The spatial arrangement method of measuring similarity can capture high-dimensional, semantic structures

Russell Richie (drrichie@sas.upenn.edu)[1]
Bryan White (bryanw@nmsu.edu)[2]
Sudeep Bhatia (bhatiasu@sas.upenn.edu)[1]
Michael C. Hout (mhout@nmsu.edu)[2]

[1] Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104
[2] Psychology Department, New Mexico State University, Las Cruces, NM 88003

Abstract

Despite its centrality to cognition, similarity is expensive to measure, spurring development of techniques like the Spatial Arrangement Method (SpAM), wherein participants place items on a 2-dimensional plane such that proximity reflects similarity. While SpAM hastens similarity measurement, its suitability for higher-dimensional stimuli is unknown. In Study 1, we collected SpAM data for eight different categories composed of 20-30 words each. Participant-aggregated SpAM distances correlated strongly (r=.71) with pairwise similarity judgments, although below SpAM and pairwise judgment split-half reliabilities (r's>.9), and cross-validation with multidimensional scaling fits at increasing dimensionalities suggested that aggregated SpAM data favored higher dimensional solutions for 7 of the 8 categories. In study 2, we showed that SpAM can recover the Big Five factor space of personality traits, and that cross-validation favors a four- or five-dimension solution on this dataset. We conclude that SpAM is an accurate and reliable method of measuring similarity for high-dimensional items.

Keywords: similarity; multidimensional scaling; spatial cognition; concepts; traits; Big Five

Introduction

Similarity is central to cognitive science. Shepard's (1987) "universal law of generalization" holds that the probability of generalizing from one item to another decreases exponentially as a function of their dissimilarity. The Generalized Context Model (GCM) of categorization holds that a stimulus will be categorized with whatever set of exemplars it is most similar to (Nosofsky, 1984). Similarity is also often theorized to be a heuristic cue for many other more complex judgments, including probability judgment, social judgment, and causality (Kahneman & Tversky, 1972), is a key variable used by individuals to make multiattribute choices (Tversky, 1969), and is thought to play a fundamental role in memory, with retrieved items cuing the subsequent retrieval of other similar items (e.g. Howard & Kahana, 2002). Finally, to the extent that similarity reflects degree of sameness of representation, human similarity judgments in a domain can be used to infer representational structure in that domain, using techniques like additive clustering or multidimensional scaling.

Despite its theoretical and methodological importance, measuring similarity numerically is not straightforward. There are many popular methods, but each has its own strengths and weaknesses (for review, see Jaworksa & Chupetlovska, 2009). For example, one standard approach, the *pairwise method*, is to ask participants to rate the similarity (usually via Likert scale) between every possible pair of items in a domain. Despite the simplicity of this method, it has several drawbacks, chief among these being its inefficiency: collecting pairwise judgments for n items requires $n(n - 1)/2$ judgments. For just 30 items, this means 435 pairwise judgments, and doubling the set to a mere 60 items would require 1,770 comparisons.

To study similarity and its attendant phenomena more easily, researchers need cheap, reliable, and construct-valid methods for collecting similarity data. While there are many recent advances in this vein (e.g. Roads & Mozer, 2019), we focus here on an empirical technique first developed by Goldstone (1994) and repopularized recently by Hout, Goldinger, and Ferguson (2013): the Spatial Arrangement Method, or SpAM.

In SpAM, multiple items are simultaneously presented to a participant on a computer screen, and the participant is tasked with rearranging the items such that inter-item proximities correlate with similarity. Each participant thus provides a dissimilarity matrix via the Euclidean distances between their item placements. SpAM presents a number of advantages. First, it is intuitive for participants, as it relies on the spatial nature in which people tend to conceptualize similarity (Casasanto, 2008). Second, it is very fast, as each movement of an item simultaneously adjusts its proximity for all other items on the screen (*n-1* items if all items are presented simultaneously).

SpAM has accordingly seen many applications, in domains from letters (Goldstone, 1994) to architectural scenes (Berman et al., 2014). However, there has been relatively little empirical investigation of SpAM's suitability as a method for collecting similarity data. Of principal concern is that a *single* trial of SpAM more or less only allows a participant to perfectly represent two dimensions of a domain (Verheyen, Voorspoels, Vanpaemel, & Storms, 2016). This aspect of SpAM might therefore limit its ability to recover higher-dimensional (>2) similarity spaces. This may be especially problematic for rich conceptual stimuli like words (cf. pictures of simple objects used in much prior work with SpAM). However, if multiple trials are conducted (within or between participants), and different stimulus dimensions are attended on each trial, then aggregated

SpAM data could still recover higher-dimensional structures. Hence, the primary goal of the current studies was to empirically evaluate the ability of SpAM to reliably and accurately recover higher-dimensional similarity structures for lexical-semantic stimuli.

Study 1

Materials

We collected similarity data for eight categories: furniture, clothing, birds, vegetables, sports, vehicles, fruit, and professions. Each category contained 20-30 words referring to category members. Where possible these items were selected to be as similar as possible to those in the Leuven Concept Database (De Deyne et al., 2008). These and other study materials can be found in our OSF respository at https://bit.ly/2tmIChy.

SpAM

Participants We recruited 57 participants (mean age = 19.76, 63% female) from the student population of a large state university. Data from three participants were lost to computer failures, yielding usable data from 54 participants.

Design and procedure. The experiment, implemented in E-Prime (see OSF repository for code), consisted of eight trials, one for each category of words. Trials were presented in random order, with each trial consisting of a display broken into three sections. In the center of the screen was the “arena”; the area in which the words could be moved around and organized at distances proportional to their perceived dissimilarity. Outside the arena was the space (to the left and right of the arena) where the words were first randomly placed in columns at the beginning of each trial. Participants moved the items into the arena (using “click and drag”) one at a time, and were given as much time to arrange the arena as they saw fit. Each trial could only be completed once all the items had been moved into the arena.

Pairwise method

Participants. We recruited 365 participants (mean age = 33 years, 55% female) through Prolific Academic. We limited our data collection to participants who were from the U.S. and had an approval rate above 80%. Participants were only allowed to participate once, and were paid approximately $10 per hour.

Design and procedure. In contrast to SpAM, whose speed affords a within-subjects design, the pairwise method reasonably admitted only a between-subjects design whereby participants were randomly assigned to one of the eight categories – furniture (N = 33), clothing (N = 61), birds (N = 54), vegetables (N = 30), sports (N = 61), vehicles (N = 28), fruit (N = 31), and professions (N = 67). Twice as many participants were required for birds, clothing, professions, and sports because each participant randomly assigned to those categories only completed pairwise judgments for half of the pairs, due to category size. For each category, participants were instructed that we were interested in how people judge similarity of word meaning; participants used a Likert scale from 1 (not at all similar) to 7 (extremely similar) to provide their ratings.

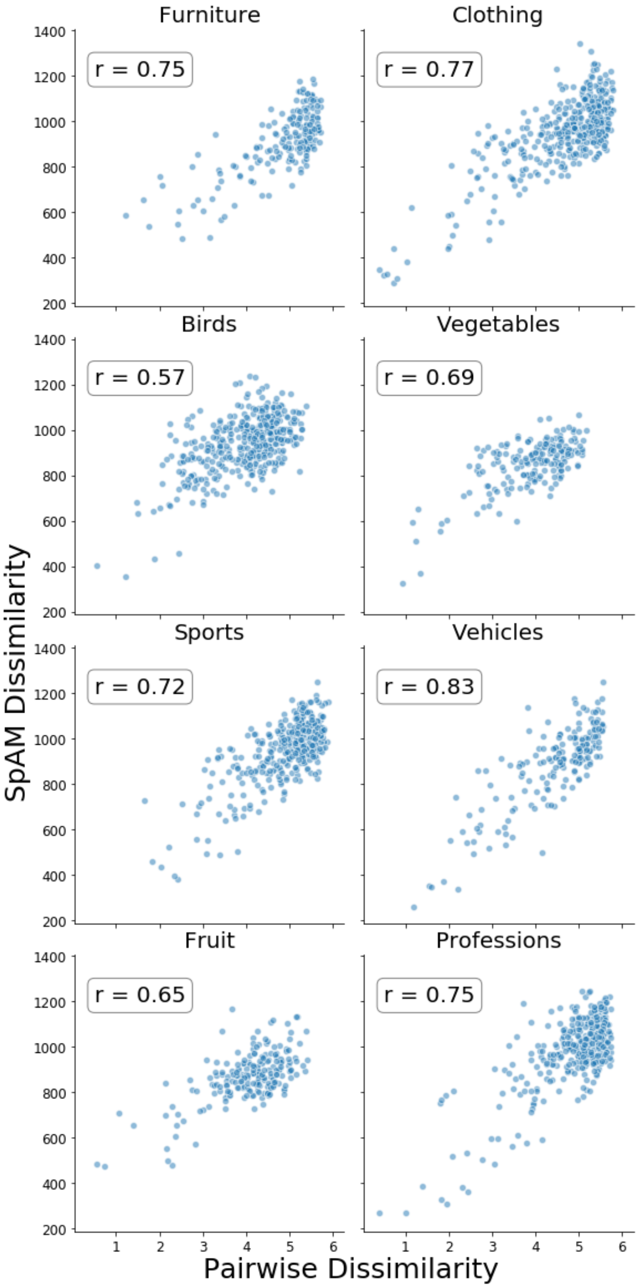

Figure 1. Average pairwise dissimilarity (x-axes) against average SpAM dissimilarity (y-axes). Pearson correlations are inset.

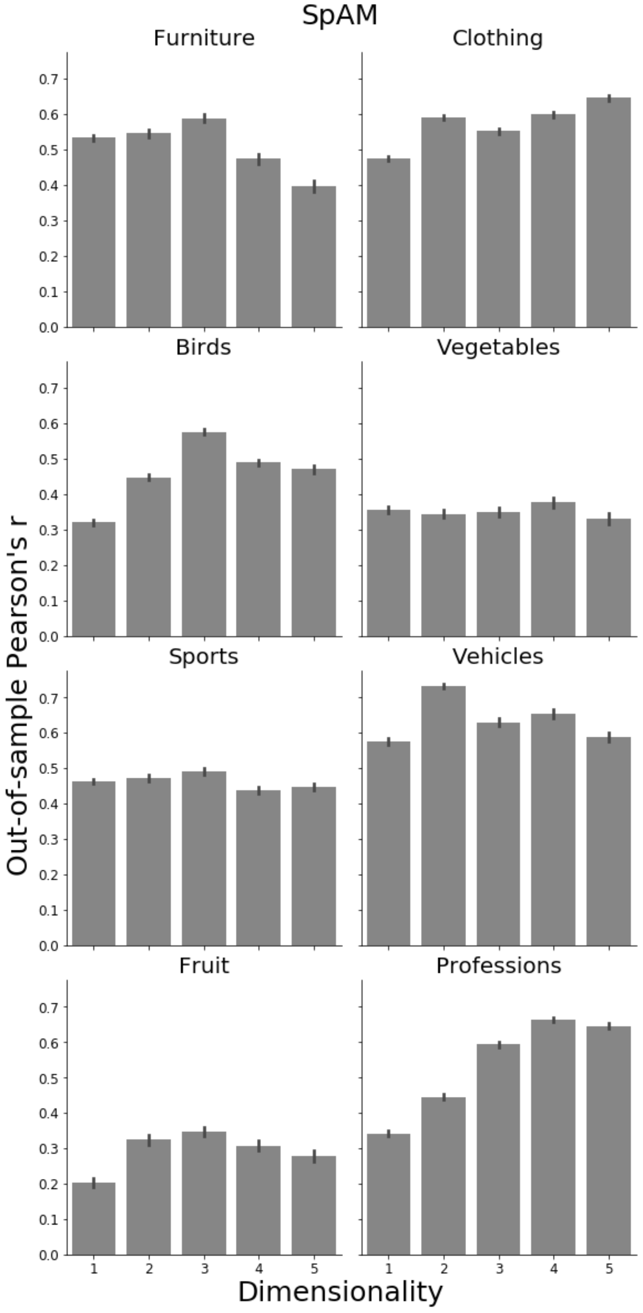

Figure 2. Cross-validation results for SpAM under MDS solutions of increasing dimensionality.

Results and discussion

First, we computed the Pearson correlation between pairwise method similarities averaged over all participants, and SpAM distances averaged over all participants. Figure 1 has scatterplots of average SpAM dissimilarities/distances plotted against average pairwise method dissimilarities for each category, with the Pearson correlation inset. These correlations are high, averaging .71 (p's<10^{-26}), but below the split-half reliability of SpAM (Pearson's r's with Spearman-Brown correction above .9 for every category) and that of the pairwise method (Pearson's r with Spearman-Brown correction of .94), suggesting that SpAM and the pairwise method measure largely, but not entirely, overlapping constructs of similarity. This is encouraging to the extent that pairwise Likert scale ratings are a standard, accepted measure of similarity whose ability to recover higher-dimensional spaces is not questioned, even though we do not argue that the pairwise method should be the "gold standard," per se.

Second, we conducted a cross-validation exercise with multidimensional scaling to determine the dimensionality latent in our SpAM data. We first averaged over participants to provide aggregate dissimilarity scores. We then conducted the following procedure 500 times for each category and dimensionality from 1 to 5 (inclusive). We randomly removed 20% of the non-diagonal (self-dissimilarity) entries in the aggregate SpAM dissimilarity matrix. We ensured that no more than half of the distances to a given word (i.e. values in a row or column) were removed, so that there was sufficient data to estimate the coordinates of every item. Using the smacof package in R (de Leeuw & Mair, 2009), we then fit MDS to this ablated matrix (smacof handles missing data by assigning a weight of 0 to those cells). Finally, we computed the Pearson's r correlation between (a) the Euclidean distances that the resulting MDS solution predicted for the held-out 20% of data, and (b) the true aggregated SpAM distances for the held-out 20% of data. Figure 2 visualizes the resulting correlations. Several (but not all) categories, like birds or professions, clearly seem to favor higher dimensional solutions. To substantiate this, we conducted t-tests comparing (a) the correlations between predicted and actual distances for the two-dimensional solution and (b) the correlations between predicted and actual distances for the higher-dimensional solution with the highest mean correlation. Every category except vehicles favored at least a three-dimensional MDS solution (p's < .05), suggesting that our aggregated SpAM data can recover higher dimensional semantic spaces.

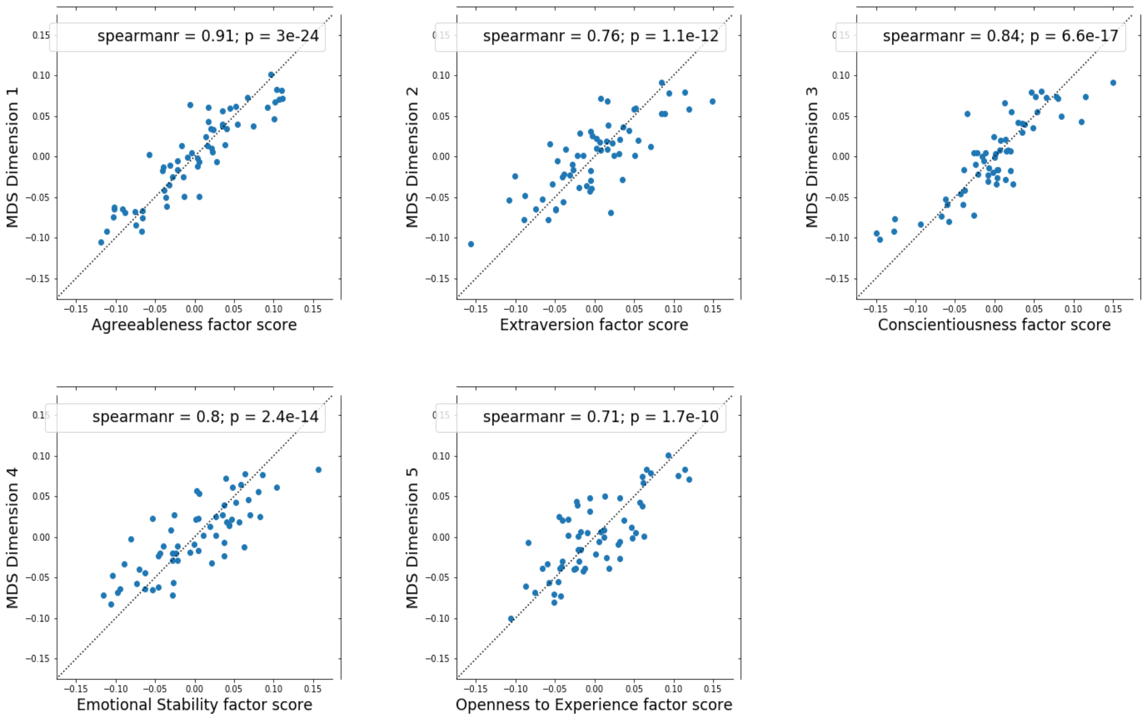

Figure 3: Scatterplots of Big Five factor loadings for personality trait adjectives against MDS dimensions from aggregated SpAM Procrustes-aligned to the Big Five factor space.

We conducted the same cross-validation exercise with participant-averaged pairwise data (after having transformed pairwise method similarities into dissimilarities by subtracting each average pairwise similarity rating from 7, the maximum value on the similarity scale). Whereas SpAM favored higher-dimensional solutions for 7 of 8 categories, pairwise methods favored higher-dimensional solutions for only 6 of 8 categories. However, within those 6, the pairwise method favored higher dimensional solutions than did SpAM (e.g., for birds, SpAM favors 3 dimensions whereas pairwise favors 5 dimensions). Thus, SpAM is not uniformly worse than the pairwise method in recovering higher-dimensional spaces. The important point is that this provides the first clear evidence that, despite the two-dimensional imposition of a single SpAM trial, aggregating over multiple SpAM trials can recover higher-dimensional lexical-semantic spaces.

Study 2

The cross-validation exercises we reported in Study 1 provide support for the idea that (aggregated) SpAM data can recover high-dimensional semantic spaces. However, another way to demonstrate SpAM's ability to recover high-dimensional spaces is to examine the extent to which aggregated SpAM data can recover a priori known spaces. One such thoroughly characterized domain of the lexicon is personality trait adjectives, which are theorized to adhere to the so-called Big Five factor structure, also known as the OCEAN model of traits (John, Naumann, & Soto, 2008). According to this model, personality trait adjectives (and people's personalities) primarily vary on five dimensions or factors: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. This model has a long history and empirical basis, largely in factor analyses or principal component analyses of participants' self- or other-ratings of large numbers of personality trait adjectives (e.g., Ashton, Lee, & Boies, 2015). One output of such analyses is factor or component loadings for each personality trait adjective – the extent to which each personality trait adjective scores high or low on each of the Big Five dimensions. Thus, our primary goal in Study 2 was to test whether aggregated SpAM data could recover these Big Five factor scores from a set of personality trait adjectives.

Method

Participants. We recruited 58 participants (mean age = 19.55, 79% female) from the student population of a large state university.

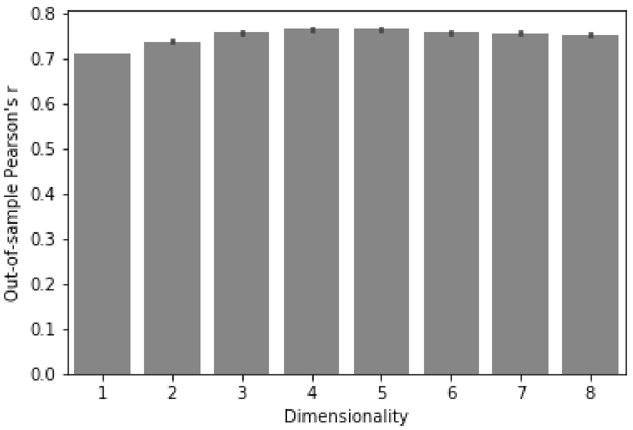

Figure 4. Cross-validation results for personality trait adjectives under MDS solutions of increasing dimensionality.

Materials. We obtained Big Five factor scores for 435 personality trait adjectives from Ashton et al., (2015). To obtain a sample that was (a) small enough to complete in a single experimental session, yet (b) maximally spanned the Big Five factor space, we employed the following procedure. We (1) randomly sampled 60 personality traits, (2) fit a PCA to their Big Five factor scores, (3) determined the percentage of variance each of the five components accounted for, and (4) computed the entropy among the distribution of variance explained from (3). We repeated this procedure 100,000 times, and took the random sample of traits with the maximum entropy in the distribution of explained variance. This ensured that the Big Five factors were maximally orthogonal in our sample.

SpAM. The SpAM procedure adopted here was largely consistent with that of Study 1 but had to be adapted to accommodate the much larger set of stimuli employed in Study 2. Sixty trait words is too many items to present to a participant simultaneously (as was done with the smaller stimulus sets in Study 1). As such, subsets of the stimuli were shown to each participant across multiple SpAM trials in the following manner.

On each trial, 25 different words were shown to the participant. Display and interface characteristics were identical to Study 1. The main difference in procedure was that rather than switching to a new category of items after each trial, the participant was simply shown a different subset of the 60 trait words across a set of ten total SpAM trials. This procedure ensured that each word was paired with every other word at least (but sometimes more than) once. Thus, each participant provided a complete similarity matrix for the set of 60 words.

Selection of words across trials was determined by employing a stimulus selection algorithm designed to minimize the number of trials or blocks in an incomplete block design like the one adopted here, such that all possible pairings of words occurred in at least one trial/block (MacDonald, Hout, & Schmidt, 2019). For most combinations of total stimulus set size and subset size a list of "blocks" does not exist in which each pair of items is presented exactly once (see Discussion in MacDonald et al., 2019). As such, in our adopted design, some items were paired with others on more than one trial, leading to multiple observations per "cell." To balance out such redundancies across participants, words were randomly assigned numeric identifiers in the algorithmic block set. This ensured that each participant saw each pair of words together at least once across the ten trials, but also ensured that different participants were presented with different redundancies in the pairings.

Results and discussion

The primary test of our SpAM data on personality traits was whether they recovered previously obtained Big Five factor scores. To this end, we first averaged SpAM distances between every pair of words, over all trials, yielding a single, aggregate 60-by-60 Euclidean distance matrix. We then submitted this dissimilarity matrix to the smacof MDS algorithm in the scikit-learn library (Pedregosa et al., 2011), setting the dimensionality to 5, and using the other default hyperparameters. We then applied Procrustes analysis to the resulting 5-dimensional coordinate solution and to the Big Five factor scores for our 60 words, to find the translation, scaling, and rotation of the MDS solution that best aligned it with the Big Five factor space. Figure 3 displays scatterplots and Spearman correlations (and p-values) of each of the five factor scores against its corresponding Procrustes-transformed dimension of the MDS space. As can be seen, the aggregated SpAM data reproduce the Big Five factor space very well; the Big Five factors correlate with the Procrustes-transformed MDS dimensions between .71 (Openness to Experience) and .91 (Agreeableness), with an average of .81 (all p's < 10-9).

We also subjected our aggregated SpAM data on personality traits to the same cross-validation exercise as in Study 1. See Figure 4 for a barplot of out-of-sample correlations for dimensionalities from 1 to 8. Again, higher-dimensional solutions (>2) were favored. Four- and five-dimensional solutions have very similar out-of-sample performance, with correlations of r=.7659 and r=.7662, respectively (t(998)=.22, p=.81). Both dimensionalities are superior to all other tested dimensionalities (all t(998)>4.67, all p's<10-6), consistent with much previous research suggesting a four or five factor structure in the trait lexicon (Ashton et al., 2015; John et al., 2008).

General Discussion

Similarity data are useful for a variety of applications in cognitive science, yet prominent methods for collecting similarity data are often time-consuming or otherwise flawed. Here, we evaluated the Spatial Arrangement Method (SpAM; Goldstone, 1994; Hout et al., 2013) as applied to the collection of similarities between words, stimuli whose high-dimensionality could have, in principal, stymied the 2-

dimensional nature of SpAM. We have two key results. First, for eight common categories of words, raw SpAM distances correspond with raw pairwise method similarities strongly (average r=.71), suggesting that the two techniques measure largely overlapping constructs of similarity even with high-dimensional lexical-semantic stimuli. Second, we showed that SpAM can in fact reliably recover higher-dimensional spaces. We demonstrated this both with a cross-validation exercise, selecting the MDS dimensionality that best predicted held-out SpAM dissimilarities (Studies 1 and 2), and by using an MDS solution applied to SpAM data to recover an a priori known high-dimensional semantic structure, the Big Five factor structure of personality traits (Study 2).

Although we suggest that SpAM can recover more than 2 dimensions in aggregate MDS solutions because different participants choose to focus on different pairs of dimensions in their individual SpAM map, we do acknowledge that the 2-dimensional nature of SpAM is likely a major factor for why SpAM correlates with the pairwise method imperfectly. In particular, it seems likely that, for a given domain, some dimensions may be more salient or meaningful than others, even if only slightly so. If participants' choice of dimensions to attend to is a (nearly) deterministic function of salience, then all but the two most salient dimensions will tend to be neglected in most SpAM trials. This will be most problematic if the salience/importance of the dimensions is more or less uniform, such that the two most salient dimensions are only barely the most salient. It is possible, in this case, for the majority of the perceived (salience-weighted) variance in the domain to go unmeasured in the aggregate SpAM data. This is not as problematic for the pairwise method, where participants can in principal (up to limits on attention and working memory, and subject to noise) give a similarity rating which is perfectly reflective of more than two dimensions.

Future work might compare SpAM to both the pairwise method and other emerging techniques like Best-Worst Scaling (Hollis & Westbury, 2018) and generalizations of the triad task (or odd-one-out task; Roads & Mozer, 2019), in both their ability to recover known spaces, and their ability to predict downstream behavior like category learning. Of particular importance will be the *amount of participant time* each technique needs to reach a given level of accuracy in recovering a known space or in predicting other behaviors. Such 'downstream' tests of SpAM's ability to predict other cognition and behavior may ultimately be the most important for evaluation of SpAM or other techniques of measuring similarity.

Acknowledgments

This work was funded by National Science Foundation grant SES-1626825, and the Alfred P. Sloan Foundation.

References

Ashton, M. C., Lee, K., & Boies, K. (2015). One-through six-component solutions from ratings on familiar English personality-descriptive adjectives. *Journal of Individual Differences, 36*(3), 183.

Berman, M. G., Hout, M. C., Kardan, O., Hunter, M., Yourganov, G., Henderson, J. M., Hanayik, T., Karimi, H., & Jonides, J. (2014). The perception of naturalness correlates with low-level visual features of environmental scenes. *PLoS ONE, 9,* e114572. doi: 10.1371/journal.pone.0114572.

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

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods, 40*(4), 1030–1048. https://doi.org/10.3758/BRM.40.4.1030

de Leeuw J., & Mair P. (2009) Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software, 31*(3), 1-30.

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

Hollis, G., & Westbury, C. (2018). When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior research methods, 50*(1), 115-133.

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013a). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General, 142*(1), 256–281. https://doi.org/10.1037/a0028860

Howard, M. W., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language, 46*, 85–98. http://dx.doi.org/10.1006/jmla.2001.2798

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), 1-10.

John, O. P., Naumann, L., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114-158). New York, NY: Guilford.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

MacDonald, J., Hout, M. C., & Schmidt, J. (2019). An algorithm to minimize the number of blocks in incomplete block designs. *Behavior Research Methods*. doi: 10.3758/ s13428-019-01326-x.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 104–114.

Pedregosa F., et al. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Roads, B. D., & Mozer, M. C. (2019). Obtaining psychological embeddings through joint kernel and metric learning. *Behavior Research Methods*. doi: 10.3758/s13428-019-01285-3

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

Tversky, A. (1969). Intransitivity of preferences. *Psychological review*, *76*(1), 31.

Verheyen, S., Voorspoels, W., Vanpaemel, W., & Storms, G. (2016). Caveats for the spatial arrangement method: Comment on Hout, Goldinger, and Ferguson (2013). *Journal of Experimental Psychology: General*, *145*, 376 – 382. http://dx.doi.org/10.1037/a0039758