided data using the pairwise method and calculated the reliability. This procedure was repeated 100 times. The average reliability across these 100 replications is reported under SpAM RS (short for Reduced Subjects) in Table 4 for each of the six stimulus sets. Except for Wheels 2D ($t(99) = 18.12, p = 1$), it is lower than when the same number of participants provide proximities using the pairwise method according to a one-sample t-test ($t = -25.64, t = -6.52, t = -28.62, t = -17.70, t = -33.89$, respectively) with $df = 99$ and $p < .001$.

Table 4: Reliability of the average dissimilarity data.

		# participants		reliability of the average			
set	# stimuli	Pairwise	SpAM	Pairwise	SpAM	SpAM RS	k
Wheels 2D	25	13	87	.76	.97	.83	.65
Wheels 3D	27	15	87	.90	.95	.77	2.69
Bugs 2D	25	10	81	.85	.97	.80	1.42
Bugs 3D	27	13	98	.91	.95	.73	3.74
Categorical Animals	25	13	94	.91	.97	.83	2.07
Continuous Animals	25	17	87	.86	.91	.67	3.03

Note: SpAM RS = SpAM Reduced Subjects.

Lord and Novick (1968) provide a formula for computing from an observed reliability the required number of participants to arrive at a desired reliability:

$$k = \frac{\rho_D(1 - \rho_O)}{\rho_O * (1 - \rho_D)} \tag{3}$$

We take as the desired reliability ρ_D the estimated reliability of the pairwise method and for the observed reliability ρ_O the estimated reliability of SpAM RS. The last column of Table 4 indicates the resulting value for k, the factor with which the number of participants needs to be multiplied. Except for Wheels 2D more participants are required. Across stimulus sets the average value for k measures 2.27, indicating that about two times the number of SpAM participants are required to arrive at the reliability level of the pairwise method.

References

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.

Dry, M. J., & Storms, G. (2009). Similar but not the same: A comparison of the utility of directly rated and feature-based similarity measures for generating spatial models of conceptual data. *Behavior Research Methods*, *41*, 889-900.

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, *142*, 256-281.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publising Company.

Pruzansky, S., Tversky, A., & Carroll, J. D. (1982). Spatial versus tree representations of proximity data. *Psychometrika*, *47*, 3-19.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, *42*, 319-345.

Takane, Y. (2007). Handbook of statistics (Vol. 26): Pyschometrics. In C. R. Rao & S. Sinharay (Eds.), (p. 359-400). Amsterdam, The Netherlands: Elsevier.

Takane, Y., Young, F., & De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, *42*, 7-67.

Tversky, A., & Hutchinson, W. (1986). Nearest neighbor analysis of semantic spaces. *Psychological Review*, *93*, 3-22.