Whether the Pairwise Rating Method and the Spatial Arrangement Method yield comparable dimensionalities depends on the dimensionality choice procedure

Steven Verheyen a, *, Gert Storms b

a Erasmus University Rotterdam, Rotterdam, the Netherlands
b KU Leuven, Leuven, Belgium

ARTICLE INFO

Keywords:
Similarity
Dissimilarity
Multidimensional scaling
Dimensionality
Reliability
Spatial representation

ABSTRACT

We investigate whether the Pairwise Rating Method (PRaM) and the Spatial Arrangement Method (SpAM) yield multidimensional scaling (MDS) solutions of comparable dimensionality. Across three studies that included twelve semantic categories with varying numbers of both pictorial and verbal exemplars, we did not find consistent dimensionality differences between the two similarity measurement methods. The results alleviate the concern that SpAM might underestimate the dimensionality of high-dimensional stimuli compared to PRaM. However, the resulting number of dimensions was found to be highly dependent on the dimensionality choice procedure, indicating the need for a more systematic investigation into dimensionality selection for MDS.

1. Introduction

Similarity is at the heart of cognitive science theories concerning identification, categorization, generalization, reasoning, and memory retrieval (Goldstone and Son, 2012; Hahn, 2014). The more similar stimuli are perceived to be, the more likely they will be identified as the same, judged to belong to the same category, inferred to display comparable properties and behaviour, and to cue each other. Spatial representations of similarity dominate the literature (Gärdenfors, 2000; Jones et al., 2018; Shepard, 1960). They represent measurements of (dis)similarity among pairs of stimuli as distances between points in a low-dimensional space. The larger the similarity between two stimuli, the closer the points representing them will be positioned in space. Spatial representations provide visual depictions of the empirical interrelations between the stimuli, which allow one to see and explore the structure of the data better than using the numerical similarity indices. Take the two-dimensional representation of vehicles in Fig. 1 as an example. The relative positioning of the different vehicles in space allows one to interpret the structure of the stimulus set much better than the corresponding numerical information in the matrix at the bottom of Fig. 1. Although the entries in the matrix accurately reflect the Euclidean distances between the stimuli in the space and therefore have the same informational content as its two-dimensional representation, they do not allow for the straightforward interpretation the space does in terms of two-wheelers, four-wheelers, public transportation, aircrafts, and the subtle relations between them. The relative ease with which a visual depiction of (dis)similarities is interpreted compared to numerical indices of similarity, explains the appeal of spatial representations (Borg and Groenen, 2005).

Spatial representations are also favoured over other visual representations. Whereas similarity in spatial representations changes with the stimuli’s positions along the continuous dimensions that constitute the space, the similarity in feature-based models increases with the addition of common features and the deletion of distinctive features (and vice versa) following Tversky’s contrast model (Tversky, 1977). Several alternative (dis)similarity representations in terms of features have been proposed (e.g., hierarchical clustering, Johnson, 1967; additive trees, Sattath and Tversky, 1977; additive clustering, Shepard and Arabie, 1979; extended trees, Corter and Tversky, 1986), but they have been rarely applied (see Johnson and Tversky, 1983; Malt, 1994; and Tenenbaum, 1996, for a few notable exceptions). The amenity of spatial representations not only shows in the paucity of its alternatives, but also when different visual representations are explicitly compared and spatial representations are found to yield superior information retrieval (e.g., Butavicius and Lee, 2007; Butavicius et al., 2012). This “competitive” advantage of spatial representations probably stems from the
spatial nature of our environment and our familiarity with maps representing it (Hout et al., 2013; Tolman, 1948). Indeed, it has been repeatedly shown that people intuitively conceptualize similarity in a spatial manner (Boot and Pecher, 2010; Breaux and Feist, 2008; Casasanto, 2008; Winter and Matlock, 2013) and even in our language we rely on our proficiency in the spatial domain, for instance when we use spatial metaphors to make sense of abstract concepts (Boroditsky, 2000; Lakoff and Johnson, 2003). Spaces are clearly a natural way for us to represent relationships.

2. Multidimensional scaling

Multidimensional scaling (MDS; Borg and Groenen, 2005; Kruskal and Wish, 1978) is the statistical procedure that is most commonly used to obtain a spatial representation based on input proximities. (The term ‘proximities’ is here used to refer both to measures of similarity and to measures of dissimilarity.) The inference problem that MDS solves in a sense is the reverse of the problem we resolve when we determine the distance between two locations using a map. If we needed to determine how far the US state capitals are apart from each other (as the bird flies), we would measure the distance between each pair of capitals on a map of the United States (in cm or inch). By multiplying this distance with the scale factor indicated on the map, we would obtain the actual distance between the capitals (in kilometres or miles). If one only had the inter-capital distances at one’s disposal, one could apply MDS to reconstruct a map of the US state capitals. MDS would then yield the map coordinates of each of the capitals. In order to achieve this, MDS uses three functions (Borg and Groenen, 2005). The distance function defines how distance is measured in the resulting space. The Euclidean distance function, for example, defines the distance between two points in space as the square root of the sum of the squared differences between their coordinates. The representation function defines the relationship between the input dissimilarities and the distances in the resulting space. This relationship is generally not completely specified, but is restricted to a particular class of transformations that define the types of MDS: a positive multiplicative transformation (ratio MDS), a positive linear transformation (interval MDS), or a monotonically increasing transformation (ordinal or non-metric MDS). The decision for a particular transformation is generally based on the measurement level of the input data (Stevens, 1946, 1951). Finally, the loss function quantifies the fit between the input data and the output distances. As was the case for the other functions, there exist several loss functions but most make use of the squared error of representation, defined as the squared difference between the optimally transformed input dissimilarities and the output distances. Usually, these squared errors of representation are summed across all stimulus pairs to yield an indication of the overall fit of the spatial configuration to the input data. The commonly used stress-1 loss function provides a normalization of this overall badness-of-fit measure.

\[ \text{stress-1} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (d_{ij} - d_{ij}^*)^2} \]

That does not mean that spaces are always the most appropriate or optimal way of representing similarities. Depending on the distribution of the (dis)similarities, feature-based representations may capture the relationships between the stimuli better (e.g., Dry and Storms, 2009; Ghose, 1998; Giordano et al., 2011; Pruzansky et al., 1982; Sattath and Tversky, 1977; Tversky and Hutchinson, 1986; Verheyen et al., 2016; Verheyen et al., 2020).

![Fig. 1. Two-dimensional spatial representation (top) of the empirical dissimilarities (bottom) between 16 exemplars of the category vehicles.](image)
by dividing it by the sum of the squared distances and taking the square root of the result. Unlike the US capitals example, where accurate measurements of the true distances between the cities can serve as input to MDS, empirical measurements of other (dis)similarity relations tend to be noisy. In most MDS applications, we therefore do not expect the stress function to be zero (nor do we want it to be since this would likely yield error fitting).

Any set of input proximities can be represented near perfectly in a space with a sufficiently high number of dimensions. Typically, MDS users will want to obtain a representation with fewer dimensions, however, to be able to explore the resulting visualization and establish the underlying structure in terms of a limited number of interpretable dimensions (Borg and Groenen, 2005; Kruskal and Wish, 1978). In cases where the dimensionality of the space is not known beforehand, MDS users will then need to obtain representations in spaces with varying numbers of dimensions. For each dimensionality set by the end-user, MDS will provide the spatial representation of the input proximities that minimizes the loss function, leaving it to the user to decide on the optimal dimensionality. A number of approaches to establish the “true” dimensionality have been proposed in the literature, but there exists no consensus as to the preferred method. A review of 125 MDS analyses of semantic concepts, for which the authors had no a priori reasons to prefer a particular dimensionality, learned that the majority of authors did not justify their choice for a particular dimensionality or based their decision on either parsimony or interpretability (Verheyen et al., 2007).

About a third of the reviewed analyses referred to Kruskal’s (1964) verbal characterizations of absolute stress values ranging from ‘poor’ to ‘perfect’ to motivate a particular dimensionality choice. It thus appears that the most commonly used dimensionality selection procedures have an element of subjectivity to it, in that parsimony, interpretability, and which verbal characterization is deemed satisfactory, are ultimately subject to the end user’s judgment. There exist more objective ways of the determining the dimensionality in MDS, but the review by Verheyen et al. learned that these are hardly used in the literature. We will return to these more objective procedures in later sections, where we will use them to establish whether the manner in which (dis)similarity is measured may also affect the dimensionality choice.

3. Comparing similarity measurement methods

In this paper, we will compare two popular methods for collecting (dis)similarity data on their ability to yield high-dimensional MDS solutions. The Pairwise Rating Method is arguably the method that has most commonly been used to obtain similarity data from participants, but is increasingly being replaced by the Spatial Arrangement Method (see Koch et al., 2020, for a citation analysis).

In the Pairwise Rating Method (PRaM), participants rate the (dis)similarity of all pairs of stimuli on a Likert scale. Participants tend to find this procedure straightforward and clear, but very tiresome (Verheyen et al., 2020). While most participants will be accustomed with the nature of the task and will not find it difficult to determine the extent to which two stimuli resemble each other, PRaM has the disadvantage that it may take very long to judge all the pairs that make up a (relatively large) stimulus set (Giordano et al., 2011; Kriegeskorte and Mur, 2012; Tsogo et al., 2000; Verheyen et al., 2020). The addition of a single stimulus to a set of n stimuli requires n additional judgments to be made, which can rapidly add up. Collecting data over an extended period carries the risk that participants will come to prioritize other dimensions, recalibrate the scale, or provide incorrect answers as the task progresses (Hout et al., 2013; Koch et al., 2020).

In the Spatial Arrangement Method (SpAM; Goldstone, 1994; Hout et al., 2013), participants organize stimuli on a computer screen so that the distance between stimuli represents their perceived dissimilarity. Participants tend to find this procedure rather pleasant, applauding the task’s incremental and comprehensive nature (Verheyen et al., 2020). In addition, SpAM is found to be less tiresome than PRaM. Because it makes use of the relations among the stimuli, it takes far less time to complete.

If two stimuli resemble each other closely, but are distinct from a third, this can be conveyed at once by placing that stimulus at a long distance from the other two. This single action simultaneously establishes its dissimilarity to the pair of original stimuli, avoiding the requirement to indicate this dissimilarity on separate trials for each of the stimuli in the pair (Goldstone, 1994; Hout et al., 2013). A disadvantage of SpAM is that the two-dimensional nature of the screen on which the stimuli ought to be organized, only allows participants to communicate two dimensions at a time (Kriegeskorte and Mur, 2012; Verheyen et al., 2016). Verheyen et al. (2016) voiced the concern that under these circumstances, participants might decide to only communicate the two most salient dimensions of variation and that less salient dimensions might remain uncovered. As a result, the use of SpAM might make MDS-users underestimate the true dimensionality of the stimuli’s representation. In order for the spatial representation of SpAM dissimilarity data to include all relevant dimensions, the SpAM data from different participants would have to be combined and the relative importance of the dimensions constituting the high-dimensional space that these participants share, would have to be systematically related to the choice of the SpAM dimensions across participants. Put differently, the relative importance of the various stimulus dimensions would have to be reflected in the relative number of participants who decide to convey these dimensions in their two-dimensional SpAM configurations.

Verheyen et al. (2016) used individual differences scaling (INDSCAL; Carroll and Chang, 1970; Takane, 2007; Takane et al., 1977) to re-analyze the PRaM and SpAM proximity data obtained by Hout et al. (2013) for two sets of visual stimuli constructed to vary along three perceptual dimensions each (27 wheels varying in their thickness, hue, and the angle of the spoke, and 27 bugs varying in their shading, the number of legs, and the curvature of their antennae). INDSCAL is a version of MDS that assumes that the stimuli are embedded in a multi-dimensional group space that is shared by all individuals, but allows the weight that is attached to each of these dimensions to vary between individuals. Individuals’ spaces can then be constructed from the group space, by multiplying the stimuli’s coordinates in the group space with the individuals’ weights for each dimension. Verheyen et al. found that while most PRaM participants had positive weights for all three dimensions of the stimuli, many SpAM participants had a weight close to zero for one of the three dimensions. This finding confirmed that individual participants can only clearly convey two dimensions of variation in SpAM, while individual PRaM participants can at least convey three dimensions of variation at once. Despite the fact that individual SpAM participants could only manoeuvre stimuli in two dimensions, a three-dimensional MDS representation that reflected the stimuli’s dimensions of variation could nevertheless be constructed by aggregating the data from multiple SpAM participants. It has now repeatedly been shown that averaging the SpAM data from several participants can yield MDS configurations for which the loss function decreases considerably beyond two dimensions (e.g., Hout and Goldinger, 2016; Koch et al., 2016; Richie et al., 2020) suggesting that despite the two-dimensional

---

2 This argumentation assumes that each participant provides a single spatial organization of the entire stimulus domain. It has been argued that one can have participants convey more than two dimensions by having them provide multiple spatial organizations for subsets of the stimulus domain (Goldstone, 1994; Kriegeskorte and Mur, 2012). The different composition of stimuli on each trial would bring other dimensions of variation into focus.

3 SpAM participants who did attempt to convey three dimensions in the two-dimensional spatial arrangement were (necessarily) found to do this in a sub-optimal manner.
nature of the task, spatial representations with more than two dimensions can be obtained using SpAM, as long as data from a sufficient number of participants is pooled together.

This study examines whether the spatial representations of proximity data collected with the Spatial Arrangement Method (SpAM) and with the Pairwise Rating Method (PRaM) have a comparable dimensionality. The observation that aggregating SpAM data across participants can yield spatial representations with more than two dimensions does not necessarily imply that the resulting number of dimensions will be as high as those obtained with other methods for measuring similarity. Conversely, the observation by Verheyen et al. (2016) that individual PRaM participants were able to convey three dimensions at once does not guarantee that participants will be able to convey more than three dimensions at once. Although in PRaM, participants are only required to judge the similarity of two stimuli at once, it is likely that there are cognitive limitations to the number of dimensions that can inform such a judgment. PRaM too, might therefore rely on the aggregation of data from several participants to obtain high-dimensional representations. Given our extensive reliance on the construct of similarity and our use of multi-dimensional spatial representations of (dis)similarity for the explanation of a host of phenomena, a good understanding of the manner in which we obtain similarity data and its implications for the dimensionality of the resulting representations is warranted.

We will present three studies that are aimed at establishing the dimensionality of conceptual similarity data obtained with PRaM and SpAM. Conceptual stimuli are well suited to test whether the spatial representations that result from these similarity methods have a comparable dimensionality (see also Richie et al., 2020; Verheyen et al., 2020). For one, the representation of semantic concepts is one of the most important domains of application of MDS. Many researchers in the domain of concepts and categories assume that category exemplars are embedded in a metric space that captures both the relations between categories and the internal structure of each separate category (e.g., Ameel and Storms, 2006; Caramazza et al., 1976; Douven, 2016; Rips et al., 1973). In addition, the spatial representations of conceptual categories tend to have a large number of dimensions (e.g., Nosofsky et al., 2018b, 2020; Richie et al., 2019b; Verheyen et al., 2007), which are not all equally salient (De Deyne et al., 2014; Nosofsky et al., 2018a; Smits et al., 2002). Crucially, evidence is amounting that one needs to move beyond the two or three dimensions that are usually retained to adequately account for phenomena that are central in the categories and concepts literature. Nosofsky and colleagues, for instance, established that even eight dimensions were insufficient to adequately account for classification learning in real-world, naturalistic domains (Nosofsky et al., 2019) and therefore supplemented their multidimensional space of rocks with five additional dimensions, for a total of thirteen (Nosofsky et al., 2020). Verheyen et al. (2007) provide similar evidence for the modeling of typicality of other conceptual categories. By investigating conceptual stimuli, we thus expect to arrive at spatial representations with more than two dimensions, that – unlike the dimensions of the perceptual stimuli on which SpAM and PRaM have hitherto been compared – might not all be considered equally important by participants. This rich but graded structure of conceptual stimuli poses an interesting base of comparison for SpAM and PRaM, in that individual participants might only be able to communicate a subset of the stimuli’s dimensions of variation on a single SpAM or PRaM trial, and a spatial representation encompassing more dimensions may thus need to come from the aggregation of data from multiple participants. If these salience differences between these dimensions are pronounced and perceived similarly across participants, the dimensionality of the resulting aggregate solutions may differ between the two methods if participants were only to communicate the most salient dimensions and PRaM would allow individual participants to communicate more of them than SpAM.5

This paper comprises three studies with four conceptual categories each. All three studies make use of existing similarity data, obtained with SpAM and PRaM. For SpAM, the entire set of stimuli was arranged in terms of similarity on a single trial. For PRaM, participants judged all stimulus pairs in terms of similarity, except in Study 3 where they only judged half of the stimulus pairs. The data sets for the studies were chosen because both SpAM and PRaM participants were aware of the stimuli’s dissimilarity range prior to providing similarity measures, so as not to confound potential dimensionality differences with contextualization differences. In SpAM, judgments are naturally contextualized since all the stimuli in the set are presented simultaneously to participants for them to organize. In PRaM, this is not necessarily the case. When pairs of stimuli are presented one after the other, participants might only become aware of the relevant dimensions of comparison after a considerable number of trials is completed, rendering the judgments on these first trials unrepresentative. To avoid this, the PRaM data in Study 1 were obtained with the total-set version of the pairwise rating method (Hout et al., 2013). In this version of PRaM, all stimuli are present on the screen at all times and the pair to judge on a particular trial is highlighted. In Studies 2 and 3, participants were familiarized with the entire set of stimuli before starting the pairwise judgment of individual pairs.

The similarity data for Study 1 were taken from Study 1 of Verheyen et al. (2020), in which participants indicated the similarity between photorealistic images representing the 16 most familiar exemplars from the categories birds, vegetables, vehicles, and sports. The study employed a within-subjects design, whereby each participant alternatingly used PRaM and SpAM to indicate the similarity between the categories’ exemplars. The similarity data for Study 2 and Study 3 were taken from Study 1 of Richie et al. (2020) in which participants indicated the similarity between words representing the exemplars from the categories furniture, vegetables, vehicles, and fruit (Study 2) and the categories professions, sports, clothing and birds (Study 3) using either PRaM or SpAM (between-subjects design). Richie et al. had participants judge the similarity of all exemplar pairs of the former categories using PRaM, but decided to only present participants with a random half of the exemplar pairs of the latter categories, because they comprised more exemplars and having participants judge all of their exemplar pairs using PRaM would therefore require a prohibitive amount of time. Studies 2 and 3 serve several purposes. They are meant to generalize the findings from Study 1 to categories with a larger set of items, potentially yielding a more complex structure of higher dimensionality. The stimuli are presented as words, rather than pictures, to alleviate the concern that participants in Study 1 were judging the visual similarity of the specific pictures instead of the conceptual similarity of the basic level categories they represent. Studies 2 and 3 were pre-registered and employ data that have been gathered by external researchers to avoid concerns of selective reporting of results. This choice does come with one disadvantage. While the similarity data in Study 1 were obtained with a within-subjects design, the similarity method was manipulated between subjects in Studies 2 and 3. As a result, the SpAM and PRaM

5 From this, it should be clear that we do not claim that PRaM yields the true dimensionality, which SpAM should approximate. As we have indicated elsewhere (Verheyen et al., 2016; Verheyen et al., 2020), we do not believe one can determine the true similarity structure of a stimulus set in a task-independent manner (see also Goldstone and Medin, 1994). At most, we expect PRaM to yield a higher dimensionality than SpAM because individual participants can and have been shown to convey more than two dimensions of variation using PRaM, and as a result less salient dimensions might be more easily uncovered. We deem it likely that both methods underestimate the true dimensionality because they only allow a limited number of dimensions to be conveyed on a single trial and participants are likely to convey the most salient ones.
obtained with MDS for the twelve conceptual categories from Studies 1 and 3, different participants provided the input similarities. Any differences in dimensionality between the methods could thus potentially be due to differences in the make-up of the respective participant samples. Finally, Study 3 allows for an investigation of spatial representations’ dimensionality when the number of items per category becomes so large that it is still feasible to obtain all pairwise comparisons on a single trial using SpAM, but only a subset of exemplar pairs can realistically be presented for PrAM.

In the next section, we will describe in detail how we will go about establishing whether the dimensionality of the spatial representations obtained with MDS for the twelve conceptual categories from Studies 1 to 3 differs between the SpAM and PrAM similarity data.

4. Research outline

For each combination of category and proximity collection method, we will average the proximity data across participants. The resulting average proximity matrices will be subjected to multidimensional scaling assuming the Euclidean distance function, a monotonically increasing (non-metric) transformation function, and the stress-1 loss function. These choices represent the standard in multidimensional scaling of conceptual data. The Euclidean distance function is preferred when the dimensions constituting the space are not a priori known because it is more robust against the dimensionality choice than other distance functions and the most straightforward one to interpret (Oh and Raftery, 2001). Non-metric MDS imposes fewer restrictions on the resulting representations than its metric counterparts. Stress-1 is the most commonly used loss function for non-metric MDS because it has practical advantages over its alternatives (Borg and Groenen, 2005). Starting from an initial configuration obtained using classical scaling (Torgerson, 1958), the smacof package (De Leeuw and Mair, 2009) in R version 3.6.1 (R Core Team, 2016) will be used to obtain the final MDS configurations. Since smacof expects dissimilarities as input, the average PrAM similarities will be subtracted from the maximum scale point +1 to yield dissimilarities. (One is added to the maximum scale point because the similarity scale starts at 1.) MDS solutions in dimensionality 2 to D will be obtained, where D is such that it does not exceed the number of scaled instances divided by 4 (confirming to a liberal reading of the guideline by Kruskal and Wish, 1978, on the ratio of stimuli and dimensions).6

We will only consider MDS solutions that converged and that can be reliably discerned from random data. To verify the latter criterion, we generated 10,000 random proximity data sets with the same number of instances as the target category. The proximities were drawn from a uniform distribution between 0 and 1. These random proximity data sets were then subjected to multidimensional scaling and the mean and the standard deviation of the resulting stress values were extracted. For an empirical MDS solution to be reliable discernible from random data, its stress should be lower than a criterion value that is equal to the mean stress of the random data sets minus three times the standard deviation of the stress of the random data sets (Spence and Ogilvie, 1973). All the multidimensional scaling solutions that are reported in this paper passed these criteria.

Rather than to use the subjective dimensionality choice procedures (parsimony, interpretability, verbal characterizations of absolute stress values) that dominate the literature (Verheyen et al., 2007), we will employ three objective procedures to determine the appropriate number of dimensions for multidimensional scaling solutions of proximities obtained with either SpAM or PrAM. In earlier work, we suggested that compared to PrAM, SpAM would be biased against high-dimensional representations Verheyen et al. (2016). We therefore believe it is important to conduct the comparison of the two proximity data collection methods using objective and thus impartial procedures. To avoid any suggestions of selection bias, we use the three objective dimensionality choice procedures we employed in an earlier investigation of multidimensional scaling of conceptual proximity data Verheyen et al. (2007). The first is based on the reliability of the input proximities, the second on simulations of proximity data with varying levels of noise, and the third on the predictive ability of the obtained MDS solutions. It is noteworthy that the latter method was used to argue for high-dimensional spatial representations of conceptual proximity data in the original work.

4.1. Reliability of the input proximities

The first dimensionality choice procedure is based on the reliability of the input proximities. The dimensionality is chosen for which 1 minus the stress resembles the reliability of the input proximities the closest (Kruskal, 1964; Borg and Groenen, 2005). The underlying idea is to obtain a spatial configuration for which the error matches the random component of the data. The reliability will be determined as the mean split-half correlation corrected with the Spearman-Brown formula, across 1000 random splits of the proximity data (Lord and Novick, 1968). The stress-1 formula will be used to quantify the representation error of the multidimensional scaling solutions.

4.2. Comparing stress with Monte Carlo studies

The second dimensionality choice procedure is based on Monte Carlo simulations. The choice for a particular dimensionality is based on a comparison of the empirical stress value with stress profiles of simulated data with known dimensionality and error level (Spence, 1983; Spence and Graef, 1974; Verheyen et al., 2007; Wagenaar and Padmos, 1971). One-thousand stimulus configurations with the same number of stimuli as the target categories will be generated in dimensionality 2 until D by sampling the stimuli’s coordinates from a uniform distribution between 0 and 1000. The Euclidean distances between the stimuli will be error perturbed by multiplying each distance with a random factor, drawn from a normal distribution with a mean of 1.0 and a particular variance (0.00, 0.05, 0.10, 0.15, 0.20, 0.25, or 0.30), corresponding to seven increasing error levels. Each of these simulated data sets will then be subjected to the same multidimensional scaling procedure as the empirical proximities, yielding MDS solutions with 2 to D dimensions. For every combination of the true underlying dimensionality, the error level, and the dimensionality of the solution, the 1000 resulting stress values will be rank ordered and averaged. The final dimensionality will then be determined by the dimensionality choice procedure detailed in the Appendix of Verheyen et al. (2007).

This procedure starts by looking at the stress values of the 2-dimensional representations of the truly 2-dimensional data. The estimated error level of the empirical data is determined by linear interpolation of the error levels whose average stress values border the empirical stress value. The stress value of the 3-dimensional representation of the empirical data is then compared to a criterion value that is arrived at by linear interpolation of the percentile 5 stress values for the error levels that border the estimated error level in the 3-dimensional representations of the truly 2-dimensional data. If the empirical 3-dimensional stress value is higher than the expected (interpolated) fifth percentile stress value, the procedure stops and the dimensionality is assumed to equal 2; if the empirical 3-dimensional stress value is lower than the expected fifth percentile, the procedure is started over by turning to the stress values of the 3-dimensional representations of the truly 3-dimensional data. The procedure terminates when the K-dimensional empirical stress exceeds the fifth percentile of the stress value distribution.
corresponding to the K-dimensional representation of the truly K-1-dimensional data (indicating a dimensionality of K-1) or when the K-dimensional empirical stress is lower than this fifth percentile and K corresponds to D, the maximum number of dimensions one can retain according to the liberal reading of the guideline by Kruskal and Wish (1978). In the latter case, D dimensions are retained.7

4.3. External criterion

The third dimensionality choice procedure is based on information that is external to the input proximities. It is common practice to externally validate MDS spaces by referring to their ability to predict an external criterion (e.g., Dry and Storms, 2009; Hout and Goldinger, 2016; Hout et al., 2013; Koch et al., 2016; Richie et al., 2020). This predictive validity can be employed to determine the dimensionality of MDS spaces, by establishing the number of dimensions that allows the best prediction of the external criterion (Jones and Rosenberg, 1974; Rosenberg and Jones, 1972). Since the prediction pertains to a criterion that is external to the proximities on which the MDS solutions are based, any increases in the prediction of the external variable with increasing dimensionality cannot be due to error fitting because the two data sets are independent.

In line with a long tradition of using semantic spaces to predict external variables such as exemplar generation (e.g., Henley, 1969), category verification time (e.g., Rips et al., 1973), and priming (e.g., Hutchinson and Lockhead, 1977), Verheyen et al. (2007) proposed to determine the dimensionality of semantic spaces based on their ability to predict the rated typicality of the represented category exemplars. Verheyen et al. opted for typicality because it is highly reliable and captures categories’ internal structure very well (Rosch and Mervis, 1975; Rosch, 2002). Following their suggestion, we will choose the dimensionality that provides the maximum correlation between the category instances’ average rated typicality across participants and their Euclidean distance to the category centroid in the MDS configuration. The category centroid will be determined by computing the average of the coordinates of all the category exemplars in the MDS configuration and can be thought of as the category’s prototype (see also Caramazza et al., 1976; Martin and Caramazza, 1980; Reed, 1972; Rips, 1975; Verheyen et al., 2007).

5. Study 1

The similarity data for Study 1 were taken from the first study in Verheyen et al. (2020) in which participants provided similarity data for four conceptual categories. Whether PRaM or SPaM was used for a particular category was manipulated within subjects. The typicality data for Study 1 were taken from De Deyne (2014) in which normative data for photorealistic images of exemplars from nine conceptual categories were provided. The same norm data were used for the selection of stimulus materials for the first study in Verheyen et al. (2020).

5.1. Participants

All participants were undergraduate students at KU Leuven (University of Leuven, Belgium) whose native language was Dutch. Forty-eight participants provided the similarity data. Thirty-seven participants provided the typicality data.

5.2. Materials

The stimuli were 16 photorealistic images of exemplars of four conceptual categories. The categories birds, vehicles, vegetables, and sports, were chosen to be representative of natural categories, artefact categories, natural-artefact categories (i.e., natural products, cultivated by man), and activity categories, respectively. Per category, the 16 exemplars with the highest mean familiarity ratings in De Deyne (2014) were chosen. An overview of the exemplars is provided in Table A1 in the appendix. See Fig. 1 for examples of the stimuli for the category vehicles.

5.3. Procedure

Both the similarity and the typicality task were conducted in Dutch. Each participant provided ratings for all four categories.

In the similarity task, a participant arranged the exemplars of two categories such that the inter-exemplar distances on the screen would be inversely proportional to their perceived similarity, and judged the similarity of all exemplar pairs of the other two categories on a Likert scale ranging from 1 (very dissimilar) to 9 (very similar). Participants alternated between SPaM and PRaM, and categories and methods were counterbalanced, to ensure that an equal number of participants (#24) completed each method-category combination. The starting configuration was similar for both methods: The 16 exemplars were randomly positioned on the screen in a 4 by 4 grid. In SPaM, participants could change the position of the exemplars by dragging them around the screen with the computer mouse. They were free to choose the order in which to change the positions of the exemplars. In PRaM, participants provided their response by pressing a numerical key. The order of pairs to judge was randomized. On every trial, the two exemplars comprising a pair were highlighted with a black border. Both in SPaM and PRaM, all category exemplars remained visible on the screen.

In the typicality task, a participant judged the typicality of the category exemplars on a 7-point Likert scale with higher values indicating higher typicality. The order of the categories and of the exemplars within a category was randomized. All of a category’s exemplars were present on the screen throughout the task.

5.4. Results

Table 1 lists the split-half reliability of the average PRaM and SPaM proximity data for each of the four categories in Study 1. The values confirm previous findings that when the number of PRaM and SPaM participants is equated, PRaM is more reliable than SPaM (Verheyen et al., 2016; Verheyen et al., 2020). Also in line with earlier work (e.g., De Deyne et al., 2008; Hampton and Gardiner, 1983), the split-half reliability of the average typicality judgments for the category exemplars was invariably high: 0.99 for sports, 0.95 for vegetables, 0.97 for vehicles, and 0.98 for birds.

Since each category comprised 16 exemplars, MDS solutions with 2, 3, and 4 dimensions were obtained. For every category, the accompanying stress-1 values are shown using black lines in Fig. 2. The stress values for the PRaM (solid) and SPaM (dotted) data are similar and decrease with increasing dimensionality, indicating the improved fit that higher dimensional MDS solutions afford because of the additional free parameters.

Also shown in Fig. 2 are the MDS solutions’ predictive correlations. Gray lines indicate the correlation between the category exemplars’

| Category | # Pairs | # Participants | PRaM Reliability | SPaM Reliability |
|----------|---------|----------------|------------------|------------------|
| sports   | 120     | 24             | .97              | .89              |
| vegetables | 120     | 24             | .94              | .83              |
| vehicles | 120     | 24             | .98              | .94              |
| birds    | 120     | 24             | .96              | .86              |
average typicality ratings and their Euclidean distance in the MDS solutions to the category centroid, which was calculated by averaging the coordinates of the exemplars in a given dimensionality. As expected, the correlations are negative: exemplars that are more typical are expected to be found closer to the category centroid. The SpAM predictive correlations (dotted gray) are more pronounced than the PRaM predictive correlations (solid gray) for birds, but the reverse holds for vegetables and vehicles. For the category sports, both similarity methods predict typicality to a similar extent. In the case of sports and vehicles, the predictive correlations start to approximate the theoretical maxima, indicated by the average typicality data’s reliability. In the case of vegetables and birds, the predictive correlation falls considerably short of the theoretical maximum.

Table 2 summarizes the outcome of the three dimensionality choice procedures. The procedure that relies on the reliability of the input proximities consistently yields the lowest dimensionality for SpAM (2 dimensions) and the highest dimensionality for PRaM (4 dimensions). This finding underscores that the reliability difference that is found between average SpAM and PRaM proximities is substantial in light of the error components of the MDS spaces that are used to represent them. The procedure that is based on Monte Carlo simulations suggests that the SpAM representations contain an additional dimension compared to the PRaM representations, except for the category of vehicles where the reverse holds. It is also notable that this procedure suggests a dimensionality higher than 2 for three of the four categories, confirming earlier findings that despite its two-dimensional nature SpAM can yield representations with additional dimensions when the data of different individuals are averaged (Hout and Goldinger, 2016; Koch et al., 2016; Richie et al., 2020). The dimensionality choice procedure that is based on the predictive correlation between exemplars’ typicality and their distance to the category centroid in spatial representations of different dimensionality, establishes the same dimensionality for PRaM and SpAM. For two of the four categories more than two dimensions are retained. What is striking is that for only one category-method combination in Table 2, the three choice procedures point to the same dimensionality (2 dimensions in the case of vehicles-SpAM). For the other combinations, there is at least one procedure that indicates a different dimensionality (3 combinations), or the different procedures indicate different dimensionalities altogether (4 combinations). This level of inconsistency between dimensionality choice procedures has been established before (Verheyen et al., 2007).

6. Study 2

The similarity data for Study 2 were taken from the first study in Richie et al. (2020) in which participants provided similarity data for eight conceptual categories. One group of participants used SpAM to provide proximity data for all categories. Another group of participants used PRaM to provide data for a single category. For Study 2 we restrict ourselves to the four categories for which PRaM participants judged the similarity of all exemplar pairs, for comparability with the results from the previous study. The dimensionality choice procedures based on the reliability of the input proximities and on the Monte Carlo simulations can be implemented based on the proximity data from Richie et al. (2020). The predictive correlation dimensionality choice procedure required the collection of typicality ratings.

Study 2 was pre-registered (osf.io/k6yjc). The pre-registration describes the procedure for collecting the typicality data, as well as the dimensionality selection, which was identical to that of Study 1, except that it accounts for the fact that the number of PRaM and SpAM participants in Study 2 differ. Note that per the pre-registration, the proximity data were only requested from Richie et al. once the collection of
the typicality data had been completed.

The purpose of Study 2 is to verify whether the findings from Study 1 can be generalized to another set of materials. The categories from Richie et al. (2020) include a larger number of exemplars that are presented as words rather than pictures. This might result in higher dimensional spatial representations if participants recognize more dimensions of variation because of the wider range of exemplars and/or the more abstract nature of the exemplars.

6.1. Participants

All participants were English speaking. Richie et al. (2020) collected proximity data by either having participants (N = 54) complete SpAM for all categories (within-subjects) or having participants provide pairwise similarity ratings for one randomly assigned category (between-subjects). Among the latter, 33 provided proximities for furniture, 30 for vegetables, 28 for vehicles, and 31 for fruit. The SpAM participants were students of New Mexico State University. The PRaM participants were U.S. workers on Prolific Academic.

To ensure comparability with Richie et al. the typicality data too were provided by participants (N = 40, mean age 30.30 years; 47.50% female, 42.50% male, 10.00% did not identify as male or female) from the U.S. with an approval rating above 80 on Prolific Academic, who were paid at a rate of $10/hour. We decided on N = 40 because the number is comparable to the number of participants in Study 1 (N = 37) and is likely to ensure that the reliability of the average typicality data is large (i.e., the minimum reliability in Study 1 equaled 0.95). This is necessary since the reliability serves as an upper bound for the correlation of the average typicality data with an external criterion. Since the maximum correlation we observed in Study 1 had a magnitude of 0.76, an expected reliability of 0.95 or higher (and therefore a theoretically maximal correlation of 0.95) was deemed sufficient.

6.2. Materials

We restrict ourselves to the four categories in Richie et al. (2020) for which all exemplar pairs were rated in terms of similarity: furniture, vegetables, vehicles, and fruit, with respectively 20, 20, 22, and 21 exemplars. An overview of the exemplars is provided in Table A2 in the appendix. When we obtained the proximity data from Richie et al. (2020), we learned that the SpAM proximities for the items truck and van from the category vehicles were missing. Because of a coding error, they had not been displayed during the collection of SpAM proximities (M. C. Hout, personal communication, October 26, 2019).

6.3. Procedure

Both the similarity and the typicality task were conducted in English. SpAM participants provided proximity data for all four categories. PRaM participants only provided ratings for a single category because PRaM takes considerably longer than SpAM to complete. Each participant in the typicality task provided ratings for all four categories.

In the similarity task, a participant either arranged the exemplars in such a manner that their Euclidean distance on the screen represented the perceived dissimilarity between the exemplars, or judged the similarity of all exemplar pairs on a Likert scale ranging from 1 (not at all similar) to 7 (extremely similar). Besides the range and the labels of the similarity scale, the pairwise rating task used by Richie et al. also differed from that of Study 1 in that a different approach was taken to ensure that participants were aware of the relevant comparison class and (dis)similarity range. Participants were shown the complete list of category exemplars, along with the category name, before the onset of the similarity judgment task. During the task, at most 50 pairs of exemplars were simultaneously visible on the screen (in a random order). The implementation of SpAM by Richie et al. too, was somewhat different to that of Study 1. The category exemplars were presented in a random order to the left and right of a 2100 by 2100 pixels arena in the middle of the screen, delineated by a black border. Participants arranged the exemplars by dragging them into the designated area with the computer mouse. Participants were free to choose the order in which to position the different exemplars.

For the typicality task, we employed the instructions of Rosch and Mervis (1975) with the exception that we reversed the original scale such that higher ratings reflected higher typicality. Participants had a 7-point scale at their disposal and were asked to respond 1 if the instance fit very poorly with their idea or image of the target category; respond 4 if the instance fit moderately well with their idea or image of the target category; respond 7 if the instance is a very good example of their idea or image of the target category; use the other numbers of the 7-point scale to indicate intermediate judgments. The order of the categories and of the exemplars within a category was randomized. All of a category’s exemplars were present on the screen throughout the task.

6.4. Results

The results were obtained according to the pre-registration. This means that we conducted the same analyses as in Study 1, with two exceptions. (i) The average PRaM proximities were subtracted from 8 instead of 10 because a 7-point Likert scale was used instead of a 9-point scale. (ii) In Richie et al. (2020) the number of PRaM participants is smaller than the number of SpAM participants for each of the four categories under investigation. Since the number of participants influences the reliability of the average proximity data, and therefore might affect the dimensionality choice, we equated the PRaM and SpAM proximity data in terms of number of participants. This was done by drawing 100 random samples of SpAM data comprising of n participants each (where n is equal to the number of PRaM participants). We will refer to these findings as SpAM RS, where RS stands for Reduced Subjects. We note one deviation from the pre-registration. As noted above, the items truck and van from the category vehicles were not included in the analyses, because they had not been presented during the collection of the SpAM proximities in Richie et al. (2020). The analyses for the category vehicles thus pertains to 20 exemplars instead of 22.

Table 3 lists the split-half reliability of the average PRaM and SpAM...
Table 3

Reliability of the average PRaM and SpAM RS proximity data for each of the four concepts in Study 2. The SpAM RS reliability corresponds to the average reliability across 100 samples with an equal number of participants as PRaM.

| Category     | # Pairs | # Participants | PRaM | SpAM RS |
|--------------|---------|----------------|------|---------|
| furniture    | 190     | 33             | .97  | .73     |
| vegetables   | 190     | 30             | .93  | .58     |
| vehicles     | 190     | 28             | .96  | .82     |
| fruit        | 210     | 31             | .94  | .58     |

RS proximity data for each of the four categories in Study 2. The SpAM RS reliability corresponds to the average reliability across 100 samples with an equal number of participants as PRaM. As was the case in Study 1, the values indicate that when the number of PRaM and SpAM participants is equated, PRaM is more reliable than SpAM. The split-half reliability of the average typicality judgments for the category exemplars was again invariably high: 0.99 for furniture (20 exemplars), 0.93 for vegetables (20 exemplars), 0.98 for vehicles (22 exemplars), and 0.97 for fruit (21 exemplars). The reliability of the average vehicles typicality data remains at 0.98 when the items truck and van are removed (20 exemplars).

Since each category comprised between 20 and 22 exemplars, MDS solutions with 2, 3, 4, and 5 dimensions were obtained. For every category, the accompanying stress-1 values are shown using black lines in Fig. 3. The SpAM RS values correspond to the median values across 100 samples with an equal number of participants as PRaM. As was the case in Study 1, the stress values for the PRaM (solid) and SpAM RS (dotted) data are similar and decrease with increasing dimensionality. The gray lines in Fig. 3 indicate the correlation between the category exemplars’ average typicality ratings and their Euclidean distance to the category centroid in the different MDS solutions. The SpAM RS values correspond to the median values across 100 samples with an equal number of participants as PRaM. The SpAM RS predictive correlations (dotted gray) are more pronounced than the PRaM predictive correlations (solid gray) for vegetables, but the reverse holds for furniture and fruit. For the category vegetables, both similarity methods predict typicality to a similar extent. The PRaM predictive correlations for furniture are the closest to the average typicality data’s reliability. In the case of vegetables, the predictive correlation of both proximity methods falls considerably short of the theoretical maximum; in the case of furniture and fruit, this is particularly true for SpAM.

Table 4 summarizes the outcome of the three dimensionality choice procedures. The values for SpAM RS correspond to the mode of the dimensionality choice procedure across 100 samples with the same number of participants as PRaM. As was the case in Study 1, the procedure that relies on the reliability of the input proximities consistently yields the lowest dimensionality for SpAM RS (2 dimensions) and the highest dimensionality for PRaM (5 dimensions). Contrary to Study 1, in which the Monte Carlo simulations suggested a higher dimensionality for SpAM than for PRaM, the procedure now indicates the same dimensionality for PRaM and SpAM RS. Whereas the predictive correlation in Study 1 yielded identical PRaM and SpAM dimensionalities for each of the four concepts, in Study 2 this result is only obtained for two of the four concepts. For furniture and fruit, the predictive correlation suggests a considerable higher dimensionality for SpAM RS than for PRaM (5 dimensions compared to 3 or 2 dimensions). Do note that the low-dimensional PRaM representations for these categories achieve a much higher predictive correlation than the five-dimensional SpAM RS representations do (see Fig. 3).

Taken together, the outcome of the different dimensionality procedures paints a picture that is similar to that of Study 1: whether PRaM and SpAM yield comparable dimensionalities depends on the dimensionality choice procedure, and the final number of dimensions that is retained differs considerably between dimensionality choice procedures. Both for PRaM and SpAM RS, there were two categories for which at least one procedure indicated a different dimensionality, and two categories for which the different procedures indicated different dimensionalities altogether. In line with Study 1, as well as earlier findings, the results do suggest that despite its two-dimensional nature, it is possible to use SpAM to obtain MDS representations with more dimensions by aggregating the spatial arrangements of several participants. Finally, the results of Study 2 - particularly those concerning the predictive correlation - indicate that in order to successfully capture the internal structure of semantic categories, their spatial representations might have to include more than the two or three dimensions that are commonly opted for in the literature (see also Richie et al., 2020; Verheyen et al., 2007).

A final question that remains is whether these conclusions depend on the equation of the SpAM and PRaM data in terms of the number of participants. The SpAM RS data are based on approximately half of the SpAM data that are available from Richie et al. (2020). As such, the analyses do not take into account the fact that one can collect much more SpAM data in the time it takes participants to complete PRaM. We therefore conducted an exploratory analysis in which we applied the dimensionality choice procedures to the average SpAM data across all 54 participants. It yielded identical dimensionalities, except for vegetables where the Monte Carlo procedure yielded 3 instead of 2 dimensions, and fruit where the predictive correlation procedure yielded 4 instead of 5 dimensions. Although including all SpAM participants increased the reliability of the average SpAM data to 0.82 for furniture, 0.72 for vegetables, 0.90 for vehicles, and 0.71 for fruit, the dimensionality choice procedure based on reliability still yielded 2 dimensions for all four categories. Basing the dimensionality choice on the data of almost twice as many participants thus did not yield consistently higher dimensionalities for SpAM. This finding resonates with the suggestion in Verheyen et al. (2016) that the inclusion of additional SpAM participants will not increase dimensionality if these participants only choose to represent two of the most salient dimensions of their mental representation of the category in their two-dimensional spatial arrangement, since these dimensions are also likely to be communicated by earlier participants. For less salient dimensions to be uncovered, a sufficient number of participants needs to include them in their spatial arrangements, but seeing that participants are likely to go with two salient dimensions if two dimensions is all they can communicate on a single SpAM trial, it might require sampling a very large number of participants to encounter a few “odd” ones who decide to convey a not so salient dimension of variation.

7. Study 3

The proximity data for Study 3 too were taken from the first study in Richie et al. (2020). In Study 3, we use the data for the four categories from their study for which PRaM participants judged only a random half of the exemplar pairs because obtaining all pairwise judgments would have taken a prohibitive amount of time. SpAM participants did provide proximity data for all category exemplars.

Study 3 was conducted at the request of an anonymous reviewer. It followed the procedure that was outlined in the pre-registration of Study 2 with the exception that to equate the number of PRaM and SpAM participants, samples now had to be drawn from the PRaM instead of the SpAM data since the number of PRaM participants outnumbered the number of SpAM participants, instead of the other way around as in Study 2. It also needs to be noted that at the time we collected the typicality ratings that are required to implement the predictive correlation dimensionality choice procedure for Study 3, we already had access to the proximity data from Richie et al. (2020) from our original request.

9 We thank an anonymous reviewer for raising this question.
The purpose of Study 3 is to assess the impact on dimensionality of applying PRaM to a subset of the exemplar pairs when the number of category instances prohibits the collection of all pairwise judgments. It constitutes a case where SpAM might have an advantage over PRaM since SpAM does allow one to collect all proximity data for a considerable number of category exemplars in a single sitting. This advantage should play out particularly when the number of PRaM and SpAM participants is once more equated, since the average PRaM proximity data will then only be based on half of the data points compared to SpAM.

7.1. Participants

Richie et al. (2020) had 54 participants complete SpAM for all categories (within-subjects) and 243 participants provide pairwise similarity ratings for one randomly assigned category (between-subjects). Sixty-seven participants provided similarity ratings for professions, 61 for sports, 61 for clothing, and 54 for birds. As for Study 2, we had 40 U.S. Prolific Academic workers with an approval rating above 80% provide typicality ratings. They were compensated at a rate of $10/hour. Eighteen participants identified as female; 22 as male. Their mean age was 34.90 years.

7.2. Materials

This study pertains to the four categories in Richie et al. (2020) for which participants rated a random half of all exemplar pairs in terms of similarity: professions, sports, clothing, and birds, with respectively 28, 28, 29, and 30 exemplars. The decision to have participants rate only half of the exemplar pairs was taken because 28 exemplars combine in 378 pairs, which Richie et al. judged to take a prohibitive amount of time to rate. An overview of the exemplars is provided in Table A3 in the appendix.

Table 4: Withheld dimensionality for PRaM and SpAM RS proximity data according to three dimensionality choice procedures for each of the four concepts in Study 2.

| Category       | # Stimuli | # Participants | Dimensionality Choice Procedure |
|----------------|-----------|----------------|---------------------------------|
|                |           |                | 1.0 Minus Reliability          |
|                |           |                | Monte Carlo                     |
|                |           |                | Predictive Correlation          |
|                | PRaM      | SpAM RS        | PRaM                           |
|                | SpAM RS   | PRaM RS        | SpAM RS                        |
|                | PRaM      | SpAM RS        | PRaM                           |
|                | SpAM RS   | PRaM RS        | SpAM RS                        |
| furniture      | 20        | 33             | 5                              |
| vegetables     | 20        | 30             | 5                              |
| vehicles       | 20        | 28             | 5                              |
| fruit          | 21        | 31             | 5                              |

Note: The reported dimensionalities for SpAM RS correspond to the dimensionality mode across 100 samples with an equal number of participants as PRaM.

7.3. Procedure

The procedures of the typicality rating task, SpAM, and PRaM were identical to those in Study 2, with the exception that PRaM participants only rated a random half of a category’s exemplar pairs in terms of...
similarity compared to all pairwise combinations in Study 2.

7.4. Results

We conducted the same analyses as in Study 2, with the exception that we equated the PRaM and SpAM proximity data in terms of number of participants by drawing 100 random samples of PRaM data instead of SpAM data because the number of PRaM participants was larger than the number of SpAM participants for three categories under investigation (professions, sports, clothing). We will refer to these as PRaM RS, where RS stands for Reduced Subjects. It was not necessary to undertake this procedure for the birds category since the number of PRaM and SpAM participants was the same for this category (N = 54).

Table 5 lists the split-half reliability of the average PRaM RS and SpAM proximity data for each of the four categories in Study 3. The PRaM RS reliability corresponds to the average reliability across 100 samples with an equal number of participants as SpAM, except for birds, for which the number of PRaM and SpAM participants was the same. As was the case in the previous studies, the values indicate that when the number of PRaM and SpAM participants is equated, PRaM is more reliable than SpAM. This result obtains despite the fact that the average PRaM data are only based on half of the data points the average SpAM data are based on. The split-half reliability of the average typicality judgments for the category exemplars was again invariably high: 0.94 for professions (28 exemplars), 0.97 for sports (28 exemplars), 0.98 for clothing (29 exemplars), and 0.91 for birds (30 exemplars).

Since each category comprised between 28 and 30 exemplars, MDS solutions with 2, 3, 4, 5, 6, and 7 dimensions were obtained. For every category, the accompanying stress-1 values are shown using black lines in Fig. 4. The PRaM RS values correspond to the median values across 100 samples with an equal number of participants as SpAM, except for birds for which sampling was not required because the number of PRaM and SpAM participants was the same. The stress values for the PRaM RS (solid) and SpAM (dotted) data are again similar and decrease with increasing dimensionality. The gray lines in Fig. 4 indicate the correlation between the category exemplars’ average typicality ratings and their Euclidean distance to the category centroid in the different MDS solutions. The PRaM RS values correspond to the median values across 100 samples with an equal number of participants as SpAM, except for the category of birds. The PRaM RS (solid gray) and SpAM (dotted gray) predictive correlations are comparable in all categories except birds, where the PRaM predictive correlations are more pronounced than the SpAM predictive correlations. The correlations do fall short of the theoretical maxima indicated by the reliability of the typicality data. This is particularly noticeable for the category of professions. The latter result might be due to a restriction of range, however, seeing that the various professions were uniformly judged to be typical.

Table 6 summarizes the outcome of the three dimensionality choice procedures. The values for PRaM RS correspond to the mode of the dimensionality choice procedure across 100 samples with the same number of participants as SpAM, except for birds for which sampling was not required to equate the number of participants. As was the case for the previous studies, the final number of dimensions that is retained does not affect the resulting reliabilities much either. The new values of 0.95 for professions, 0.93 for sports, and 0.95 for clothing are close to the average values reported for PRaM RS in Table 5, which were already quite high.

8. General discussion

In this paper, we investigated whether the similarity measurement procedure affects the dimensionality of the resulting multidimensional scaling (MDS) representation. In three studies, we compared the dimensionality of spatial representations of semantic categories for which the underlying (dis)similarity data had been obtained either with the Pairwise Rating Method (PRaM) or the Spatial Arrangement Method (SpAM). The results were similar, irrespective of the number and the presentation format (verbal or pictorial) of the category exemplars: A systematic dimensionality difference between PRaM and SpAM was only found when the dimensionality choice procedure was based on the reliability of the input (dis)similarities, not when it was based on Monte Carlo simulations or predictive correlations. In the former case, PRaM was consistently found to yield higher-dimensional spatial representations of the semantic categories than SpAM. This is a direct consequence of the recurring observation that the reliability of PRaM is higher than relatively large number of SpAM participants and the decreased number of data points on which the PRaM data are based (due to participants only rating half the pairs), which respectively increase and decrease the reliability of the average proximity data. As such, the dimensionality choice procedure based on reliability appears to be the only one that is affected by the choice to have PRaM participants only rate a subset of a category’s exemplar pairs.10 In response to the increased number of instances per category, the other dimensionality choice procedures tend to indicate a higher dimensionality in Study 3 than in Study 2. Although it stands to reason that this reflects the increased variability the additional exemplars bring, it cannot be ruled out that this difference is due to other (potentially richer) categories being included in Study 3 compared to Study 2. In line with Study 1 and Study 2, the results of Study 3 indicate that it is possible to use SpAM to obtain MDS representations with more than two dimensions by aggregating the spatial arrangements of several participants, and that the inclusion of more than two dimensions is generally warranted to adequately account for the internal structure of semantic categories.

Note: PRaM participants only judged half of a category’s exemplar pairs. *This value is not the result of sampling because the number of PRaM and SpAM participants for birds is the same.

Table 5

| Category | # Pairs | # Participants | PRaM RS | SpAM |
|----------|---------|----------------|---------|------|
| Professions | 378 | 54 | .94 | .84 |
| sports | 378 | 54 | .93 | .79 |
| clothing | 406 | 54 | .95 | .84 |
| birds | 435 | 54 | .88* | .72 |

10 Repeating the dimensionality choice procedure on the entirety of the PRaM data (as opposed to on samples matched to the SpAM data in terms of the number of participants) did not consistently increase the withheld dimensionality. Only the predictive correlation procedure for sports yielded 7 instead of 6 dimensions. The Monte Carlo procedure established the dimensionality of the complete professions data at 4. This outcome is perhaps not surprising seeing that the increase in the number of participants is only modest, from 54 to 61 in the case of sports and clothing, and from 54 to 67 in the case of professions. It does not affect the resulting reliabilities much either. The new values of 0.95 for professions, 0.93 for sports, and 0.95 for clothing are close to the average values reported for PRaM RS in Table 5, which were already quite high.
Because the average PRaM data are more precise than the average SpAM data, they afford a more complex (that is: higher dimensional) interpretation (Kruskal, 1964; Borg and Groenen, 2005). Proponents of SpAM will rightfully point out that the reliability difference between the two similarity measurement methods is rather easy to overcome since SpAM takes less time to complete (Hout and Goldinger, 2016; Hout et al., 2013; Verheyen et al., 2020) and is therefore arguably easier to recruit additional participants for. Using the Spearman-Brown prediction formula (Brown, 1910; Spearman, 1910), Verheyen et al. (2020) estimated that for the 16 exemplar stimulus sets from Study 1, at least 50 SpAM participants would be required to obtain a reliability that is comparable to the PRaM reliability obtained with 24 participants. Note that it will not always come down to a doubling of the sample size. The exact multiplication factor will depend both on the number and the nature of the stimuli (Verheyen et al., 2020), and the dimensionality of the stimuli might be a factor that substantially increases the required number of SpAM participants to match the PRaM reliability. For instance, Verheyen et al. (2016) showed that for 25 2-dimensional perceptual stimuli the multiplication factor was about 1, compared to a factor of about 3 for 27 3-dimensional perceptual stimuli. Similarly, in Study 2 we showed that doubling the number of SpAM participants increased the reliability of the average SpAM data, but not to the extent that the dimensionality choice procedure based on reliability indicated that more than the minimum number of two dimensions should be retained.

The lower reliability of SpAM has been attributed to individual differences in the dimensions participants choose to convey when they can only communicate two from a potentially larger set, participants sub-optimally conveying additional dimensions in the two-dimensional spatial arrangement, participants not fully realizing they are simultaneously changing the distances to all the other stimuli when moving a single stimulus, and participants alternatively interpreting SpAM as a discrete sorting task rendering within-cluster distances meaningless (Hout et al., 2016; Verheyen et al., 2016).

Table 6

Withheld dimensionality for PRaM RS and SpAM proximity data according to three dimensionality choice procedures for each of the four concepts in Study 3.

| Category     | # Stimuli | # Participants | Dimensionality Choice Procedure |
|--------------|-----------|----------------|---------------------------------|
|              |           |                | 1.0 Minus Reliability           | Monte Carlo                   | Predictive Correlation |
|              |           |                | PRaM RS | SpAM | PRaM RS | SpAM | PRaM RS | SpAM |
| professions  | 28        | 54             | 7       | 3    | 3/4     | 5    | 2        | 7    |
| sports       | 28        | 54             | 6       | 2    | 3       | 3    | 6        | 7    |
| clothing     | 29        | 54             | 7       | 3    | 4       | 5    | 4        | 2    |
| birds        | 30        | 54             | 4*      | 2    | 5*      | 3    | 7*       | 7    |

Note: PRaM participants only judged half of a category’s exemplar pairs. The reported dimensionality for PRaM RS correspond to the dimensionality mode across 100 samples with an equal number of participants as SpAM. The Monte Carlo procedure indicated that half of the PRaM RS samples for professions should be represented in three dimensions and half in four dimensions. *These values are not the result of sampling because the number of PRaM and SpAM participants for birds is the same.

The lower reliability of SpAM has been attributed to individual differences in the dimensions participants choose to convey when they can only communicate two from a potentially larger set, participants sub-optimally conveying additional dimensions in the two-dimensional spatial arrangement, participants not fully realizing they are simultaneously changing the distances to all the other stimuli when moving a single stimulus, and participants alternatively interpreting SpAM as a discrete sorting task rendering within-cluster distances meaningless (Hout et al., 2016; Verheyen et al., 2016).
Methods in Psychology 5 (2021) 100060

13

S. Verheyen and G. Storms

well to PRaM-derived distances as to SpAM-derived distances (Ramsay, model that is assumed in the Monte Carlo simulations applies equally procedures are equally appropriate for the two similarity measurement -affect SpAM

Wagenaar and Padmos, 1971). Similarly, whether the category prototype is appropriately characterized as a centroid, might depend on the distributional characteristics of the proximity data, which are known to differ between PRaM and SpAM (Verheyen et al., 2016; Verheyen et al., 2020). Setting these caveats aside for a moment, the doubling of the number of SpAM participants in Study 2 showed that having more participants provide spatial arrangements of semantic categories, does not guarantee that the resulting spatial representations will include more dimensions, at least not when the three dimensionality choice procedures from this paper are used.

Seeing that (i) the dimensionality difference between PRaM and SpAM that results from the reliability difference of the input proximities can be fairly easily resolved through the collection of additional SpAM data, and that (ii) with comparable numbers of participants, no systematic differences in dimensionality between PRaM and SpAM data sets were found based on Monte Carlo simulations or predictive correlations, the results from the current studies alleviate the concern that because of its two-dimensional nature, SpAM might underestimate the dimensionality of high-dimensional stimuli compared to PRaM (Verheyen et al., 2016). Through the aggregation of two-dimensional spatial arrangements from different participants, one can obtain spatial representations of semantic categories with more than two dimensions. It appears that the relative importance or salience of the dimensions constituting the high-dimensional space that these participants presumably share, is reflected in the relative number of participants who convey these dimensions in their two-dimensional SpAM configurations. As counterintuitive as it may appear, it thus seems that the pronounced individual differences in similarity judgments that sometimes make average similarity data unrepresentative for the individual data (Ashby et al., 1994; Lee and Pope, 2003; Summers and MacKay, 1976; Verheyen et al., 2016) are the very means that help aggregate SpAM data attain more than two dimensions.

When one is interested in obtaining multidimensional scaling representations at the aggregate level, there thus appears to be no reason to favour PRaM over SpAM or vice versa, since for twelve semantic categories that were deliberately chosen because of their large number of dimensions with varying salience, no systematic dimensionality differences arose between the two similarity measurement methods. The results of Study 3 show that when the number of category exemplars becomes too large to have participants rate all pairs in terms of similarity, one can suffice with presenting participants with a random subset of pairs to rate, without detrimental consequences for the dimensionality of the resulting spatial representation (when the dimensionality of the corresponding SpAM representations are taken as a reference, that is). The current results demonstrate this in categories of approximately 30 exemplars for which half the exemplar pairs were rated. It needs to be shown whether this extrapolates to larger categories for which it might be necessary to present a much smaller proportion of exemplar pairs. When the number of category exemplars increases beyond 30, it might also be necessary to make changes to SpAM, since it will become increasingly difficult for participants to generate a spatial arrangement that provides a satisfactory account of all the inter-exemplar relations, up till the point that such an arrangement might no longer fit on a single screen. Having participants spatially arrange subsets of exemplars on consecutive trials might then be a viable solution (Goldstone, 1994; Kriegeskorte and Mur, 2012; Verheyen et al., 2020).

Using objective dimensionality choice procedures, the spatial representations of several of these semantic categories in Studies 1, 2, and 3 were found to include more dimensions than what is commonly found in the concepts and categories literature (where the dimensionality choice is largely determined by subjective criteria; see Verheyen et al., 2007). However, there is no way of establishing what the true dimensionality of these spatial representations is and whether PRaM and SpAM might not both underestimate it. There is no task-independent manner of determining the true dimensionality of a set of stimuli, unless one would have the stimuli vary by design along a fixed number of dimensions. If the dimensionnalities retained across similarity measurement methods and/or dimensionality choice procedures would be the same, this would provide converging evidence for a particular number of psychological dimensions. Unfortunately, the results across both similarity methods and dimensionality choice procedures were (wildly) inconsistent (see also below). Perhaps both SpAM and PRaM might be constrained in the number of dimensions that they allow participants to communicate (see also Nosofsky et al., 2020). At the moment, the only way we foresee to test this possibility is to conduct a study with artificial stimuli, designed to vary on a large number of salient dimensions. Subjecting PRaM and SpAM similarity data for these stimuli to individual differences scaling (as Verheyen et al., 2016, did for the three-dimensional stimuli from Hout et al., 2013) would probably constitute the most straightforward way of assessing the limits to the number of dimensions that can be conveyed by a single participant using PRaM or SpAM.

Perhaps the most striking as well as sobering finding of this paper are the inconsistent results the different dimensionality choice procedures yielded. This inconsistency was found within similarity measurement methods (and are thus not the results of different data characteristics) and occurred despite the fact that the dimensionality choice procedures were chosen because of their objective nature. The issue of variable dimensionality choice outcomes is probably even more pronounced than this paper suggests, since authors in the concepts and categories literature tend to rely on more subjective criteria to determine the number of dimensions of categories’ spatial representations. According to Verheyen et al. (2007), the majority of MDS practitioners in this field rely on considerations of parsimony and/or interpretability to settle on a dimensionality. Alternatively, they rely on verbal characterizations of absolute stress values (following Kruskal, 1964). These too may lead to widely varying dimensionality choices since stress values are dependent upon the number of dimensions and the number of stimuli, and different MDS users might consider different verbal characterizations satisfactory. Although we do not have data concerning the dimensionality choice procedures that are used in other fields that rely on MDS, it is to be expected that the situation is not much different given the variability in the outcome of established, objective dimensionality choice procedures. This is worrisome, because without clear guidelines or useable procedures, MDS users are likely to fall back on a solution that suits their purposes. Researcher degrees of freedom like these are generally...
considered to have contributed considerably to psychology’s replication crisis (Simmons et al., 2011).

Since our original observation of the inconsistency between MDS dimensionality choice procedures over 10 years ago Verheyen et al. (2007), there have been a number of developments. Both Bayesian (Gronau and Lee, 2020; Lee, 2001; Oh and Raftery, 2001) and cross-validation approaches (Richie and Verheyen, 2020; Richie et al., 2020; Steyvers, 2006) to dimensionality selection have been proposed. These proposals share a number of characteristics with the maximum likelihood models that were proposed in the seventies and eighties and allowed the dimensionality to be determined using likelihood ratio tests (e.g., Ramsay, 1977; Takane and Carroll, 1982), but were not used by any MDS practitioner in the concepts and categories literature up till 2007 (Verheyen et al., 2007). They make a number of assumptions that limit their applicability, are difficult to understand for casual MDS users, and/or are not available in popular statistical software, and therefore may succumb to the same fate as their maximum likelihood predecessors. Based on the literature review by Verheyen et al., a dimensionality choice procedure is most likely to be adopted when it is conceptually easy and practically implemented. Because of its simplicity, cross-validation is perhaps the most promising approach to dimensionality selection in this respect, but because the method remains absent in the empirical literature (with the exception of Richie et al., 2020) little is known about its effectiveness. Before it is implemented in commonly used statistical software, extensive research will be needed to establish the conditions under which it is able to recover the true dimensionality of simulated data sets. Note that these alternative dimensionality selection techniques will not necessarily do away with the issues raised in this paper. For instance, both the Bayesian approach of Lee (2001) and the cross-validation approach of Richie and Verheyen (2020) are informed by the precision of the proximity measures across participants.

Open practices statement

The data that support the findings of this study are openly available on the Open Science Framework at https://osf.io/yqpd4/?view_only=8964bfa145f549e8bd060c22a9716ed2 (https://osf.io/wazqy/). Study 2 was pre-registered (https://osf.io/t6iyjc). The data and materials used in this article are licensed under a Creative Commons Attribution 4.0 International License (CC-BY), which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original authors and the source, provide a link to the Creative Commons license, and indicate if changes were made. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Author note

SV and GS conceptualized the study. SV gathered and analyzed the data, and drafted the manuscript. GS provided critical revisions. Both authors discussed the findings thoroughly, read, and approved the final version of the manuscript. We would like to thank Britt Wuyts for her help with data collection, and Simon De Deyne and Russell Richie for sharing their data.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This work was supported by the KU Leuven Research Council under Grant C14/16032. The funders of this research did not have a role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. SV is temporarily employed by Erasmus University Rotterdam. His evaluation, further employment, and ability to obtain research funding are partly dependent on the number of articles he publishes. In previous work, SV and GS have raised concerns regarding the ability of SpAM to yield high-dimensional spatial representations.

Appendix

Table A1
Overview of the exemplars per category in Study 1, in decreasing order of typicality. Exemplar names were translated from Dutch, the language in which the study was conducted.

| exemplar id | sports     | vegetables | vehicles | birds       |
|-------------|------------|------------|----------|-------------|
| 1           | soccer     | lettuce    | car      | blackbird   |
| 2           | running    | carrot     | bus      | sparrow     |
| 3           | swimming   | cauliflower| van      | maggie      |
| 4           | cycling    | tomato     | train    | crow        |
| 5           | basketball | green beans| metro    | tit         |
| 6           | tennis     | broccoli   | jeep     | pigeon      |
| 7           | volleyball | spinach    | tram     | gull        |
| 8           | long jump  | cucumber   | taxi     | stork       |
| 9           | judo       | choiory    | truck    | peheasant   |
| 10          | badminton  | bell pepper| bicycle  | duck        |
| 11          | squash     | zucchini   | dirt bike| peacock     |
| 12          | horse riding| parsley   | motorcycle| swan        |
| 13          | table tennis| onion     | plane    | rooster     |
| 14          | hiking     | potato     | helicopter| chicken     |
| 15          | billiards  | mushroom   | tractor  | ostrich     |
| 16          | chess      | corn       | air balloon| penguin     |

Note that these alternative dimensionality selection techniques will not necessarily do away with the issues raised in this paper. For instance, both the Bayesian approach of Lee (2001) and the cross-validation approach of Richie and Verheyen (2020) are informed by the precision of the proximity measures across participants.
### Table A2
Overview of the exemplars per category in Study 2, in decreasing order of typicality.

| exemplar id | furniture | vegetables | vehicles | fruit |
|-------------|-----------|------------|----------|-------|
| 1           | sofa      | lettuce    | car      | strawberry |
| 2           | bed       | carrot     | truck\(^1\) | apple |
| 3           | chair     | broccoli   | van\(^1\) | banana |
| 4           | dresser   | asparagus  | jeep     | peach |
| 5           | table     | celery     | bus      | grapes |
| 6           | desk      | cabbage    | motorcycle | blueberry |
| 7           | bookcase  | spinach    | train    | pear |
| 8           | cabinet   | cauliflower | airplane  | cherry |
| 9           | bench     | radish     | tractor  | pineapple |
| 10          | footstool | corn       | helicopter | plum |
| 11          | lamp      | cucumber   | boat     | tangerine |
| 12          | cushion   | beet       | trolley  | watermelon |
| 13          | mirror    | turnip     | trailer  | apricot |
| 14          | rug       | onion      | hovercraft | melon |
| 15          | vase      | eggplant   | bicycle | grapefruit |
| 16          | piano     | squash     | submarine | honeydew |
| 17          | stove     | artichoke  | carriage | pomegranate |
| 18          | clock     | potato     | scooter  | lemon |
| 19          | radio     | tomato     | rocket   | coconut |
| 20          | telephone | beans      | cart     | date |
| 21          |          |            | skateboad | olive |
| 22          |          |            | sled     |       |

\(^1\) Because of a coding error, the items truck and van were not displayed during the collection of SpAM proximities in Richie et al. (2020) and therefore not included in our analyses.

### Table A3
Overview of the exemplars per category in Study 3, in decreasing order of typicality.

| exemplar id | professions | sports | clothing | birds |
|-------------|-------------|--------|----------|-------|
| 1           | doctor      | basketball | jeans   | woodpecker |
| 2           | lawyer      | baseball | dress   | crow |
| 3           | dentist     | soccer  | pants   | pigeon |
| 4           | psychologist | hockey | shorts | sparrow |
| 5           | accountant  | tennis  | sweatshirt | parrot |
| 6           | veterinarian | volleyball | sweater | dove |
| 7           | teacher     | rugby   | blouse  | eagle |
| 8           | judge       | boxing  | skirt   | robin |
| 9           | pilot       | swimming | jacket | seagull |
| 10          | architect   | golfing | suit    | canary |
| 11          | fireman     | badminton | overalls | falcon |
| 12          | pharmacist  | gymnastics | tracksuit | blackbird |
| 13          | educator    | cycling  | tuxedo  | parakeet |
| 14          | carpenter   | rowing  | coat    | swallow |
| 15          | policeman   | skiing  | pajamas | vulture |
| 16          | plumber     | archery  | gown    | owl |
| 17          | butcher     | running  | panties | swan |
| 18          | physiotherapist | fencing | boxers | duck |
| 19          | actor       | handball | socks   | pelican |
| 20          | postman     | surfing  | bra     | heron |
| 21          | baker       | squash  | boots   | magpie |
| 22          | minister    | judo    | scarf   | cuckoo |
| 23          | cook        | sailing | hat     | peacock |
| 24          | manager     | ballet  | sneakers | pheasant |
| 25          | secretary   | fishing | tie     | rooster |
| 26          | stewardess  | billiards | gloves | chicken |
| 27          | clerk       | chess   | beanie  | stork |
| 28          | cashier     | walking | mittens | turkey |
| 29          |            |         | belt    | ostrich |
| 30          |            |         |         | penguin |

### References

Ameel, E., Storms, G., 2006. From prototypes to caricatures: geometrical models for concept typicality. J. Mem. Lang. 55 (3), 402-421. [https://doi.org/10.1016/j.jml.2006.05.005](https://doi.org/10.1016/j.jml.2006.05.005).

Ashby, F.G., Maddox, W.T., Lee, M.D., 1994. On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. Psychol. Sci. 5 (3), 144-151. [https://doi.org/10.1111/j.1467-9280.1994.tb00651.x](https://doi.org/10.1111/j.1467-9280.1994.tb00651.x).

Barsalou, L.W., 1990. On the indistinguishability of exemplar memory and abstraction in category representation. In: Srull, T.K., Wyer, R.S. (Eds.), Advances in Social Cognition: Content and Process Specificity in the Effects of Prior Experiences, vol. 3. Erlbaum, Hillsdale, NJ, pp. 61-88.

Boot, I., Pecher, D., 2010. Similarity is closeness: metaphorical mapping in a conceptual task. Q. J. Exp. Psychol. 63 (5), 942-954. [https://doi.org/10.1080/17470210903134351](https://doi.org/10.1080/17470210903134351).

Borg, I., Groenen, P.J.F., 2005. Modern Multidimensional Scaling. Springer, New York, NY. [https://doi.org/10.1007/0-387-28981-x](https://doi.org/10.1007/0-387-28981-x).
S. Verheyen and G. Storms

Methods in Psychology 5 (2021) 100060

Rosch, E.H., 2002. Principles of categorization. In: Levitin, D.J. (Ed.), Foundations of Cognitive Psychology: Core Readings. MIT Press, Cambridge, MA, pp. 251–270. https://doi.org/10.7551/mitpress/2068.003.0017.

Rosch, E.H., Mervis, C.B., 1975. Family resemblances: studies in internal structures of categories. Cognit. Psychol. 7 (4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9.

Rosenberg, S., Jones, R., 1972. A method for investigating and representing a person’s implicit theory of personality: Theodore Dreiser’s view of people. J. Pers. Soc. Psychol. 22 (3), 372–386. https://doi.org/10.1037/0022-3514.22.3.372.

Sattath, S., Tversky, A., 1977. Additive similarity trees. Psychometrika 42, 319–345. https://doi.org/10.1007/BF02293654.

Shepard, R.N., Arabie, P., 1979. Additive clustering: representation of similarities as combinations of discrete overlapping properties. Psychol. Rev. 86 (2), 87–123. https://doi.org/10.1037/0033-295X.86.2.87.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22 (11), 1359–1366. https://doi.org/10.1177/0956797611417632.

Smits, T., Storms, G., Rosseel, Y., De Boeck, P., 2002. Fruits and vegetables categorized: an application of the generalized context model. Psychometric Bull. Rev. 9 (4), 836–844. https://doi.org/10.3758/BF03196343.

Spence, I., Ogilvie, J.C., 1973. A table of expected stress values for random rankings in nonmetric multidimensional scaling. Multivariate Behav. Res. 8, 511–527. https://doi.org/10.1207/s15327906mbr0804_8.

Spence, I., Graef, J., 1974. The determination of the underlying dimensionality of an empirically obtained matrix of proximities. Multivariate Behav. Res. 9 (3), 331–341. https://doi.org/10.1207/s15327906mbr0903_8.

Spence, I., 1983. Monte Carlo simulation studies. Appl. Psychol. Meas. 7 (4), 405–438. https://doi.org/10.1177/014662168300700403.

Storms, G., 1995. On the robustness of maximumlikelihood scaling for violations of the error model. Psychometrika 60 (2), 247–258. https://doi.org/10.1007/bf02294115.

Summers, J.O., MacKay, D.B., 1976. On the validity and reliability of direct similarity judgments. J. Market. Res. 13 (3), 289–295. https://doi.org/10.1177/002224377601300311.

Takane, Y., 2007. Applications of multidimensional scaling in psychometrics. In: Rao, C. R., Sinharay, S. (Eds.), Handbook of Statistics, Psychometrics, vol. 26. Elsevier, Amsterdam, The Netherlands, pp. 359–400. https://doi.org/10.1016/S0168-7414(07)36142-5.

Takane, Y., Carroll, J.D., 1982. Analysis of individual differences in multidimensional scaling via an N-way generalization of MCA. Psychometrika 47, 645–674. https://doi.org/10.1007/bf02293654.

Takane, Y., Young, F., De Leeuw, J., 1977. Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. Psychometrika 42 (1), 7–67. https://doi.org/10.1007/bf02293745.

Tenenbaum, J.B., 1996. Learning the structure of similarity. In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (Eds.), Neural Information Processing Systems. MIT Press, Cambridge, MA, pp. 3–39.

Tolman, E.C., 1948. Cognitive maps in rats and men. Psychol. Rev. 55 (4), 189–208. https://doi.org/10.1037/h0061626.

Torgerson, W.S., 1958. Theory and Methods of Scaling. John Wiley & Sons, New York, NY.

Tversky, A., 1977. Features of similarity. Psychol. Rev. 84 (4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327.

Tversky, A., Hutchinson, W., 1986. Nearest neighbor analysis of psychological spaces. Psychol. Rev. 83 (3), 375–401. https://doi.org/10.1037/0033-295X.83.3.375.

Tsogo, L., Masson, M., Bardot, A., 2000. Multidimensional scaling methods for many object-sets: a review. Multivariate Behav. Res. 35 (3), 307–319. https://doi.org/10.1207/s15327906mbvr3503_3.

Tsogo, L., Masson, M., Bardot, A., 2000. Multidimensional scaling methods for many object-sets: a review. Multivariate Behav. Res. 35 (3), 307–319. https://doi.org/10.1207/s15327906mbvr3503_3.

Wagenaar, W.A., Padmos, P., 1971. Quantitative interpretation of stress in Kruskal’s multidimensional scaling: an alternating least squares method with optimal scaling features. Psychometrika 42 (1), 7–67. https://doi.org/10.1007/bf02293745.

Wagenar, W.A., Padmos, P., 1971. Quantitative interpretation of stress in Kruskal’s multidimensional scaling technique. J. Math. Stat. Psychol. 24 (1), 101–110. https://doi.org/10.1177/002225187102400129.

Winter, B., Matlock, T., 2013. Making judgments based on similarity and proximity. Metaphor Symbol 28 (4), 219–232. https://doi.org/10.1080/10926488.2013.826529.