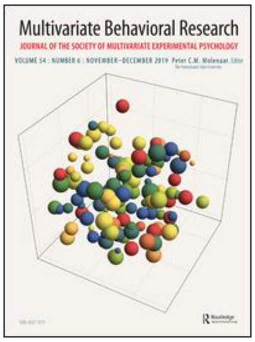

Routledge
Taylor & Francis Group

Multivariate Behavioral Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

A Comparison of the Spatial Arrangement Method and the Total-Set Pairwise Rating Method for Obtaining Similarity Data in the Conceptual Domain

Steven Verheyen , Anne White & Gert Storms

To cite this article: Steven Verheyen , Anne White & Gert Storms (2020): A Comparison of the Spatial Arrangement Method and the Total-Set Pairwise Rating Method for Obtaining Similarity Data in the Conceptual Domain, Multivariate Behavioral Research

To link to this article: https://doi.org/10.1080/00273171.2020.1857216

Published online: 17 Dec 2020.

MULTIVARIATE BEHAVIORAL RESEARCH
https://doi.org/10.1080/00273171.2020.1857216

Routledge
Taylor & Francis Group

OPEN ACCESS

A Comparison of the Spatial Arrangement Method and the Total-Set Pairwise Rating Method for Obtaining Similarity Data in the Conceptual Domain

Steven Verheyen[a,b], Anne White[b], and Gert Storms[b]

[a]Erasmus University Rotterdam; [b]KU Leuven

ABSTRACT

We compare two methods for obtaining similarity data in the conceptual domain. In the Spatial Arrangement Method (SpAM), participants organize stimuli on a computer screen so that the distance between stimuli represents their perceived dissimilarity. In Total-Set Pairwise Rating Method (PRaM), participants rate the (dis)similarity of all pairs of stimuli on a Likert scale. In each of three studies, we had participants indicate the similarity of four sets of conceptual stimuli with either PRaM or SpAM. Studies 1 and 2 confirm two caveats that have been raised for SpAM. (i) While SpAM takes significantly less time to complete than PRaM, it yields less reliable data than PRaM does. (ii) Because of the spatial manner in which similarity is measured in SpAM, the method is biased against feature representations. Despite these differences, averaging SpAM and PRaM dissimilarity data across participants yields comparable aggregate data. Study 3 shows that by having participants only judge half of the pairs in PRaM, its duration can be significantly reduced, without affecting the dissimilarity distribution, but at the cost of a smaller reliability. Having participants arrange multiple subsets of the stimuli does not do away with the spatial bias of SpAM.

KEYWORDS
Spatial arrangement; pairwise rating; similarity; dissimilarity; proximity; reliability; representation; skewness

Introduction

According to William James (1980, p. 459) the "*sense of sameness is the very keel and backbone of our thinking.*" Similarity is indeed assumed to be at the basis of fundamental cognitive processes such as object recognition (Humphreys et al., 1988; Humphreys & Forde, 2001), categorization (Nosofsky, 1988, 1992), and generalization (Shepard, 1987, 2004). As a result, many cognitive models operate on a representation that captures the similarity of the entities that are being processed (e.g., Gärdenfors, 2000; Navarro & Lee, 2004; Nosofsky, 1986; Shoben, 1983; Tversky, 1977). Given the importance that is attributed to similarity in numerous cognitive theories and models, it is important that researchers are able to obtain accurate measurements of similarity.

The measurement of similarity is not without challenges. Some of these are independent of the method that is chosen to obtain similarity measures. There exist, for instance, pronounced inter- and intra-individual differences in similarity perception that need to be acknowledged (Ashby et al., 1994; Lee & Pope, 2003; Summers & MacKay, 1976). These individual differences result from the context-dependent nature of similarity (Goldstone et al., 1997; King & Atef-Vahid, 1986; Medin et al., 1993; Tversky, 1977) and from individuals' differing experience with the entities under consideration (Charest et al., 2014; Coltheart & Evans, 1981; Medin et al., 1997). Some challenges are specific to the stimulus domain that is being assessed. For instance, when the goal is to assess how similar different wines smell, the samples need to be presented in dark glasses to ensure that visual information such as the wines' color does not influence the judgments (Ballester et al., 2005). Other challenges are specific to the method that is being used to asses similarity. There is no single method that provides the ideal measurement of similarity in all circumstances. When considering which method to use to measure similarity, researchers should carefully consider both the advantages and the disadvantages of the available methods.

In the following section, we will compare the characteristics of the Pairwise Rating Method (PRaM) and the Spatial Arrangement Method (SpAM) for

CONTACT Steven Verheyen verheyen@essb.eur.nl Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Post Box 1738, 3000 DR, Rotterdam, The Netherlands.

measuring similarity. The former is the predominant method for measuring similarity in the conceptual domain (e.g., De Deyne et al., 2016; Dry & Storms, 2009; Hill et al., 2015; Migo et al., 2013; White et al., 2014). According to Dry and Storms, 65% of similarity data sets in the semantic literature are obtained with this method. SpAM (Goldstone, 1994; Hout et al., 2013) is, however, increasingly being used, presumably because it allows similarity data to be obtained in a much faster manner than PRaM. As software to collect similarity data online through SpAM has recently become available in the form of JavaScript code implemented in the browser-based survey software Qualtrics (Koch et al., 2020), it is to be expected that the use of the method will only increase.

Comparison

PRaM and SpAM are both direct methods for collecting similarity data, meaning that the similarity indices are directly obtained from participants rather than derived from other data (Borg et al., 2013). In PRaM, all pairs of stimuli are presented to participants, who judge their perceived similarity on a Likert scale. In SpAM, all stimuli are presented to participants, who spatially organize them so that their distances are inversely related to their perceived similarity (Goldstone, 1994; Hout et al., 2013).

PRaM is a rather straightforward method: Participants are presented with pairs of stimuli and have to rate the stimuli's similarity on a Likert scale. This is the type of rating task that most participants in surveys and experiments are likely to be familiar with. The common criticism that Likert scales have an arbitrary precision therefore also applies to PRaM. When there is a mismatch between the granularity of a participant's similarity distinctions and the number of alternatives that is offered by the Likert scale, there is a concern that the resulting similarity judgments may become unreliable (Borg et al., 2013). According to Hout et al. (2013), the resolution that typical Likert scales offer is too limited for participants to convey their similarity perceptions.[1] Despite these concerns, average similarity data obtained with PRaM tend to be reliable (Bijmolt & Wedel, 1995; Giordano et al., 2011; Verheyen et al., 2016).

The steep expansion of the number of pairs to judge as the size of the stimulus set increases is considered the biggest drawback of PRaM. It makes the method ill-suited for large stimulus sets (Giordano et al., 2011; Kriegeskorte & Mur, 2012; Tsogo et al., 2000) and use in patient populations (White et al., 2014) where there is a genuine concern for detrimental effects of fatigue, inattention, boredom, and disengagement on data quality. Lengthy data collection protocols also increase the chance that participants will change their judgment strategy within a session (Hout et al., 2013). As they encounter more and more stimulus pairs to judge, participants might recalibrate the scale or attach different weights to the different stimulus dimensions. Related to this latter concern is the fact that in most implementations of PRaM, participants only see two stimuli at a time. Stimulus pairs are thus judged in isolation, and participants might only become aware of the full (dis)similarity range after having judged several stimulus pairs (Goldstone, 1994; Hout et al., 2013). This not only obliges participants to develop a rating strategy over time (making the first judgments unrepresentative) as more information regarding the stimulus domain becomes available to them, but also seems at odds with the observation that similarity is context dependent (see Goldstone et al., 1997; King & Atef-Vahid, 1986; Medin et al., 1993; Tversky, 1977) and most researchers would likely have the similarity of individual stimulus pairs be judged in the context of the relevant comparison class.

Note that the isolated presentation of stimulus pairs is not an inherent characteristic of PRaM and several researchers have in practice accommodated this potential concern by providing participants with an overview of the stimuli that will be judged prior to the pairwise similarity judgment task (e.g., Richie et al., 2020; Verheyen & Storms, 2011), with a sample of the pairs that will be judged (Goldstone, 1994), or to have ratings remain visible so that participants can refer back to previous judgments (Hutchinson & Lockhead, 1977). Hout et al. (2013) have recently proposed a variant on PRaM, which they termed Total-Set PRaM. In Total-Set PRaM, participants get to see the entire stimulus set at all times. On each trial, two stimuli are highlighted for pairwise similarity rating (see left panel of Figure 1). This way, the context of the judgments is clear from the onset and the similarity of two stimuli can be judged against the background of the entire comparison class.[2]

[1] But see Green and Wind (1973) who show in a simulation study that even with a coarse scale one can recover the underlying similarity structure using multidimensional scaling.

[2] Nakatsuji et al. (2016) had participants rank order all pairs of stimuli in terms of similarity. Kriegeskorte and Mur (2012) introduced yet another method, having participants arrange all stimulus pairs on a one-dimensional dissimilarity scale. Both these methods allow participants to appreciate the entire range of (dis)similarity at once as well. The latter method appears to be a sort of crossover between PRaM and SpAM.

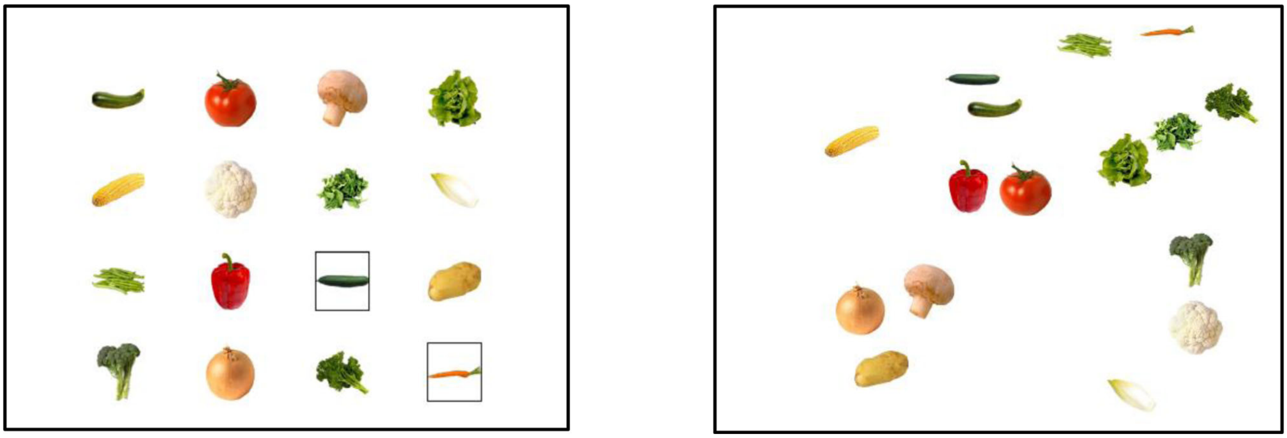

Figure 1. Illustration of Total-Set Pairwise Rating Method (PRaM, left) and the Spatial Arrangement Method (SpAM, right) for the category *vegetables* ($n = 16$). In PRaM, all items are shown simultaneously and on every trial, two of them are highlighted to be judged in terms of similarity on a Likert scale. In SpAM, participants spatially organize the simultaneously presented items so that their distances are inversely related to their perceived similarity. The right panel shows a completed example.

SpAM was developed to overcome the most apparent problems of PRaM. By organizing stimuli on a surface according to their perceived similarity, participants can convey more nuanced levels of (dis)dissimilarity compared to when they use a Likert scale (see right panel of Figure 1 for a completed example). When the stimulus organization is done on a computer screen, for instance, the level of precision corresponds to that of the screen resolution (Hout et al., 2013). The data collection also occurs in a more efficient manner because one does not need to go through all pairs of stimuli separately. Moving a single stimulus on the surface immediately adjusts its distance to all other stimuli (Goldstone, 1994; Hout et al., 2013). Proponents of SpAM argue that because of this, data collection with SpAM will not only proceed much quicker, it will be far more engaging and far less repetitive, reducing the risk of boredom and the ensuing detrimental effects on data quality even more. Moreover, SpAM is an inherently contextualized procedure in which all relevant stimuli are simultaneously present, making the (dis)similarity range immediately apparent to participants (Goldstone, 1994; Hout et al., 2013). Note that the contextual nature and the efficiency of SpAM go hand in hand. Because the relations between the stimuli are spatially represented, participants are not required to provide seemingly redundant answers. Wherein PRaM participants need to indicate explicitly that a *pineapple* is dissimilar from both a *lemon* and a *lime*, in SpAM this can be achieved at once by moving the *pineapple* away from the highly similar and thus closely positioned citrus fruits.

SpAM is not without disadvantages, however. Verheyen et al. (2016) have formulated a number of caveats for the method. When the number of PRaM and SpAM participants is equated, the average SpAM similarity data tend to be less reliable. Participants might be quicker to finalize a spatial arrangement of n stimuli than to judge the similarity of $n * (n\text{–}1)/2$ stimulus pairs; they also demonstrate more variability in the similarity judgment of the stimuli. A combination of factors might be responsible for this. When moving a stimulus in SpAM, participants might not give due consideration to the effects this has on all of the $n\text{–}1$ similarity measures it affects. In PRaM, on the other hand, participants are obliged to consider every similarity measure separately. Participants might also approach SpAM as a discrete sorting task, clustering highly similar stimuli together without much consideration for the within- or between-cluster distances. Other factors signaled by Verheyen and colleagues pertain to the inability to convey more than two stimulus dimensions on a two-dimensional surface (requiring participant to make a selection when more dimensions are available) and the obligation for participants to convey similarity in a geometric space with continuous dimensions, while they might in fact entertain (discrete) feature representations. Together, these factors might explain why average SpAM data have a more modest reliability[3], although the individual contribution of each of the factors might vary

[3] When we use the term *reliability* in this paper, we use it to indicate how comparable the similarity data of different participants are, not to quantify how stable the similarity data of a single participant are in time. We focus on the former because aggregating similarity data is common practice in the literature.

between applications. For instance, Verheyen et al. show that these factors are less of a concern for simple perceptual stimuli than they are for complex conceptual stimuli (see also below). Moreover, they can be alleviated by providing participants with instructions or examples on how to convey featural information or additional dimensions in their two-dimensional configurations (Hout & Goldinger, 2016) or by having participants arrange subsets of the stimuli on subsequent trials so that they may convey additional information (e.g., Berman et al., 2014; Coburn et al., 2019; Goldstone, 1994; Horst & Hout, 2015). In the latter case, the context shifts from trial to trial, allowing more complex relationships between stimuli to be captured (see also below). It also needs to be acknowledged that because SpAM takes little time to complete, it is fairly easy to obtain data from additional participants in order to increase the reliability (Hout & Goldinger, 2016). Although the reliability of similarity data is not routinely assessed, it is not without consequences. Representations of unreliable average similarity data are not a good reflection of the shared structure among the participants (Ashby et al., 1994; Lee & Pope, 2003), are less likely to be reproduced (Sturidsson et al., 2006; Verheyen & Peterson, 2020; Voorspoels et al., 2014; White et al., 2014), are not necessarily representative for the individual similarity patterns (Bocci & Vichi, 2011; Okada & Lee, 2016), and limit the predictive ability of the data (White et al., 2014).

The spatial nature of SpAM has also been said to impose structure on the resulting similarity data. Verheyen et al. (2016) suggested that SpAM would have a bias for spatial representations, regardless of whether the underlying stimuli are truly spatially embedded. That is, SpAM similarities would display the typical characteristics of geometric spaces, biasing their representations against alternative, non-spatial representations. This would make SpAM less suited for use in exploratory studies, where the goal of the similarity data collection is to uncover the nature of a stimulus domain of which the representational structure is unknown. If these concerns were to prove valid, this would not bode well for SpAM, as data exploration and the testing of structural hypotheses are among the main applications of similarity data collection methods (Borg & Groenen, 2005). Although proponents of SpAM argue that the method offers an intuitive way of providing similarity data because we tend to conceptualize similarity in a spatial manner (Hout et al., 2013; Richie et al., 2020), Verheyen et al. argue that this claim does not hold across all stimulus domains.

Outline

Since the use of SpAM is on the rise, we deem it important to empirically evaluate the two main points of criticism that have been offered against the method: (i) SpAM's speed trades off with its reliability and (ii) SpAM favors spatial over feature representations.[4] To assess these claims, we will compare SpAM with Total-Set PRaM. The latter method has everything of the classic PRaM, but judgments are made in a context-dependent manner, just like in SpAM. Any differences found between the methods can therefore not be attributed to a lack of contextualization. Because in both methods all stimuli are simultaneously present on the screen, they also compare favorably in terms of visual appearance. SpAM remains the more interactive of the two methods, though. When we henceforth use the abbreviation PRaM, we use it to refer to Total-Set PRaM.

PRaM and SpAM will be applied to four sets of conceptual stimuli, comprised of photorealistic images of exemplars of the categories *birds*, *vehicles*, *vegetables*, and *sports*. We chose one category for each of the domains of natural categories, artifact categories, natural artifact categories, and activity categories (Verheyen et al., 2019) to have a sample of categories that would be representative for conceptual categories as a whole.[5] We will employ a within-subjects design whereby every participant provides similarity data for two categories using SpAM and for the other two categories using PRaM. Categories and methods will be counterbalanced, ensuring an equal number of participants per method–category combination. It is warranted that SpAM be evaluated on conceptual stimuli because it is unclear whether the method is equally appropriate for perceptual and conceptual stimuli (Hout et al., 2013; Verheyen et al., 2016). The richness of conceptual stimuli might be a problem for the two-dimensional SpAM because participants might want to communicate more than two dimensions of variation (Richie et al., 2020). When participants make different choices as to which dimensions to communicate and/or employ idiosyncratic strategies for conveying additional information, this might be detrimental for the reliability of the data. The use of conceptual stimuli also allows any representational issues to be

[4] Verheyen et al. (2016) formulated a third caveat for SpAM, suggesting that it might invoke a bias against high-dimensional representations. Since it would require multidimensional scaling analyses to assess this, we defer this topic to another paper. See Hout and Goldinger (2016) and Richie et al. (2020) for counterarguments.

[5] We will not go into differences between categories or domains in this paper and defer an investigation of such differences to future work in which the domains can be systematically compared using several instances.

checked since conceptual stimuli are generally considered to be represented in terms of features, as opposed to perceptual stimuli that tend to be represented in a spatial manner (Dry & Storms, 2009; Pruzansky et al., 1982; Tversky & Hutchinson, 1986; Verheyen et al., 2016). Paradigmatic examples of perceptual stimuli are forms, colors, and sounds (Pruzansky et al., 1982). Although we use photorealistic images of category exemplars, we consider our stimuli conceptual as they pertain to semantic categories. The perceptual-conceptual distinction should thus not be equated with a difference in presentation format (pictorial vs. verbal).

SpAM yields as output Euclidean distances between stimuli, measured in pixels. For comparability, PRaM similarities will be converted into dissimilarities by subtracting the similarity ratings on the nine-point Likert scales (1 = very dissimilar; 9 = very similar) from 10. This way, both PRaM and SpAM yield measures of dissimilarity. We will compare SpAM and PRaM in terms of completion time (duration in seconds), reliability (split-half reliability), and distributional characteristics of the ensuing dissimilarity data (skewness and centrality). No transformation or standardization will be applied to the dissimilarity data, as this is not common practice in the similarity measurement literature[6] and because individual differences in absolute similarity appraisal may be of interest.

By comparing the completion time and reliability of the two methods, we can evaluate the first caveat that has been raised for SpAM: While it might be faster to obtain dissimilarity data with SpAM than with PRaM, the reliability of the former will be lower than that of the latter when an equal number of participants provide PRaM and SpAM data. For completion time, we will report per method and category combination the mean and standard deviation of the task duration (in seconds), conduct Mann–Whitney tests to establish whether SpAM takes significantly less time to complete than PRaM, and indicate the task duration ratio. Per combination of method and category, we will also report the reliability, which we establish by computing the split-half correlation between the dissimilarity measures across exemplar pairs and correcting it with the Spearman-Brown formula (Lord & Novick, 1968). The reported reliability values are averages across 10,000 random splits of the data. Taken PRaM reliability as the standard, we also indicate the number of participants who need to be tested using SpAM to attain the same level of reliability. To this end, we compute the factor k, with which the current number of participants needs to be multiplied, using the formula provided by Lord and Novick (1968):

$$k = \frac{\rho_D(1-\rho_O)}{\rho_O(1-\rho_D)},$$

with the desired reliability ρ_D equal to PRaM's reliability and the observed reliability ρ_O equal to SpAM's reliability.

By comparing the distributional characteristics of the dissimilarity data of the two methods, we can evaluate the second caveat that has been raised for SpAM: because of its spatial nature, SpAM might be biased against feature representations. The distributional characteristics of dissimilarity data can be used to establish in what way stimuli are best represented (Dry & Storms, 2009; Ghose, 1998; Giordano et al., 2011; Verheyen et al., 2016). The most widely used characteristics are skewness and elongation (Sattath & Tversky, 1977) and centrality and reciprocity (Tversky & Hutchinson, 1986). We will restrict our discussion to skewness and centrality, because unlike elongation and reciprocity, the results of these distributional characteristics are not affected by differences in the granularity of dissimilarity data, a characteristic on which SpAM and pairwise data differ.[7] Positively skewed dissimilarity data accord well with spatial representations, while negatively skewed dissimilarity data accord better with feature representations (Sattath & Tversky, 1977). When stimuli vary continuously along dimensions, the majority is positioned relatively close together. Only the stimuli at opposite ends of the dimensions are far apart. Feature representations, on the other hand, are particularly well suited to capture hierarchical structures, comprised of many large between-cluster dissimilarities and few small within-cluster dissimilarities. These representations are typical for the mutually exclusive stimulus organizations people spontaneously introduce and the increasingly divergent structures that result from evolutionary processes (Sattath & Tversky, 1977). Typically, hierarchical structures also include focal stimuli that form the centers of the clusters or the starting point of the evolutionary process. The centrality of these focal stimuli can be expressed as the

[6] Unless spatial arrangements are obtained on screens of different sizes (see Koch et al., 2020), which was not the case here.

[7] All analyses were also repeated on SpAM dissimilarity measures of reduced granularity. To this end, exemplars' distance in pixels was rounded to the nearest hundred (e.g., 713 pixels becomes 7; see also Hout et al., 2013, and Verheyen et al., 2016). This yielded results comparable to those reported here, indicating that any differences between SpAM and PRaM are not due to precision differences.

number of times they are the nearest neighbor of other stimuli. Stimuli at the center of a cluster are clearly more often the nearest neighbor of other stimuli than stimuli at the border of a cluster. In continuous spatial representations, on the other hand, stimuli will generally only be the nearest neighbor of one or a few other stimuli. That is, compared with the feature representations that are apt at capturing hierarchical structures, fewer stimuli will stand out as focal or highly central in spatial representation. Centrality values higher than 2 are therefore taken to indicate that the data are better represented by feature models than by spatial ones (Tversky & Hutchinson, 1986). We compute the centrality of each participant's dissimilarity data using the formula from Tversky and Hutchinson (1986):

$$C = \frac{1}{n+1}\sum_{e=0}^{n} N_e^2.$$

where $S = \{0, 1, \ldots, n\}$ is the set of exemplars and N_e reflects the focality of exemplar e with $N_e = 0$ if there is no element in S whose nearest neighbor is e and $N_e = n$ if e is the nearest neighbor of all other stimuli. Because of the occurrence of multiple ties in the pairwise dissimilarity data and its potential influence on the results, the computation was repeated 100 times, each time breaking ties at random.

We will present the results of three studies comparing PRaM and SpAM. In Study 1, both methods are compared in terms of completion time, reliability, and distributional characteristics of the dissimilarity data, for the conceptual stimuli *sports*, *vegetables*, *vehicles*, and *birds*. In Study 2, we investigate to what extent the results of Study 1 generalize to conceptual categories of differing sizes. To that effect, the number of exemplars of the four conceptual categories is varied. Where all categories in Study 1 comprise 16 exemplars, the number of exemplars per category in Study 2 varies between 8 and 32, which spans the typical set size in similarity measurement studies (Hout et al., 2018). In Study 3, variants of PRaM and of SpAM are compared on the same materials as those used in Study 2. Both variants are aimed at accommodating a shortcoming of their respective methods. By only presenting half of the exemplar pairs for judgment, the completion time of PRaM is expected to be halved. By subsequently arranging various subsets of the exemplars, more information can presumably be communicated than on a single SpAM trial. We report how these variants compare to each other, and to the results obtained with the original methods in Study 2.

All three studies were conducted in Dutch. All participants were undergraduate students at the University of Leuven (KULeuven, Belgium) who were native speakers of Dutch. They were compensated either with course credit or at a rate of 8 euros/hour. All three studies were implemented in the E-Prime software for behavioral research (Schneider et al., 2002). The analyses were conducted with JASP Team (2019). A significance level of $\alpha = .05/4 = .0125$ is used in all significance tests to acknowledge the fact that testing is done for multiple categories. The materials and the data that support the findings of Studies 1–3 are openly available on the Open Science Framework at https://osf.io/9s2qe/.

Study 1

Participants

Forty-eight undergraduate students (39 women, 9 men), aged between 17 and 55 years old[8], participated in Study 1. They were offered the choice to be compensated financially (25%) or with course credit (75%).

Materials

For each of the four categories (*birds*, *vegetables*, *vehicles* and *sports)* we included photorealistic images of the 16 most familiar exemplars according to the De Deyne (2014) norms. The choice of the most familiar exemplars was based on the average familiarity rating across 20 raters (50% female, aged between 20 and 28 years, $M = 23.05$, $SD = 1.85$), who had a seven-point Likert scale at their disposal with higher values indicating higher familiarity. An overview of the exemplars is provided in Table A1 in the Appendix A. See Figure 1 for examples of the stimuli for the category *vegetables*. The decision to include 16 exemplars per category was based on the consideration that 16 images can be comfortably fit on a screen in 4-by-4 grid and having participants judge the similarity of all 16*15/2 = 120 pairs of exemplars of a category is still feasible.

Procedure

After completing an informed consent, every participant provided similarity data for two categories using PRaM and for two categories using SpAM. In this

[8] The original file with demographic information was lost, preventing us from reporting the mean and standard deviation for age.

Table 1. Mann–Whitney test comparing PRaM and SpAM on completion time (seconds) per category of 16 exemplars in Study 1.

	PRaM		SpAM					
Category	*M*	*SD*	*M*	*SD*	*W*	*p*	*r*	*k*
sports	478.667	131.530	196.792	95.960	547.50	<.001	.90	2.43
vegetables	401.208	199.408	195.000	109.667	507.00	<.001	.76	2.06
vehicles	442.875	167.447	172.583	69.098	561.00	<.001	.95	2.57
birds	452.333	177.794	183.417	87.756	557.50	<.001	.94	2.47

Note. Effect size is given by the rank biserial correlation *r*. The value *k* represents the duration ratio (PRaM/SpAM).

manner, we obtained 24 similarity data sets per combination of method and category. Four categories can be presented in 24 different orders. Each order was completed by two participants, alternating SpAM with PRaM, and with one of the participants starting with SpAM, while the other one started with PRaM (resulting in two method orders for every ordered set of categories: SpAM – PRaM – SpAM – PRaM vs. PRaM – SpAM – PRaM – SpAM). For every new participant, the stimuli were randomly positioned in a 4-by-4 grid on the screen.

In PRaM, participants were invited to judge the similarity of all 120 pairwise exemplar combinations on a nine-point Likert scale (1 = very dissimilar, 9 = very similar). Participants indicated their response by pressing a numerical key. On every trial, the exemplars that were to be rated in terms of similarity were indicated by a black border (see left panel of Figure 1). Throughout the rating of a pair, all other exemplars remained visible on the screen without black order, along with the rating scale on the bottom of the screen. The highlighting of exemplar pairs occurred in a random order for every new participant.

In SpAM, participants were invited to position the exemplars in such a way that the distance between any two exemplars on the screen reflected how similar they perceived them: the more similar they were found to be, the closer they needed to be positioned; the more dissimilar they were regarded, the further apart they needed to be positioned. Participants could position exemplars anywhere on the screen by dragging them with the computer mouse. By right clicking the mouse, participants could indicate that they were satisfied with the stimulus configuration. As a safe guard against unintended premature completions, "*Have you finished organizing the stimuli?*" was presented upon right clicking the mouse. If participants pressed the Y key, indicating that they were finished ("*Yes, I am finished.*"), they were directed to the next category (or the experiment finished when it was the last category). If they pressed the N key, indicating that they needed more time ("*No, I need more time.*"), they were returned to the configuration in the state they had left it. Finally, participants who pressed the S key ("*I want to start over.*") were returned to the 4 × 4 starting configuration.

Once participants had provided similarity data for all four categories, they were presented with a survey intended to assess their experiences with both methods. Participants were invited to indicate which method they found most (1) clear, (2) pleasant, (3) easy, and (4) tiresome. The survey concluded with two open questions asking to list the perceived (dis)-advantages of both methods, and a binary question, asking about participants' preferred method: "*If you were to repeat this study, with just one method, which one would you choose?*" We constructed two versions of this survey: one in which SpAM was always mentioned before PRaM, and one in which PRaM was always mentioned before SpAM. The former was administered to participants who used SpAM for their first category; the latter was administered to participants who started with PRaM. We defer the discussion of the survey data to a later section (see section *Survey responses*) in which the results of Studies 1–3 are treated simultaneously.

Results

Duration

Table 1 lists the average completion time (in seconds) per combination of method and category. A Mann–Whitney test established that PRaM took longer to complete than SpAM, for each of the four categories. On average, participants spent just over 7 minutes judging the 120 exemplar pairs of a category, and just over 3 minutes organizing 16 exemplars on the screen. The value *k* in Table 1 indicates the average duration ratio of PRaM vs. SpAM per category. With 16 stimuli per category, SpAM is about 2.3 times faster than PRaM.

Reliability

Table 2 lists the estimated reliability of the average dissimilarity data per combination of method and category. PRaM's reliability is higher than that of SpAM, for each of the four categories, with an average of .96 compared to .88. The value *k* in Table 2 represents the factor with which the number of SpAM participants needs to be multiplied to obtain the same reliability as PRaM. With 16 stimuli per category, an average of 84 participants needs to be tested with SpAM to obtain a similar reliability as PRaM with 24 participants. That is, about 3.5 times more

Table 2. Reliability of the average dissimilarity data in Study 1.

Category	PRaM	SpAM	*k*	*N*
sports	.97	.89	3.96	96
vegetables	.94	.83	3.30	80
vehicles	.98	.94	2.73	66
birds	.96	.86	4.03	97

Note. k represents the factor with which the number of SpAM participants needs to be multiplied to obtain the same reliability as PRaM. *N* represents the resulting number of participants.

Table 3. Mann–Whitney test comparing the skewness of PRaM and SpAM dissimilarity data in Study 1.

	Individual proximities							Average proximities	
	PRaM		SpAM					Skewness	
Category	*M*	*SD*	*M*	*SD*	*W*	*p*	*r*	PRaM	SpAM
sports	−1.25	.57	.38	.24	5.00	<.001	−.98	−1.59	−.68
vegetables	−.95	.69	.26	.18	3.00	<.001	−.99	−1.27	−1.34
vehicles	−.87	.56	.31	.25	18.00	<.001	−.94	−.96	−.65
birds	−1.22	1.19	.34	.24	2.00	<.001	−.99	−1.36	−.80

Note. Effect size is given by the rank biserial correlation *r*.

participants are required to obtain equally reliable results. Of course, these numbers are dependent on the level of reliability one wants to obtain and the current analysis assumes that researchers considering using SpAM intend to obtain the level of reliability they are accustomed to using PRaM. We acknowledge that the reported SpAM reliabilities are already considerable. The correlations between the average dissimilarity data of the two methods are at the maximum level one could expect given SpAM reliabilities, with the Pearson correlation equal to .87 for *sports*, .91 for *vegetables*, .95 for *vehicles*, and .90 for *birds*.

Bias

Per combination of method and category, Tables 3 and 4 respectively list the average skewness and centrality across the individual dissimilarity data sets. Mann–Whitney tests were used to establish that PRaM dissimilarity data are more negatively skewed and have a higher centrality than SpAM dissimilarity data sets. The difference was significant at $\alpha = .0125$ for each of the four categories, except for centrality in the case of *vehicles* ($p = .015$). The average skewness was negative for PRaM dissimilarity data (–1.07) and positive for SpAM dissimilarity data (.32). The average centrality was higher for PRaM dissimilarity data (1.90) than for SpAM dissimilarity data (1.63), but did not exceed the critical value of 2 that was put forward by Tversky and Hutchinson (1986) in the majority of data sets (75% of PRaM data sets compared to 88.54% of SpAM data sets).

Table 4. Mann–Whitney test comparing the centrality of PRaM and SpAM dissimilarity data in Study 1.

	Individual proximities							Average proximities	
	PRaM		SpAM					Centrality	
Category	*M*	*SD*	*M*	*SD*	*W*	*p*	*r*	PRaM	SpAM
sports	2.10	.36	1.69	.20	491.50	<.001	.71	2.32	1.75
vegetables	1.95	.34	1.60	.25	469.00	<.001	.63	1.93	1.50
vehicles	1.78	.22	1.64	.21	406.00	.015	.41	1.63	1.50
birds	1.76	.21	1.58	.24	420.50	.006	.46	1.63	1.50

Note. Effect size is given by the rank biserial correlation *r*.

Tables 3 and 4 also indicate for each category the skewness and the centrality of the average PRaM and SpAM dissimilarity data, obtained by averaging the individual dissimilarity data sets across participants. Both for PRaM and SpAM, the averaging leads to dissimilarity data with a lower skewness compared to the average skewness of the individual data. The difference is much more pronounced for SpAM (–.87 compared to .32 across categories) than it is for PRaM (−1.30 compared to −1.07). Where the individual SpAM dissimilarities tended to be positively skewed, the average SpAM dissimilarity data are negatively skewed. As a result, the distributions of the average PRaM and SpAM dissimilarity data are much more comparable. The average PRaM data remain more negatively skewed than the average SpAM data, however. The results of the averaging on centrality are less consistent. The centrality of the average dissimilarity data tends to be lower than the average centrality of the individual dissimilarity data, both for PRAM and SpAM, except for the category of *sports*. The average PRaM data still have a higher centrality than the average SpAM data, however (1.88 compared to 1.56 across categories, a difference comparable to that of the average centrality: 1.90 vs. 1.63). Only one of the centrality values for the average dissimilarity data exceeds 2 (PRaM *sports*).

Discussion

The findings from Study 1 empirically confirm the caveats that were raised regarding SpAM by Verheyen et al. (2016). Participants were much faster to complete an organization of the exemplars of a conceptual category than they were to rate the similarity of all pairs of exemplars. This increase in efficiency came at the cost of a decrease in reliability. With 16 exemplars per category, SpAM was about 2.3 times faster to complete than PRaM, but requires about 3.5 times the number of participants to attain the reliability that is obtained by having 24 participants complete all pairwise judgments. It thus seems that researchers choosing to use either PRaM or SpAM are faced with a tradeoff between speed and accuracy. This choice only presents itself when one wants to attain the high level of reliability that PRaM affords (> .94 in all categories). Our results indicate that if researchers are satisfied with a reliability of .80 (a common lower limit in psychological studies[9]), they can suffice with running 24 SpAM participants for categories comprising 16 exemplars. Note that under these circumstances, the overall completion time (the number of participants times average completion time) is comparable for PRaM and SpAM since PRaM, while taking more time to complete, requires fewer participants than SpAM to attain a .80 reliability.

The positive skewness of SpAM dissimilarity data is in line with known distributional characteristics of distances obtained from spatial representations such as the one used in SpAM (Sattath & Tversky, 1977). The fact that the skewness of the individual PRaM dissimilarity data was found to be negative suggests that the conceptual categories need not necessarily be represented in a spatial manner, and feature representations should be considered.[10] We are confident that this discrepancy is the result of bias in SpAM rather than PRaM, since Verheyen et al. (2016) established in a comparison of perceptual and conceptual categories that SpAM consistently yielded dissimilarity data with a positive skewness, while the sign of the skewness of PRaM dissimilarity data depended on the nature of the category: negative in the case of conceptual categories and positive in the case of perceptual categories. The results for centrality are largely in line with those for skewness, in that SpAM dissimilarity data tended to have a lower centrality than PRaM similarity data, which again suggests that SpAM is biased toward spatial representations. More PRaM than SpAM data sets had a centrality higher than 2, which is the cutoff point for considering a feature rather than a spatial configuration. The evidence on the basis of centrality was not as strong as that on the basis of skewness, however, in that most data sets did not demonstrate a centrality higher than 2.

Averaging tended to have an effect on the skewness of both PRaM and SpAM dissimilarity data, but it was more pronounced for SpAM than for PRaM. While the skewness of the average data was always more negative than the average skewness of the individual data, for SpAM it involved a change in sign from positive to negative. That is, while the individual SpAM data were characterized by a relatively small number of large dissimilarities, the average SpAM data were characterized by a relative large number of large dissimilarities. Averaging also tended to decrease centrality, which is somewhat at odds with its effects on skewness, in that it provides less evidence for a feature representation, while a decrease in skewness provides more evidence in favor of such a representation. The average PRaM and SpAM dissimilarity data were found to be more similar to each other than the individual dissimilarity data in terms of skewness, but not centrality. The increased distribution similarity was also reflected in the pronounced correlation between the average PRaM and SpAM dissimilarities (all > .87) compared to the average correlations of the individual dissimilarity ratings (.36 for *sports*, .27 for *vegetables*, .51 for *vehicles*, .32 for *birds*). It thus appears that for conceptual stimuli, PRaM and SpAM do not provide equivalent dissimilarity data, but that the discrepancy decreases when the data are averaged across participants. Researchers are expected to draw similar conclusions for the average SpAM and PRaM data from Study 1. Although this is an encouraging finding for researchers who intend to use aggregate SpAM data, it is curious that average SpAM data are not representative of individual SpAM data. A related observation was made by Richie et al. (2020). They found that although participants can only convey two dimensions in SpAM, aggregating the data and subjecting it to multidimensional scaling could nevertheless yield more than two dimensions, presumably because different participants convey different dimensions (see also Verheyen & Storms, 2020).[11] It thus appears that average SpAM data are not representative for individual SpAM data because the amount of information individuals can convey in a single spatial arrangement is limited. Researchers might therefore want to refrain from using SpAM to study individual differences, unless their goal explicitly is to understand which information participants convey when the circumstances only allow a limited number of dimensions to be communicated. Based on Study 1, we recommend the use of PRaM for the study of individual differences in similarity perception.

[9] What constitutes an acceptable reliability is dependent on the nature of the data set and the purpose of the study. The reliability increases with the number of stimuli it is computed over. It is also the upper boundary for correlations with external variables. Since conceptual similarity is often used to predict other variables (see Verheyen, Ameel, & Storms, 2007, for an overview) it is desirable that the reliability is as high as possible.

[10] Note that this difference presents despite the fact that we used photorealistic images for the conceptual category exemplars, indicating that it is not the presentation format that is at the basis of the perceptual-conceptual distinction.

[11] This source of individual differences might explain why the reliability of SpAM is lower than that of PRaM when the number of participants is equated.

Table 5. Mann–Whitney test comparing PRaM and SpAM on completion time (in seconds) per category in Study 2.

		PRaM		SpAM					
Category	# exemplars	*M*	*SD*	*M*	*SD*	*W*	*p*	*r*	*k*
sports	8	131.708	46.379	129.375	75.961	350.00	.205	.22	1.02
*vegetables**	16	193.875	76.378	182.542	78.048	320.50	.509	.11	1.06
vehicles	24	741.750	238.175	260.208	80.590	567.00	.001	.97	2.85
birds	32	1378.542	508.898	428.958	229.345	567.00	.001	.97	3.21

Note. Effect size is given by the rank biserial correlation *r*. The value *k* represents the duration ratio (PRaM/SpAM). *Only half of the exemplar pairs of *vegetables* were presented in PRaM.

Table 6. Reliability of the average dissimilarity data in Study 2.

Category	# exemplars	PRaM	SpAM	*k*	*N SpAM*
sports	8	.98	.94	2.74	66
*vegetables**	16	.89	.86	1.21	30
vehicles	24	.98	.91	4.45	107
birds	32	.94	.84	3.19	77

Note. *k* represents the factor with which the number of SpAM participants needs to be multiplied to obtain the same reliability as PRaM. *N* represents the resulting number of participants. *Only half of the exemplar pairs of *vegetables* were presented in PRaM.

Study 2

The purpose of Study 2 is threefold. Since task duration and reliability are dependent on the number of stimuli, we will repeat the comparison between PRaM and SpAM with a different number of stimuli per category to see how this affects the duration and sample size ratio. As such, Study 2 also serves as a replication of the previous findings regarding the spatial bias and lack of representativity of SpAM. Finally, we will investigate whether it is possible to reduce PRaM completion time without affecting the data quality, by only presenting 50% of a category's exemplar pairs. As was indicated in the rationale for the development of SpAM, many pairs provide redundant information (Goldstone, 1994; Hout et al., 2013; see also Young & Cliff, 1972). This can be capitalized on by using incomplete rating tasks in which only a subset of pairs is presented to participants for rating. We will apply this procedure to the category of *vegetables* comprised of the 16 exemplars of Study 1, while we will apply the standard Total-Set PRaM to the categories *sports*, *vehicles*, and *birds*, but with a different number of exemplars than in Study 1 (8, 24, and 32, respectively).

Participants

Forty-eight undergraduate students (42 women, 6 men), aged between 17 and 36 years old ($M = 19.94$, $SD = 3.94$), participated in Study 2. They were financially compensated for their participation at a rate of 8 euros/hour.

Materials

We used the same categories that were used in Study 1, but with a different number of exemplars each: 8 for *sports*, 16 for *vegetables*, 24 for *vehicles*, and 32 for *birds*. The selected stimuli again corresponded to the most familiar exemplars according to De Deyne (2014). An overview can be found in Table A1 in the Appendix A.

Procedure

Study 2 followed the same procedure as Study 1, with one exception. For the category *vegetables* participants were only presented with half of the exemplar pairs (60 instead of $16*15/2 = 120$) for judgment in PRaM (all 16 exemplars were presented in SpAM). Which half was presented, was randomly determined for every new participant, meaning that different participants judged different pairs.

As before, the 16 *vegetable* exemplars were randomly organized in a 4-by-4 grid on the starting screen of both PRaM and SpAM. The 8 *sport* exemplars were presented in a 2-by-4 grid; the 24 *vehicle* exemplars in a 5-by-5 grid with the right bottom corner left empty; and the 32 *bird* exemplars in a 6-by-6 grid with the right four positions on the bottom row left empty. Because the number of exemplars differed between categories, the number of exemplar pairs to judge in PRaM also differed from category to category: 28 for *sports*, 276 for *vehicles*, and 496 for *birds*.

Results

Duration

Table 5 lists the average completion time (in seconds) per combination of method and category. A Mann–Whitney test established that PRaM took longer to complete than SpAM for the two categories with the highest number of exemplars (24 and 32). On average, participants spent just over 12 minutes judging the 276 *vehicle* pairs, and just over 4 minutes arranging the 24 *vehicles* on the screen. Judging the

Table 7. Mann–Whitney test comparing the skewness of PRaM and SpAM dissimilarity data in Study 2.

		Individual proximities							Average proximities	
		PRaM		SpAM					Skewness	
Category	# exemplars	*M*	*SD*	*M*	*SD*	*W*	*p*	*r*	PRaM	SpAM
sports	8	−1.24	.78	.42	.27	5.00	<.001	−.98	−1.24	−.27
*vegetables**	16	−1.77	1.25	.36	.17	0.00	<.001	−1.00	−1.77	−1.01
vehicles	24	−1.92	1.00	.32	.19	0.00	<.001	−1.00	−1.84	−.85
birds	32	−1.57	1.91	.34	.20	2.00	<.001	−.99	−1.30	−1.06

Note. Effect size is given by the rank biserial correlation *r*. *Only half of the exemplar pairs of *vegetables* were presented in PRaM.

Table 8. Mann–Whitney test comparing the centrality of PRaM and SpAM dissimilarity data in Study 2.

		Individual proximities							Average proximities	
		PRaM		SpAM					Centrality	
Category	# exemplars	*M*	*SD*	*M*	*SD*	*W*	*p*	*r*	PRaM	SpAM
sports	8	1.62	.28	1.54	.30	336.00	.319	.17	1.75	1.25
*vegetables**	16	1.93	.42	1.58	.25	464.50	<.001	.61	1.50	1.50
vehicles	24	1.87	.24	1.54	.16	499.00	<.001	.73	1.58	1.50
birds	32	2.01	.20	1.60	.17	542.00	<.001	.88	1.50	1.69

Note. Effect size is given by the rank biserial correlation *r*. *Only half of the exemplar pairs of *vegetables* were presented in PRaM.

496 *bird* pairs took on average about 23 minutes, while organizing the 32 exemplars only took 7 minutes. That is, for these categories PRaM takes about three times as long as SpAM. When the number of category exemplars is small, as was the case for *sports* with eight exemplars, judging all exemplars pairs and arranging all exemplars take about equally long. Likewise, judging half of the pairs of a 16-exemplar category (60 instead of 120) lasts about as long as organizing the 16 exemplars. On average, both tasks took little over 3 minutes, which roughly corresponds to half of the time it took participants in Study 1 to judge all pairs, and compares to the time taken in Study 1 to organize the same exemplars spatially (see Table 1). For neither of these categories did the Mann–Whitney test indicate a significant difference in completion time between PRaM and SpAM.

Reliability

Table 6 lists the estimated reliability of the average dissimilarity data per combination of method and category. Having an equal number of participants judge all exemplar pairs of a category yields a higher reliability than having participants spatially arrange the exemplars in terms of similarity. The average reliability for PRaM across the categories *sports*, *vehicles*, and *birds* is .97 compared to .90 for SpAM. As a result, more participants need to be tested using SpAM to obtain a reliability that is comparable to that of PRaM, although with 24 participants the reliability of SpAM is already higher than the .80 threshold that is commonly used in psychology. The correlations between the average proximity data of the two methods are at the maximum level one could expect given SpAM reliabilities, with the Pearson correlation equal to .94 for *sports*, .91 for *vehicles*, and .85 for *birds*.

When participants judge only half of the exemplar pairs of a 16-exemplar category, the across-participant reliability for all 120 dissimilarity pairs is comparable to that obtained by having participants organize the 16 exemplars in terms of similarity. For the 16-exemplar category *vegetables*, PRaM reliability was .89 compared to a .86 SpAM reliability. The correlation between the average PRaM and SpAM *vegetables* data equaled .78. The Pearson correlation between the average *vegetables* data from Study 1 and Study 2 was .89 for PRaM and .83 for SpAM.

Bias

Per combination of method and category, Tables 7 and 8 respectively list the average skewness and centrality across the individual dissimilarity data sets. Mann–Whitney tests were used to establish that PRaM dissimilarity data are more negatively skewed and have a higher centrality than SpAM dissimilarity data sets, with the exception of centrality for *sports* ($p = .319$). The average skewness was negative for PRaM dissimilarity data (−1.63) and positive for SpAM dissimilarity data (.36). The average centrality was higher for PRaM dissimilarity data (1.86) than for SpAM dissimilarity data (1.57), but did not exceed the critical value of 2 in the majority of data sets (70.83% of PRaM data sets compared to 92.71% of SpAM data sets).

The skewness and centrality values for *vegetables* are comparable to those in Study 1 (see Tables 3 and 4). The average skewness values were −.95 and −1.77 for PRaM and .26 and .36 for SpAM in studies 1 and 2, respectively. The decrease in the average skewness of PRaM dissimilarities appears to be in line with a general trend for more negatively skewed dissimilarity judgments in this sample compared to that of Study 1, and is not necessarily the result of participants only judging half the exemplar pairs for this category (see below for further discussion). The average centrality values were 1.95 and 1.93 for PRaM and 1.60 and 1.58 for SpAM in studies 1 and 2, respectively. The number of *vegetable* dissimilarity sets attaining a centrality value higher than 2 was also similar in studies 1 and 2 (both 25% for PRaM, and 16.67% and 12.5% for SpAM).

Tables 7 and 8 also indicate for each category the skewness and the centrality of the average PRaM and SpAM dissimilarity data, obtained by averaging the individual dissimilarity data sets across participants. For SpAM, averaging leads to dissimilarity data with a lower skewness compared to the average skewness of the individual data, while for PRaM we observed similar skews. Across categories, the skewness of the average SpAM data was −.80 compared to an average skewness of .36 across individual SpAM data sets. For PRaM, these values measured −1.54 and −1.63. While the individual SpAM dissimilarity tended to be positively skewed, the average SpAM dissimilarity data were negatively skewed. As a result, the distributions of the average PRaM and SpAM dissimilarity data are more similar than the individual distributions, although the average PRaM data remain more negatively skewed than the average SpAM data. The results of the averaging on centrality are less consistent. The centrality of the average dissimilarity data tends to be lower than the average centrality of the individual dissimilarity data for PRAM, though less so for SpAM (but individual categories defy this pattern). Across categories, the centrality of the average SpAM data was 1.49 compared to an average skewness of 1.57 across individual SpAM data sets. For PRaM, these values measured 1.58 and 1.86. The average difference in centrality between PRaM and SpAM across categories is greater for the average centrality (.42) than for the centrality of the average (.11). This is mostly the result of the decrease in centrality for PRaM as a result of averaging being more pronounced compared to the decrease in centrality for SpAM (–.28 across categories for PRaM compared with −.08 for SpAM). None of the centrality values for the average dissimilarity data exceeds 2. A final noteworthy observation is that the average of PRaM *vegetable* data behaves similarly as the averages of the other PRaM categories: Skewness is unaffected and centrality decreases. It thus appears that having participants only judge half of the exemplar pairs does not affect the skewness or centrality of the average dissimilarity data differently compared to having participants judge all exemplar pairs.

Discussion

Together with the findings from Study 1, the results of Study 2 indicate that from 16 exemplars per category onward, SpAM constitutes a significant time gain over PRaM. Given that most conceptual categories count over 16 exemplars, it follows that SpAM will generally be the most time efficient method for obtaining conceptual similarity data. The duration ratio of PRaM vs. SpAM increases from about 2.3 with 16 exemplars (value *k* in Table 1) to about 3 for categories with 24 and 32 exemplars (Table 5). This increase was to be expected in light of the dramatic increase of exemplar pairs with category size *n*. While each of these pairs needs to be explicitly judged in PRaM, in SpAM participants can adjust *n* − *1* distances simultaneously by moving a single exemplar. With increasing sample size, the decision where to position an exemplar does become more taxing as participants need to take into account more relationships, making for a steeper than linear increase in task duration for SpAM as well. While organizing the 16 category exemplars in Study 1 took about 3 minutes to complete, organizing 32 category exemplars in Study 2 took about 7 minutes to complete. As was the case in Study 1, this increase in efficiency came at the cost of a decrease in reliability. While SpAM was much faster to complete than PRaM, it requires more participants to attain a comparable level of reliability. It should be noted, however, that while PRaM/SpAM duration ratio increased considerably with the number of category exemplars, the differences in reliability remained within limits. In terms of the speed-accuracy tradeoff, this result tips the balance in favor of SpAM for categories with a large number of exemplars. While one can estimate the overall completion time (the number of participants times average completion time) of the two methods to be comparable for a set size equal to 32, we expect SpAM to attain reliabities comparable to that of PRaM in a more time efficient manner once additional exemplars per category are considered. When the number of category exemplars was small

($n = 8$ for the category *sports*), we did not find a difference in completion time between PRaM and SpAM. A difference in reliability remained, however, which was due to the very high PRaM reliability.

Regardless of the number of exemplars per category, we found PRaM dissimilarity data to be negatively skewed and SpAM dissimilarity data to be positively skewed. Centrality was higher for PRaM than for SpAM in all categories, except the one with the smallest number of exemplars. Averaging the dissimilarity data across participants again tended to bring the distributional characteristics of the dissimilarity data resulting from the two methods closer together. While the individual SpAM data were characterized by a positive skewness, the average SpAM data were characterized by a negative skewness. Because averaging decreased the centrality of PRaM data more than it decreased the centrality of SpAM data, the average PRaM and SpAM dissimilarity data were also found to be more similar to each other than the individual dissimilarity data in terms of centrality. The resemblance of the methods' aggregate data also showed in their correlation, which approached the maximal attainable values given their reliabilities.

Taken together, the results of Study 2 are comparable to those of Study 1 and confirm that the caveats that were raised regarding SpAM by Verheyen et al. (2016) apply across categories of varying sizes. Participants were much faster to arrange the exemplars of a conceptual category according to similarity than they were to rate the similarity of all pairs of exemplars, and the difference in task duration between PRaM and SpAM increased with the number of category exemplars. Although this increase in efficiency came at the cost of a decrease in reliability, the reliability difference did not appear to change with category size, presumably making SpAM the most interesting choice in terms of the speed-accuracy tradeoff for conceptual categories with a large number of exemplars, especially in light of the observation that SpAM data always attained the commonly used .80 lower limit for reliability with 24 participants. As was the case in Study 1, we found that the spatial nature of SpAM biased the resulting dissimilarity data against feature representations. While PRaM dissimilarity data demonstrated a negative skewness in line with the known feature representational format of conceptual categories (Dry & Storms, 2009; Pruzansky et al., 1982; Tversky & Hutchinson, 1986; Verheyen et al., 2016), SpAM dissimilarity data were positively skewed, a characteristic of spatial representations (Pruzansky et al., 1982; Sattath & Tversky, 1977). Similarly, PRaM dissimilarity data demonstrated a higher centrality than SpAM data, but only a minority of the data sets attained a centrality of 2 or higher, the cutoff value that was used in previous studies to argue for feature representations (Tversky & Hutchinson, 1986). Average SpAM data appeared not to be representative of individual SpAM data in that they displayed a negative skewness, while the skewness of the individual data was positive. On the plus side, this did make aggregate PRaM and SpAM data resemble each other more, both qualitatively (in terms of distributional characteristics) and quantitatively (in terms of inter-correlation). Whereas the average PRaM and SpAM dissimilarities for *sports*, *vehicles*, and *birds* respectively correlated .94, .91, and .85, the corresponding average correlations of the individual dissimilarity ratings were .51, .42, and .26. The conclusions from Study 1 not to use SpAM for the study of individual (differences in) dissimilarity data and not to regard average SpAM data as representative for individual SpAM data, thus also applies to Study 2, generalizing this recommendation to conceptual categories of varying sizes. However, Study 2 is limited to categories with up to 32 exemplars. For categories with more exemplars, it remains to be determined whether or not the limitations of SpAM outweigh PRaM's extensive completion time and its potential ensuing detrimental effects, provided it proves at all possible to collect all pairwise ratings in a single sitting.

We found that the time it takes to obtain pairwise similarity judgments could be drastically shortened by only having participants judge half of the exemplar pairs. For the 16-exemplar category *vegetables*, the resulting completion time was comparable to that of SpAM. The difference in reliability between PRaM and SpAM was equally reduced because of this change in procedure. We believe this is due to a reduction in the reliability of PRaM data since it is only based on half of the observations. Having participants judge all exemplar pairs of 16 exemplar categories in Study 1 resulted in an average reliability of .96 across categories (.94 for *vegetables*), whereas having the participants in Study 2 only judge half of the *vegetable* pairs resulted in a reliability of .89. This change in procedure does not appear to affect the centrality and skewness values of the resulting dissimilarity data considerably. The average PRaM centrality measure was comparable in studies 1 and 2, and although the average skewness was more negative in Study 2 than it was in Study 1, we believe this to be due to a sample difference rather than the result of participants judging only half of the *vegetable* exemplar pairs. We carried out a simulation study to confirm that having participants judge only half of the pairs is not expected to affect the average skewness or centrality of the resulting dissimilarity distributions. We drew 10,000 samples from the Study 1 *vegetable* dissimilarities by randomly selecting half of each participant's ratings. This yielded an average skewness of −.94 (95% reference interval [−.92, −.89]) and an average centrality of 1.98 (95% reference interval [1.87, 2.10]). These values are comparable to the average values of −.95 and 1.95 reported in Tables 3 and 4 for the entire distribution. Having participants judge only half of the exemplar pairs was also found not to affect the skewness and centrality of the average dissimilarity data differently, compared to having participants judge all exemplar pairs. This alteration to PRaM might thus allow one to obtain pairwise ratings in a rather time efficient manner even in categories with many exemplars, especially if the percentage of pairs that is to be judged were found to be further reducible because of the additional constraints imposed by additional category exemplars.

Study 3

Study 3 intents to investigate whether some of the limitations of PRaM and SpAM that have been identified in the previous studies can be overcome. The main issue that PRaM faces seems to be the time it takes to complete, especially when the number of stimuli to compare is large. A lengthy task can have all kinds of negative effects on the quality of the resulting data, due to participants becoming tired, bored, distracted, or disengaged, and should therefore be avoided if possible. It also makes the method ill-suited for use in samples of patients, children, or elderly participants. Separating data collection across multiple occasions might not be an ideal solution to this problem, as the information that is retrieved from semantic memory is not necessarily invariant across occasions (see Verheyen et al., 2019, for an overview of studies on the probabilistic nature of the semantic retrieval process). It is therefore not guaranteed that participants will make the same consideration across data collection sessions. The results of Study 2 for the category *vegetables* seem to suggest that presenting participants with only 50% of a category's exemplar pairs is a viable strategy to improve PRaM's efficiency. It reduces the method's completion time considerably without affecting the resulting data's distributional characteristics.[12] In Study 3, we will investigate whether this finding generalizes to categories with varying numbers of exemplars.

The main problem facing SpAM is that it appears less suited to study individual (differences in) dissimilarity data. Studies 1 and 2 yielded quite comparable aggregate PRaM and SpAM data, but while the former were representative of the individual data, the latter were not. This showed in the lower reliability of SpAM data compared to PRaM data, but most notably in the distributional properties of the individual data sets. For SpAM, these properties differed both from the properties of the individual PRaM data (with SpAM data demonstrating a positive skewness and lower centrality than the negatively skewed PRaM data) *and* the average SpAM data (which were negatively skewed). Verheyen et al. (2016) speculated this may be due to participants interpreting the spatial organization task in different manners (see also Hout et al., 2013), being restricted to only communicate two out of a potentially much larger number dimensions of variation, and/or communicating additional dimensions in an idiosyncratic manner. In Study 3, we will investigate whether SpAM can also be used to obtain representative individual level data by presenting participants with multiple subsets of stimuli to organize spatially in terms of similarity. Such a procedure has been used before to allow participants to convey information beyond two dimensions or when the number of stimuli did not fit onto a single screen (e.g., Berman et al., 2014; Coburn et al., 2019; Goldstone, 1994; Horst & Hout, 2015; see also Kriegeskorte & Mur, 2012). In studies 1 and 2, we found that averaging the data from several SpAM participants yielded average dissimilarity data sets that were comparable to the average PRaM data. Does averaging multiple arrangements by a single participant yield an average dissimilarity data set that is comparable to judged individual dissimilarity data?

Participants

Forty-eight undergraduate students (42 women, 6 men), aged between 17 and 24 years old ($M = 18.77$, $SD = 1.57$), participated in Study 3. They were

[12] As for *vegetables* in Study 2, we conducted a simulation study to see whether the average skewness and centrality for random halves of the Study 2 PRaM dissimilarity distributions would be comparable to those of the entire distributions. With average skewness and centrality values of −1.00 (95% reference interval [−1.19, −.80]) and 1.90 (95% reference interval [1.74, 2.08]) for *sports*, −1.92 (95% reference interval [−2.04, −1.82]) and 1.99 (95% reference interval [1.89, 2.09]) for *vehicles*, and −1.48 (95% reference interval [−1.65, −1.34]) and 2.05 (95% reference interval [1.97, 2.13]) for *birds*, this proved to be the case except for *sports*.

Table 9. Mann–Whitney test comparing PRaM and multi-arrangement SpAM on completion time (in seconds) per category in Study 3.

		PRaM		SpAM					
Category	# exemplars	*M*	*SD*	*M*	*SD*	*W*	*p*	*r*	*k*
sports	8	106.708	33.835	251.542	97.270	23.00	<.001	−.92	.42
vegetables	16	220.583	61.572	351.000	85.676	48.00	<.001	−.83	.63
vehicles	24	481.833	160.309	579.375	209.514	205.50	.091	−.29	.83
birds	32	779.000	200.531	732.417	263.929	350.00	.205	.22	1.06

Note. Effect size is given by the rank biserial correlation *r*. The value *k* represents the duration ratio (PRaM/SpAM). Only half of the exemplars pairs were presented in PRaM. Six trials with half of the exemplars were presented in SpAM.

financially compensated for their participation at a rate of 8 euros/hour.

Materials

The materials were identical to the ones used in Study 2, that is: photorealistic images of the 8 most familiar exemplars of *sports*, the 16 most familiar exemplars of *vegetables*, the 24 most familiar exemplars of *vehicles*, and the 32 most familiar exemplars of *birds*, according to De Deyne (2014).

Procedure

As was the case in studies 1 and 2, every participant provided similarity data for two categories using PRaM and for two categories using SpAM. Participants alternated between PRaM and SpAM, half of them starting with PRaM and the other half starting with SpAM. These two orders of presenting the methods were crossed with the 24 possible orders of presenting the categories, for a total of 48 combinations. Each of these combinations was completed by one participant. The similarity tasks were preceded by an informed consent and followed by a survey intended to assess participants' experiences with both methods.

In PRaM, each participant judged a randomly selected half of the category's exemplar pairs. This reduces the number of judgments from 28, 120, 276, and 496 to 14, 60, 138, and 248 for *sports* (8 exemplars), *vegetables* (16 exemplars), *vehicles* (24 exemplars), and *birds* (32 exemplars), respectively. All exemplars were always present on the screen. The exemplars that were to be judged in terms of similarity were highlighted using black rectangles (see left panel of Figure 1). The selected exemplar pairs were highlighted in a random order. As in studies 1 and 2, participants had a nine-point Likert scale (1 = very dissimilar, 9 = very similar) at their disposal to indicate their answers.

In SpAM, we had participants organize multiple subsets of the category's exemplars in terms of similarity. We will refer to this procedure as multi-arrangement SpAM. We opted for six trials with half of a category's exemplars on screen per trial. This decision was made on practical grounds. These parameters were chosen so that the average duration of the total study would be similar to that of Study 2. We estimated that it would allow participants to complete the study within the scope of one hour. We employed a Steiner system to distribute exemplars across trials. For the categories *sports*, *vegetables*, *vehicles*, and *birds*, we thus determined six Steiner series with 4, 8, 12, and 16 stimuli each, respectively. The employed Steiner series can be found in Tables A2–A5 in the Appendix A. The six Steiner series of a category were always completed consecutively (i.e., no other task or other category intervened). The order in which the series were presented was randomized for every participant. The physical stimulus that was assigned to the stimulus number in the Steiner series was also randomized for every participant. The combination of trials and number of exemplars per trial necessitates that some exemplar pairs are repeated across trials. Because of the randomization that is in place, the particular pairings that are repeated are different across participants. For repeated pairs, the average distance across repetitions will be used in the analyses. Note that because only half of a category's exemplars are presented on a trial, multi-arrangement SpAM loses one of the attractive features of SpAM, namely that the entire stimulus range is immediately apparent to participants.

Depending on the combination of method and category, 4, 8, 12, 16, 24, or 32 stimuli were simultaneously presented on the screen. Four exemplars were presented in 2-by-2 grid, 8 exemplars in a 2-by-4 grid, 12 exemplars in a 3-by-4 grid, 16 exemplars in a 4-by-4 grid, 24 exemplars in a 5-by-5 grid with the right bottom corner left empty, and 32 exemplars in a 6-by-6 grid with the right four positions on the bottom row left empty.

Results

Duration

Table 9 lists the average completion time (in seconds) per combination of method and category. A

Table 10. Reliability of the average dissimilarity data for PRaM and across trials 1:T for SpAM T in Study 3.

Category	# exemplars	PRaM	SpAM 6	SpAM 5	SpAM 4	SpAM 3	SpAM 2	SpAM 1
Sports	8	.95	.94	.92	.89	.88	.79	.59
vegetables	16	.91	.91	.89	.86	.80	.68	NA
Vehicles	24	.95	.94	.93	.91	.88	.79	NA
Birds	32	.89	.88	.86	.83	.77	.67	NA

Note. Due to randomization only 1 value available for at least one pair so no reliability available (NA). Only half of the exemplars pairs were presented in PRaM.

Mann–Whitney test established that multi-arrangement SpAM took longer to complete than PRaM for categories with a relatively small number of exemplars ($n = 8$ for *sports*, $n = 16$ for *vegetables*). The differences in completion time for the categories with a larger number of exemplars ($n = 24$ for *vehicles*, $n = 32$ for *birds*) were not significant. The results of the significance tests and the values of the duration ratio k in Table 9 are not particularly important as they depend on the specific methodological decisions that were made (i.e., presenting half of the exemplar pairs in PRaM and presenting six trials with half of the exemplars each in multi-arrangement SpAM). They do indicate that since the completion time of PRaM and multi-arrangement SpAM was comparable for *vehicles* and *birds*, we can make a time-equated comparison of the reliability and distributional characteristics of the resulting data in the following sections.

Compared to Study 2, where all exemplar pairs instead of 50% were presented, participants on average needed little more than half of the time to complete PRaM tasks: 106.708 seconds vs. 131.708 seconds for *sports*, 481.833 vs. 741.750 for *vehicles*, and 779.000 vs 1378.542 for *birds* (see Table 5). The average completion time for *vegetables* (220.583 seconds) was also just over half of the average completion time in Study 1 (401.208 seconds), in which all *vegetable* exemplar pairs were administered (see Table 1), and comparable to that in Study 2 (193.875 seconds) where half of the *vegetable* exemplar pairs were administered in PRaM as well (see Table 5).

Sixteen *birds* had to be organized in each of the six trials of multi-arrangement SpAM. With an average completion time of 732.417 seconds, this makes for an estimated average trial duration of 122.070 seconds. This is somewhat lower than the average value of 183.417 seconds for organizing 16 *birds* in Study 1 (or any of the other categories in Study 1, which all had 16 stimuli; see Table 1). It is also quicker than the average time that was needed to organize 16 *vegetable* exemplars in Study 2 (182.542 seconds; see Table 5). A similar finding was obtained for the category *vegetables*. Eight *vegetables* had to be organized in each of the six trials of multi-arrangement SpAM. With an average completion time of 351.000 seconds, this makes for an estimated average trial duration of 58.800 seconds, which is considerable faster than the 129.375 seconds taken to organize eight *sports* in Study 2 (see Table 5)[13].

Reliability

Table 10 lists the estimated reliability of the average dissimilarity data per combination of method and category. For multi-arrangement SpAM, the cumulative reliability across trial 1 until 6 is provided.

The first thing to note is that all reliabilities for PRaM with only half of the pairs presented (Study 3) are lower than the reliabilities for PRaM with all pairs presented (.91 vs .94 for *vegetables* from Study 1, see Table 2; .95 vs .98 for *sports*, .95 vs. .98 for *vehicles*, .89 vs. .94 for *birds* from Study 2; see Table 6). This was to be expected since the average PRaM dissimilarity data are based on fewer observations, though the difference is small given that only half the amount of data was obtained.

The reliability of multi-arrangement SpAM improves with trials up to the sixth and final trial. Although this was to be expected as the Steiner system was set up such that only after the sixth trial a participant would have judged all exemplar combinations and adding trials thus involves basing the average SpAM dissimilarity data on more observations, the increase in reliability is nevertheless not a necessity. The increase suggests that even on the later trials participants are providing meaningful, non-redundant information that helps make the estimates of exemplar distances provided by other participants on earlier trials more precise. After six trials in which participants provided a distance for every exemplar pair at least once, the level of reliability attained by multi-arrangement SpAM was comparable to that of PRaM with half the number of exemplar pairs presented to

[13] Unlike the comparison for *birds*, this comparison is between categories, which might be adding to the difference. It might, for instance, be the case that *sports* are more difficult to judge than *vegetables* because they are more abstract. The average SpAM completion times in Study 1 (196.792 seconds for *sports* vs. 195.000 seconds for *vegetables*; see Table 1) in which participants arranged 16 exemplars of each category, suggest otherwise, however.

Table 11. Mann–Whitney test comparing the skewness of PRaM and multi-arrangement SpAM dissimilarity data in Study 3.

		Individual proximities							Average proximities	
		PRaM		SpAM					Skewness	
Category	# exemplars	*M*	*SD*	*M*	*SD*	*W*	*p*	*r*	PRaM	SpAM
sports	8	−1.07	.92	.20	.58	45.00	<.001	−.84	−1.43	−.27
vegetables	16	−1.22	.84	.19	.35	12.00	<.001	−.96	−1.35	−.92
vehicles	24	−1.52	.87	.33	.33	8.00	<.001	−.97	−1.58	−.52
birds	32	−1.53	.64	.23	.23	1.00	<.001	−1.00	−1.52	−1.21

Note. Effect size is given by the rank biserial correlation *r*. Only half of the exemplars pairs were presented in PRaM. Six trials with half of the exemplars were presented in multi-arrangement SpAM.

Table 12. Mann–Whitney test comparing the centrality of PRaM and multi-arrangement SpAM dissimilarity data in Study 3.

		Individual proximities							Average proximities	
		PRaM		SpAM					Centrality	
Category	# exemplars	*M*	*SD*	*M*	*SD*	*W*	*p*	*r*	PraM	SpAM
sports	8	1.83	.39	1.75	.49	343.00	.255	.19	1.75	1.75
vegetables	16	1.87	.27	1.79	.39	340.50	.283	.18	1.88	1.75
vehicles	24	2.11	.35	1.85	.26	411.00	.011	.43	1.83	1.67
birds	32	2.04	.22	1.85	.34	418.50	.007	.45	1.56	1.81

Note. Effect size is given by the rank biserial correlation *r*. Only half of the exemplars pairs were presented in PRaM. Six trials with half of the exemplars were presented in multi-arrangement SpAM.

participants. In the case of *sports* ($n = 8$) and *vegetables* ($n = 16$), this is achieved in about twice the time needed to complete PRaM. In the case of *vehicles* ($n = 24$) and *birds* ($n = 32$), this is achieved with both methods taking roughly the same amount of time to complete. From trial 4 of multi-arrangement SpAM onward, the reliability of all categories' average dissimilarity data exceeds .80.

The reliability of multi-arrangement SpAM after six trials is higher than the reliability of regular SpAM (for *sports* .95 vs. .94 in Study 2; for *vegetables* .91 vs. .83 in Study 1 and .86 in Study 2; for *vehicles* .94 vs. .91 in Study 2; for birds .88 vs. .84 in Study 2, see Tables 2 and 6).

The Pearson correlations between the average dissimilarity data of PRaM and multi-arrangement SpAM in Study 3 are .87 for *sports*, .80 for *vegetables*, .83 for *vehicles*, and .77 for *birds*. These values are below the maximum level one could expect given the reliabilities in Table 10 and are lower than the inter-method correlations observed in Study 1 (.91 average across categories) and Study 2 (.88 average across categories). The inter-study correlations between average data sets obtained with comparable methods suggest that it is the multi-arrangement SpAM data that correspond the least with the other similarity data. The correlations between data sets obtained by judging all or half of the exemplar pairs in PRaM are invariantly high: .95 for *sports* (Study 2–Study 3); .92 for *vegetables* (Study 1–Study 3); .96 for *vehicles* (Study 2–Study 3); and .92 for *birds* (Study 2–Study 3). The correlations between data sets obtained using regular (single-trial) SpAM and multi-arrangement SpAM are considerably lower: .92 for *sports* (Study 2–Study 3); .85 (Study 1–Study 3) and .81 (Study 2–Study 3) for *vegetables*; .87 for *vehicles* (Study 2–Study 3); and .74 for *birds* (Study 2–Study 3).

Bias

Per combination of method and category, Table 11 lists the average skewness across the individual dissimilarity data sets. For PRaM, these are comprised of judgments for half of the exemplar pairs. For multi-arrangement SpAM, these are comprised of all inter-exemplar distances across the six trials (the average distance is used for any repeated exemplar pairs). Mann–Whitney tests were used to establish that PRaM dissimilarity data are more negatively skewed than multi-arrangement SpAM dissimilarity data. The difference was significant at $\alpha = .0125$ for each of the four categories. Across categories, the average skewness was negative for PRaM dissimilarity data (−1.34) and positive for the multi-arrangement SpAM dissimilarity data (.24). Table 11 also indicates for each category the skewness of the average PRaM and multi-arrangement SpAM dissimilarity data, obtained by averaging the individual dissimilarity data sets across participants. Both for PRaM and multi-arrangement SpAM, the averaging leads to dissimilarity data with a lower skewness compared to the average skewness of the individual data (with the exception of PRaM for *birds*). The difference is much more pronounced for SpAM (–.73 compared to .24 across categories) than it is for PRaM (–1.47 compared to −1.34). While the individual multi-arrangement SpAM dissimilarity data tended to be positively skewed, the average multi-arrangement SpAM dissimilarity data are negatively skewed. As a result, the distributions of the average PRaM and multi-arrangement SpAM dissimilarity data are much more comparable than the distributions of the individual PRAM and multi-arrangement SpAM dissimilarity data. The average PRaM data remain more negatively skewed than the average multi-arrangement SpAM data, however.

Only presenting participants with half of the exemplar pairs to judge does not appear to affect the average skewness of the resulting PRaM dissimilarity data much. For the categories *sports*, *vehicles*, and *birds*, the average negative skewness was slightly less pronounced than in Study 2, where all exemplar pairs were judged (see Table 7). For the category *vegetables*, the average negative skewness was slightly more pronounced than in Study 1, where all *vegetable* pairs were judged (see Table 3; see also Table 7 of Study 2 for a comparable finding). None of these differences were significant according to a Mann–Whitney test (all $p > .0125$; not shown). The skewness of multi-arrangement SpAM data sets, too, was comparable to the skewness of regular SpAM data sets. For the categories *sports*, *vegetables*, and *birds*, the average centrality was smaller than in Study 2, where all exemplars were arranged on a single trial (see Table 7; see also Table 3 of Study 1 for a comparable finding for *vegetables*). For the category *vehicles,* the average centrality was slightly higher than in Study 2 (see Table 7). None of these differences were significant according to a Mann–Whitney test, however (all $p > .0125$; not shown).

Per combination of method and category, Table 12 lists the average centrality across the individual dissimilarity data sets. Mann–Whitney tests were used to establish that PRaM dissimilarity have a significantly higher centrality than multi-arrangement SpAM dissimilarity data sets for categories with many exemplars (*vehicles*: $n = 24$; *birds*: $n = 32$). The difference was not significant at $\alpha = .0125$ for the categories with a smaller number of exemplars (*sports*: $n = 8$; *vegetables*: $n = 16$). The average centrality exceeded the critical value of 2 for the categories *vehicles* and *birds* when PRaM was used. Across categories, 43.75% of individual PRaM dissimilarity data sets had a centrality value of 2 or higher, compared to 29.17% of multi-arrangement SpAM data sets. Table 12 also indicates for each category the centrality of the average PRaM and multi-arrangement SpAM dissimilarity data, obtained by averaging the individual dissimilarity data sets across participants. Both for PRaM and multi-arrangement SpAM, the averaging lead to dissimilarity data with a lower centrality compared to the average centrality of the individual data (with the exception of multi-arrangement SpAM for *sports* and PRaM for *vegetables*; note that averaging did lead to a decrease in centrality in the case of PRaM for *vegetables* in Study 2, see Table 8). The difference is more pronounced for PRaM (1.76 compared to 1.96 across categories) than it is for SpAM (1.75 compared to 1.81) and appears to be mostly driven by a decrease in PRaM centrality due to averaging in the categories with many exemplars (*vehicles* and *birds*). None of the centrality values for the average dissimilarity data exceeds 2.

Presenting participants with half of the exemplar pairs to judge increased the average centrality of the resulting PRaM dissimilarity data for the categories *sports*, *vehicles*, and *birds* compared to the average values reported in Table 8 for Study 2 in which all exemplar pairs were judged. The average centrality for *vegetables* was lower compared to the average centrality reported in Table 4 for *vegetables* in Study 1 in which all exemplars pairs were judged. (The average centrality of the Study 2 *vegetable* PRaM data – for which also half of the exemplar pairs were judged – was also lower than the corresponding value in Study 1; see Table 4). None of these differences were significant according to a Mann–Whitney test (all $p > .0125$; not shown). The centrality of the multi-arrangement SpAM data sets was higher than the centrality of regular SpAM data sets (compare the Study 3 centrality results in Table 12 for *sports*, *vehicles*, and *birds* with the Study 2 results in Table 8; also compare the Study 1 centrality result for *vegetables* in Table 4 with Study 2 and Study 3 centrality results for *vegetables* in Tables 8 and 12, respectively). In the case of *vehicles* ($W = 81.50$, $p < .001$, $r = -.717$) and *birds* ($W = 151.5$, $p = .005$, $r = -.474$), these differences were significant according to a Mann–Whitney test (all other $p > .0125$; not shown).

Discussion

Evidently, halving the number of exemplar pairs that participants had to judge in PRaM, had a comparable effect on the task duration, with participants only needing little more than half of the time to complete the task[14]. It appears that when participants engage in multi-arrangement SpAM, they tend to spend less time on an individual trial than if they were to

[14] Note that we do not expect the completion time to be exactly half since a constant time interval needed to read and process the task instructions has to be taken into account.

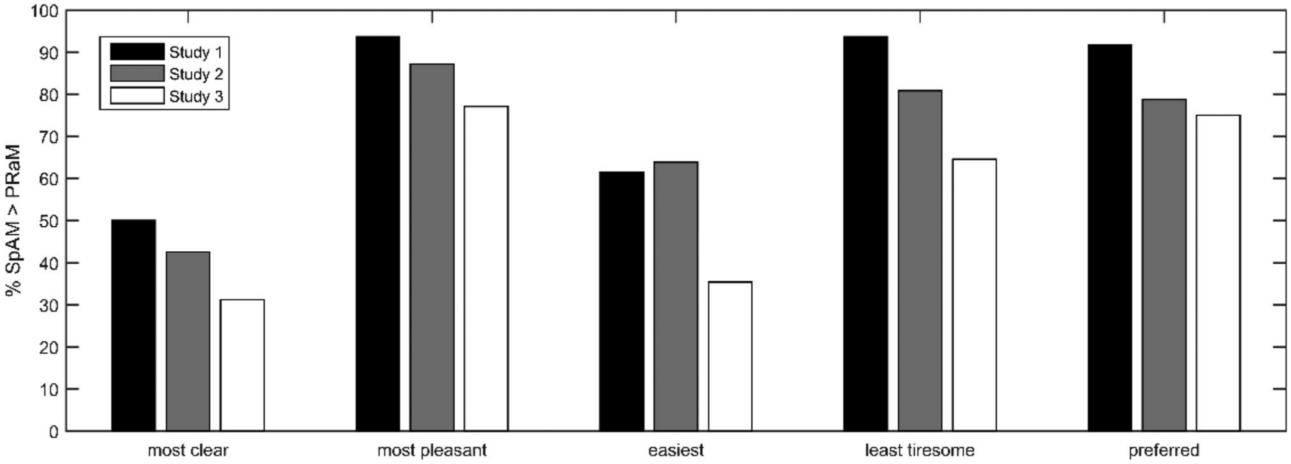

Figure 2. Percentage of participants choosing SpAM instead of PRaM for each of the survey questions. Note that one participant in Study 1 indicated that both methods were equally easy. One participant in Study 2 did not complete the survey. In Study 3, only half of the exemplars pairs were presented in PRaM, and six trials with half of the exemplars were presented in SpAM.

organize a comparable number of stimuli only once. This does not imply that participants necessarily act less deliberately, but rather that participants become more apt at arranging the exemplars because they have already figured out on earlier trials which dimensions of variation to use. The increasing reliability of multi-arrangement SpAM across trials supports this.

The reliability of multi-arrangement SpAM increased up till the sixth trial, indicating that even on the final SpAM trial participants are communicating meaningful information. The reliability level attained by multi-arrangement SpAM was comparable to the reliability level of PRaM with half of the exemplar pairs presented for judgment. If we take PRaM reliability as researchers' desired target (in light of their experience with PRaM and the effects reliability has on reproducibility and predictive ability), the speed-accuracy tradeoff changes somewhat compared to the previous studies, in that PRaM might be preferred over multi-arrangement SpAM when the number of stimuli to judge is small, and both methods are comparable in terms of duration and reliability when the number of category exemplars is large. If researchers are satisfied with a reliability of .80, then the balance might once again tip in favor of multi-arrangement SpAM, as for the larger categories this target was established after four of the six trials in our study.

We found the reliability of multi-arrangement SpAM to be consistently higher than the reliability of regular SpAM. This suggests that to increase reliability, having participants organize multiple subsets of stimuli can be used as an alternative to having additional participants complete single-trial SpAM. Here, researchers will again have to consider what they regard to be the preferred alternative based on the resources that are available to them. When they have a large number of participants at their disposal for a limited time only (e.g., when conducting the spatial arrangement method online), they might opt for single-trial SpAM, while when they have access to a limited number of participants, but for a longer time period (e.g., when running studies in the lab) they might opt for multi-arrangement SpAM.

In studies 1 and 2, we found that averaging SpAM dissimilarity data across participants yielded a negative skewness, while the individual dissimilarity data had a positive skewness. The expectation that averaging various subset arrangements of an individual participant would yield dissimilarity data with a negative skewness did not bear out. In Study 3, too, the average skewness of the individual multi-arrangement SpAM dissimilarity data was positive, and averaging across participants yielded a data set with negative skewness. One reason for this might be that when one averages across participants, one might be aggregating data from individuals who employed different considerations (Hout & Goldinger, 2016; Richie et al., 2020), while when one averages across an individual's trials in multi-arrangement SpAM, the same considerations are repeatedly used. The skewness of the individual PRaM dissimilarity data was once again found to be significantly lower than the skewness of the individual SpAM dissimilarity data.

Having participants arrange multiple subsets of stimuli did increase the centrality of the resulting dissimilarity data. For categories with a large number of exemplars, the centrality of multi-arrangement SpAM was significantly higher than the centrality of single trial SpAM. The resulting centrality values were nevertheless still significantly lower than those of PRaM. While the average centrality of PRaM for the categories *vehicles* ($n = 24$) and *birds* ($n = 32$) surpassed the critical value of 2, the average centrality of multi-arrangement SpAM did not. Note that the skewness and centrality differences for *vehicles* and *birds* hold despite them being time-equated, excluding the possibility that time on task is responsible for the differences.

The inter-method correlations of the average dissimilarity data suggest that multi-arrangement SpAM yields somewhat different results than the other methods. The correlations between the average dissimilarity data obtained with multi-arrangement SpAM and the average dissimilarity data obtained with other methods are the smallest among the inter-method and inter-study correlations. The distributional characteristics of the average dissimilarity data do not immediately give an indication of why this might be the case. When the skewness and centrality of multi-arrangement SpAM were found to differ from regular, single-trial SpAM, they tended to be more in line with the values of PRaM. One reason for the departure of multi-arrangement SpAM might be that we merely aggregated the distances from multiple trials, averaging the distances of repeated pairs without any rescaling in case these distances differed from trial to trial.[15]

In conclusion, having participants arrange multiple subsets of exemplars does not do away with the spatial bias and lack of representativity of SpAM. The distributional characteristics of multi-arrangement SpAM dissimilarity data still reflect the spatial nature of the task. Averaging multi-arrangement SpAM data across participants does away with this bias to some extent, but yields a distribution that is not representative of the individual dissimilarity distributions. Our advice not to use SpAM for individual data collection therefore extends to multi-arrangement SpAM unless the number of stimuli is prohibitively large for PRaM. When might multi-arrangement SpAM be preferred over regular, single-trial SpAM? When aggregate data need to be obtained for categories with many exemplars that cannot be presented simultaneously on a single screen, or when one wants to obtain a higher reliability with a limited number of participants, multi-arrangement SpAM can be useful. The distributional characteristics of the average dissimilarity data of both SpAM methods are comparable. Since the distributional characteristics of PRaM with 100% or 50% of the exemplar pairs judged are comparable as well, having participants only judge half of the stimulus pairs is a sensible manner to speed up data collection, both for individual- and aggregate-level analysis. One does need to take into consideration that the reliability will be lower, although this decrease might in practice not outweigh the considerable time gain.

Survey responses

In this section, we discuss the results of the closed survey questions that were filled out once participants had completed PRaM and SpAM for two categories each. Participants made a binary choice between PRaM and SpAM, indicating which method they found to be most clear, most pleasant, the easiest, and the most tiresome. They also indicated which method they would prefer to use if they were to repeat the study. Figure 2 represents the percentage of participants in studies 1, 2, and 3 (out of 48 participants per study) who chose SpAM over PRaM in response to each question. Responses are presented as ‘least tiresome’ instead of ‘most tiresome’ so that all questions in Figure 2 have a favorable connotation. In what follows, we will use the responses to the survey’s open questions to understand the quantitative results.

It is clear from Figure 2 that most participants found SpAM more pleasant and less tiresome than PRaM. These findings support the claim by Hout et al. (2013) that SpAM is more user-friendly than PRaM, which also shows clearly in participants’ preference for SpAM over PRaM in a dichotomous choice and in participants’ responses to the open survey questions. Twenty-nine participants in Study 1 and 23 participants in Study 2 indicated the duration of PRaM to be a disadvantage, often adding that it made the task boring (#17) or tiring (#11), making them lose concentration (#10) or disengage altogether (#9). In the listed advantages of SpAM, participants in studies 1 and 2 only referred to the pleasant nature of the task 15 times. They mostly praised the comprehensiveness of the resulting arrangement (#33) and the freedom they experienced in completing the task (#33), including the ability to incrementally build a configuration and overturn earlier decisions, thus highlighting the interactive nature of the task.

Although the results for pleasantness, tiresomeness, and preference hold across studies, we do observe a

[15] But note that such a consideration also holds for averaging SpAM distances across participants or even for averaging PRaM ratings across participants. A particular distance or rating does not need to have the same meaning across participants. We return to this issue in the *General Discussion*.

decrease from Study 1 to Study 2 and from Study 2 to Study 3. We believe the former decrease is mostly due to participants finding it challenging to organize many exemplars on the screen. Whereas participants in the Study 1 SpAM tasks had to spatially organize 16 exemplars per category, participants in Study 2 had to organize either 8 (*sports*), 16 (*vegetables*), 24 (*vehicles*), or 32 (*birds*) exemplars in terms of similarity. The participants' responses to the open questions of the survey support this conjecture. Among the disadvantages of SpAM, 17 participants in Study 2 directly or indirectly referred to the challenge of arranging many exemplars on the screen, either by pointing to the difficulty of arranging many exemplars in a limited space (#9), by saying that it is challenging to get all distances simultaneously right (#7), or by indicating that the task becomes tedious (#1). Only 11 participants in Study 1 referred to these issues among the disadvantages of SpAM. Participants in Study 1 each completed 240 pairwise judgments. Participants in Study 2 on average completed 430 pairwise judgments. We believe it unlikely that the decrease in SpAM appreciation from Study 1 to Study 2 is due to participants having to judge more exemplar pairs as this would intuitively add to the boredom and unpleasantness of PRaM. We cannot rule out, however, that the decrease is due to the increased duration of SpAM in Study 2 because of the additional category exemplars. Whereas 10 participants in Study 1 mentioned the speed of the method as one of its advantages, only 5 participants in Study 2 did. This corresponds with the observation that all categories were judged faster using SpAM than PRaM in Study 1, while this only held for the categories with 24 and 32 exemplars in Study 2.

The decrease in SpAM's appreciation from Study 2 to Study 3 does not appear to be the exclusive result of the longer duration of multi-arrangement SpAM used in Study 3. Whereas participants in Study 2 had to spatially arrange *n* category exemplars, participants in Study 3 had to organize *6 * n/2* category exemplars. Despite this difference, only six participants in Study 3 referred to the shorter duration of PRaM compared to multi-arrangement SpAM among the advantages of PRaM. Nineteen participants still indicated the disadvantages of PRaM to include its repetitive and lengthy nature. Only two participants mentioned the length and the repetitive nature of multi-arrangement SpAM as a disadvantage.[16] As was the case for studies 1 and 2, participants in Study 3 praised SpAM for its pleasantness (#11), the comprehensiveness of the resulting arrangement (#8), and the interactive, flexible nature of the task (#10). Participants in Study 3 on average completed 230 pairwise judgments. This is comparable to the 240 pairwise judgments participants in Study 1 provided. We therefore believe that the changes to PRaM are not the main driver of the decrease in SpAM's appreciation across studies in Figure 2. Rather, we believe an explanation can also be found in the results for clarity and ease in Figure 2.

A small majority of participants found SpAM to be easier than PRaM in studies 1 and 2, but most participants in Study 3 found providing pairwise judgments easier than providing multiple spatial arrangements. Participants in studies 1 and 2 were divided as to which method was most clear, but a majority of participants in Study 3 agreed that PRaM was clearer than multi-arrangement SpAM. Among the disadvantages of SpAM, 11 participants in Study 1, 9 participants in Study 2, and 17 participants in Study 3 explicitly used the word 'difficult', mostly referring to the difficulty of mapping perceived dissimilarities to distances (#8 in Study 3) or deciding which of the variety of dimensions to use (#5 in Study 3). This feature of SpAM might have been more salient for participants in Study 3, who had to arrange multiple sets of exemplars for each category compared to the participants in studies 1 and 2, who only had to arrange a single set of exemplars per category. The former participants would have to ensure that distances are comparable across trials, and the presence of new exemplars on consecutive trials might bring other dimensions of comparison to the foreground. This experienced difficulty might explain why participants in Study 3 tended to prefer PRaM to multi-arrangement SpAM. For the sake of completion, we mention that the most frequently provided advantages of PRaM (across studies 1, 2, and 3) contrast with the recurrent disadvantages of SpAM: finding it easier to provide a number judgment (compared with a distance, which was often suggested to be less precise – contrary to claims by SpAM proponents) and finding it easier to have to focus on only two exemplars at a time (without having to take into account the various dimensions along which these exemplars compare with other exemplars).

General discussion

We investigated across three studies whether having participants provide pairwise similarity judgments and

[16] Note that we cannot be certain whether participants in the survey judged single SpAM trials or the entirety of SpAM trials for a particular category.

having participants convey pairs' dissimilarities in a spatial manner yield comparable results for conceptual categories. Similarities were obtained with the total set version of PRaM so that any differences observed with SpAM cannot be due to unfamiliarity with the complete range of (dis)similarity. Both methods display all the category exemplars simultaneously on the screen and thus afford contextualized indications of similarity. We chose to focus on conceptual categories because these are highly dimensional (Nosofsky et al., 2018; Verbeemen et al., 2007; Verheyen et al., 2007) and often require a representation in terms of features rather than in terms of the continuous dimensions of a geometric space (Dry & Storms, 2009; Pruzansky et al., 1982; Verheyen et al., 2016). They make for interesting comparison material in light of the fact that SpAM only allows stimuli to be judged along two dimensions in an explicitly spatial fashion. Our results confirm some of the limitations of PRaM (for an overview see Hout et al., 2013) and some of the caveats that have been formulated with respect to SpAM (for an overview see Verheyen et al., 2016). Our studies also indicate a number of manners in which both methods' shortcomings can be overcome.

Our results confirm that the time to complete PRaM rises sharply with the number of category exemplars. This carries the risk of inattentive responses and disengagement, as participants tend to find PRaM tiresome. This issue can be accommodated to some extent by having participants only judge a subset of the stimulus pairs. When only 50% of stimulus pairs are rated by every participant, the task duration almost halves. While this hardly affects the nature of the dissimilarity data, the reliability decreases as fewer observations per pair are collected. Future work could experiment with percentages smaller than 50% to explore the limits of the extent to which PRaM's efficiency can be improved without compromising data quality. While large stimulus sets impose additional constraints on the similarity relationships (Hout et al., 2018) and might therefore afford that fewer than 50% of pairs are judged, PRaM with pairs missing at random may still be unfeasible when the set size becomes very large. Researchers might then want to look into smarter ways of subset selection, such as the use of cyclic designs, a type of partially balanced incomplete block design (e.g., Burton, 2003; Spence & Domoney, 1974). While ideally the methods we use are as user-friendly as possible, we also need to ensure that we obtain the type of data we are in need of. If this requires a tiresome and/or lengthy task to be completed, we need to compensate participants appropriately or provide other incentives to convince them that their best effort is required. We have found that when experimenters take the time to carefully explain why the study is being conducted and what can be learned from it, even participants who are obliged to participate (e.g., as a mandatory course requirement) are motivated to provide high-quality data: Although participants in our studies found PRaM to be tiresome, they nevertheless provided data with a high reliability. We therefore believe that our participants were attentive throughout the better part of the task. Still, since the number of pairs to judge in PRaM dramatically increases with the set size, there must be a point at which the method is no longer feasible to use, however optimized. There is a trend in the psychological literature for larger stimulus set sizes, that is likely to spill over to the similarity measurement literature, seeing that larger set sizes would facilitate stimulus selection and decrease the chance of observed relationships being due to stimulus-specific idiosyncracies (Hout et al., 2018). When similarity measures for hundreds or even thousands of stimuli need to obtained, PRaM will have to be spread out across participants and/or sessions or – more likely – alternative methods such as SpAM will have to be used to make the data collection process feasible (see also De Deyne et al., 2018, for a promising alternative that involves ranking the most similar items to a category exemplar).

Our results also confirmed that when the number of participants is equated, SpAM produces less reliable data than PRaM. There are probably multiple explanations for this. When Goldstone (1994) and Hout et al. (2013) introduced SpAM, they already observed that participants interpreted the instructions in different ways, having them arrange the stimuli in distinct manners. Participants, for instance, order stimuli one-dimensionally or form groups of stimuli without much regard of the distances between the items in the groups and/or the distance between groups (akin to the sorting method; Borg et al., 2013). Participants might also not realize they are changing $n - 1$ distances simultaneously when moving a single stimulus. Finally, it might be the case that participants explicitly treat SpAM as a partial judgment task, in which they position the stimuli with respect to a small number of reference exemplars. If these anchor stimuli would differ between participants, this would also affect the reliability negatively.[17] Additional instructions or practice time with potential feedback could easily be used

[17] We thank an anonymous reviewer for this suggestion.

to accommodate these issues and increase the reliability[18], but this would of course be disadvantageous for the task's duration. This change could also meet the observation by several participants that SpAM is not always clear.

Another way of increasing SpAM's reliability could be to have participants arrange multiple subsets of the stimuli. We found that multi-arrangement SpAM yielded a higher reliability than regular SpAM with the same number of participants, but at the cost of a longer duration. It stands to reason that multi-arrangement SpAM could also become more taxing than single-trial SpAM, especially when the relations that need to be considered are rather complex (see Ichien et al., 2019). It is therefore important to consider alternative, more efficient ways of obtaining multiple arrangements, for instance through the use of incomplete block designs (a promising method to approximate Steiner systems is proposed in MacDonald et al., 2019) or through the adaptive selection of stimulus sets for presentation (Charest et al., 2014; Kriegeskorte & Mur, 2012). Future work should also look into alternatives for the mere averaging of data from several trials, for instance through rescaling, because the meaning of distances might be trial-dependent (Mur et al., 2013). The same distance may not represent the same dissimilarity on different trials since it is subject to the overall level of (dis)similarity of the stimuli on a given trial. We do not think this posed a major problem in our study, however, because randomly selecting half of the exemplars will generally yield a broad representation of the category (opposed to zooming in on clusters of similar stimuli as in Kriegeskorte & Mur, 2012). More simulation and empirical work is also required to determine when to stop collecting additional data. When the number of trials increases, participants might again become distracted, bored, and finally disengaged, or noise might start to become added to the data because participants exhausted the information they had to convey or deemed important for the stimuli, and start coming up with additional configurations because they are being forced to. A good stopping criterion is also needed to ensure that SpAM remains time efficient compared to PRaM for larger stimulus sets. We found that with 24 or 32 stimuli, PRaM and SpAM yielded comparable reliabilities and completion times, when only half of the pairs were presented in PRaM and six subsets – each comprising half of the stimuli – were subsequently arranged using SpAM. As the set size increases, it is therefore to be expected that SpAM too will require a larger overall time investment (either because more participants will have to arrange different subsets of stimuli, or because individual participants will each have to arrange more subsets of stimuli). An alternative that we have not considered here is having participants repeat SpAM for the entire stimulus set so that participants may convey alternative structures, involving different dimensions. This has not been attempted to our knowledge, perhaps because it might come across as less natural than having to arrange various subsets comprised of different stimuli, and might invoke a demand characteristic in that participants are explicitly asked to come up with alternative organizations than the one they had originally provided. Presenting subsets might not only yield more spontaneous arrangements, but also allows participants to continue to use their global organization, while paying attention to additional sources of variation that might become apparent because of the specific exemplars that are being contrasted. This seems to be more in line with the contextual nature of similarity (judgment) than asking participants for alternative similarity structures. Although repeating SpAM would yield more observations per stimulus pair, it remains to be seen whether it would also increase the reliability of the average data. This would require the systematic communication of additional information across repetitions and participants (e.g., a combination of 2×2 different dimensions across 2 repetitions when the stimulus domain is comprised of 4 dimensions).

Finally, our results demonstrate that the spatial nature of SpAM places constraints on the resulting dissimilarity data. Whereas PRaM yields dissimilarity data with distributional characteristics that depend on the stimulus type (Verheyen et al., 2016), SpAM dissimilarity data display a positive skewness and centrality lower than 2, regardless of the stimulus domain. PRaM thus allows researchers to capture dissimilarities that do not meet spatial constraints, while SpAM does not. It is difficult to imagine how this might be overcome as the collection of similarities in a spatial manner constitutes the essence of SpAM. In the past, SpAM's bias toward spatial representations has been brushed aside in light of its efficiency and the sizeable correlations of its average dissimilarity data with the average dissimilarity data obtained with other methods. However, for high-dimensional conceptual

[18] Hout et al. (2013) suggest that PRaM is susceptible to individual differences in scale use. Presumably, this could also be accommodated by clearer instructions with meaningful labels for the different Likert scale points, to the benefit of PRaM's reliability. Alternatively, individual data could be standardized to do away with individual differences in scale use.

stimuli, the constraints that SpAM imposes are so remote from its actual representation that it might not outbalance the efficiency considerations. We venture to say that this holds for any study with the goal to uncover the latent structure of a stimulus domain or to study individual similarity perceptions, particularly when the stimuli are not simple, perceptual ones.

Perhaps the most compelling finding from the three reported studies is that there is no task-independent manner of determining the true similarity structure of a set of stimuli (Goldstone & Medin, 1994; Verheyen et al., 2016). Each direct similarity method comes with properties that might be more or less suited for a particular application. Although no one method is suited to be adopted across all possible applications, some method might be preferred (and others to be avoided) in a particular situation. Conversely, researchers should always take into account the manner in which similarities were obtained when interpreting semantic structures. This holds for PRaM and SpAM, but also for other direct similarity methods such as free sorting, conditional ranking, and triad comparisons (see, for instance, Bijmolt & Wedel, 1995). In what follows, we conclude with an overview of the situations in which the use of PRaM and SpAM is recommended vs. to be avoided.

Conclusions

When it comes to obtaining similarity data for conceptual categories with 8 to 32 exemplars, Total-Set PRaM takes significantly more time to complete than SpAM, but yields more reliable data. SpAM appears to have a bias toward spatial representations, and the average SpAM data are not representative of the individual data. Participants judge SpAM to be less tiring, more pleasant, and the method of choice. When PRaM and SpAM data are averaged across participants, the results of both methods become more in line. By only presenting half of the pairs to judge, PRaM can be speeded up without affecting the resulting dissimilarity distributions, but with a necessary decrease in reliability as fewer observations are collected. Having participants arrange multiple subsets of the stimuli does not do away with the spatial bias of SpAM or the lack of representativity of the average SpAM data for the individual SpAM data, but does increase the reliability.

It is clear that there is no gold standard for measuring similarity. All direct similarity methods come with advantages and disadvantages, which should be weighed against each other prior to data collection. Our recommendation is not to use SpAM when one is interested in individuals' similarities or individual differences in similarity perception. At the aggregate level, PRaM and SpAM yield results that are reasonably comparable in terms of distributional characteristics, but not in terms of duration and reliability. The choice between the methods will therefore ultimately depend on the number and nature of stimuli that need to be related, the available resources (the number of participants and time per participant), and on the research purposes. When the number of stimuli is large, time per participant is limited, and/or the stimuli are simple perceptual ones, SpAM provides a reasonable choice, especially if one has plenty of participants at one's disposal. When the goal is data exploration, the testing of structural hypotheses, or the study of individual differences, PRaM is to be preferred (especially when the stimuli are non-perceptual), provided the number of stimuli is not too large.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by the KU Leuven Research Council under Grant C14/16032.

Role of the funders/sponsors: None of the funders or sponsors of this research had any 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.

Acknowledgments: SV conceptualized the study and analyzed the data. SV and AW gathered the data. SV drafted the manuscript. AW and GS provided critical revisions. All three authors discussed the findings thoroughly, read, and approved the final version of the manuscript. The authors would like to thank Britt Wuyts for her help with data collection, and Simon De Deyne for making his data available. The data that support the findings of this study are openly available on the Open Science Framework at https://osf.io/jg3xm/. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the author's institution or KU Leuven is not intended and should not be inferred.

ORCID

Steven Verheyen http://orcid.org/0000-0002-6778-6744

References

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. *Psychological Science*, *5*(3), 144–151. https://doi.org/10.1111/j.1467-9280.1994.tb00651.x

Ballester, J., Dacremont, C., Le Fur, Y., & Etiévant, P. (2005). The role of olfaction in the elaboration and use of the Chardonnay wine concept. *Food Quality and Preference*, *16*(4), 351–359. https://doi.org/10.1016/j.foodqual.2004.06.001

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

Bijmolt, T., & Wedel, M. (1995). The effects of alternative methods of collecting similarity data for multidimensional scaling. *International Journal of Research in Marketing*, *12*(4), 363–371. https://doi.org/10.1016/0167-8116(95)00012-7

Bocci, L., & Vichi, M. (2011). The K-INDSCAL model for heterogeneous three-way dissimilarity data. *Psychometrika*, *76*, 691–714. https://doi.org/10.1007/s11336-011-9225-5

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling*. Springer. https://doi.org/10.1007/0-387-28981-x

Borg, I., Groenen, P. J. F., & Mair, P. (2013). *Applied multidimensional scaling*. Springer. https://doi.org/10.1007/978-3-319-73471-2

Burton, M. L. (2003). Too many questions? The uses of incomplete cyclic designs for paired comparisons. *Field Methods*, *15*(2), 115–130. https://doi.org/10.1177/1525822X03015002001

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, *111*(40), 14565–14570. https://doi.org/10.1073/pnas.1402594111

Coburn, A., Kardan, O., Kotabe, H., Steinberg, J., Hout, M. C., Robbins, A., MacDonald, J., Hayn-Leichsenring, G., & Berman, M. (2019). Psychological responses to natural patterns in architecture. *Journal of Environmental Psychology*, *62*, 133–145. https://doi.org/10.1016/j.jenvp.2019.02.007

Coltheart, V., & Evans, J. S. (1981). An investigation of semantic memory in individuals. *Memory & Cognition*, *9*(5), 524–532. https://doi.org/10.3758/bf03202346

De Deyne, S. (2014). *Dutch complexity, familiarity, mental agreement, and typicality norms for photo-realistic images from 9 semantic categories*. Unpublished manuscript.

De Deyne, S., Perfors, A., & Navarro, D. J. (2016). Predicting human similarity judgments with distributional models: The value of word associations. *Proceeding of the 26th International Conference on Computational Linguistics* (pp. 1861–1870). The COLING 2016 Organizing Committee.

De Deyne, S., Perfors, A., & Navarro, D. J. (2018). Learning word meaning with little means: An investigation into the inferential capacity of paradigmatic information. *Proceeding of the 35th Annual Conference of the Cognitive Science Society* (pp. 1608–1613). Cognitive Science Society.

Dry, M. J., & Storms, G. (2009). Similar, but not the same: A comparison of the utility of directly rated and feature-based similarity measures for generating spatial models of conceptual data. *Behavior Research Methods*, *41*(3), 889–900. https://doi.org/10.3758/BRM.41.3.889

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. MIT Press. https://doi.org/10.7551/mitpress/2076.001.0001

Ghose, S. (1998). Distance representations of consumer perceptions: Evaluating appropriateness by using diagnostics. *Journal of Marketing Research*, *35*(2), 137–153. https://doi.org/10.2307/3151843

Giordano, B. L., Guastavino, C., Murphy, E., Ogg, M., Smith, B. K., & McAdams, S. (2011). Comparison of methods for collecting and modeling dissimilarity data: Applications to complex sound stimuli. *Multivariate Behavioral Research*, *46*(5), 779–811. https://doi.org/10.1080/00273171.2011.606748

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*(4), 381–386. https://doi.org/10.3758/BF03204653

Goldstone, R. L., & Medin, D. L. (1994). The time course of comparison. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *20*(1), 29–50. https://doi.org/10.1037//0278-7393.20.1.29

Goldstone, R. L., Medin, D. L., & Halberstadt, J. (1997). Similarity in context. *Memory & Cognition*, *25*(2), 237–255. https://doi.org/10.3758/bf03201115

Green, P. E., & Wind, Y. (1973). *Multivariate decisions in marketing: A measurement approach*. Dryden.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695. https://doi.org/10.1162/COLI_a_00237

Horst, J. S., & Hout, M. C. (2015). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*, 1393–1409. https://doi.org/10.3758/s13428-015-0647-3

Hout, M. C., Cunningham, C. A., Robbins, A., & MacDonald, J. (2018). Simulating the fidelity of data for large stimulus set sizes and variable dimension estimation in multidimensional scaling. *SAGE Open*, *8*(2), 215824401877314. https://doi.org/10.1177/2158244018773143

Hout, M. C., & Goldinger, S. D. (2016). SpAM is convenient, but also satisfying: Reply to Verheyen et al. (2016). *Journal of Experimental Psychology: General*, *145*(3), 383–387. https://doi.org/10.1037/xge0000144

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

Experimental Psychology: General, *142*(1), 256–281. https://doi.org/10.1037/a0028860

Humphreys, G. W., & Forde, E. M. E. (2001). Hierarchies, similarity, and interactivity in object recognition: "Category-specific" neuropsychological deficits. *Behavioral and Brain Sciences*, *24*(3), 453–509. https://doi.org/10.1017/S0140525X01004150

Humphreys, G. W., Riddoch, M. J., & Quinlan, P. T. (1988). Cascade processes in picture identification. *Cognitive Neuropsychology*, *5*(1), 67–103. https://doi.org/10.1080/02643298808252927

Hutchinson, J. W., & Lockhead, G. R. (1977). Similarity as distance: A structural principle for semantic memory. *Journal of Experimental Psychology: Human Learning & Memory*, *3*(6), 660–678. https://doi.org/10.1037//0278-7393.3.6.660

Ichien, N., Lu, H., & Holyoak, K. J. (2019). Individual differences in judging similarity between semantic relations. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 464–470). Cognitive Science Society.

James, W. (1980). *Principles of psychology* (Vol. 1). Holt. https://doi.org/10.1037/10538-000

JASP Team. (2019). JASP (Version 0.9.2.0) [Computer software]. https://jasp-stats.org/

King, D. L., & Atef-Vahid, M.-K. (1986). Two extensions of the anchor-range effect. *Perception & Psychophysics*, *39*, 96–104. https://doi.org/10.3758/bf03211491

Koch, A., Speckmann, F., & Unkelbach, C. (2020). Q-SpAM: How to efficiently measure similarity in online research. *Sociological Methods and Research*, 1–23. https://doi.org/10.1177/0049124120914937

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*, 245. https://doi.org/10.3389/fpsyg.2012.00245

Lee, M. D., & Pope, K. J. (2003). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical Psychology*, *47*(1), 32–46. https://doi.org/10.1016/S0022-2496(02)00019-6

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

MacDonald, J. A., Hout, M. C., & Schmidt, J. (2019). An algorithm to minimize the number of blocks in incomplete block designs. *Behavior Research Methods*, *52*, 1459–1468. https://doi.org/10.3758/s13428-019-01326-x

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254–278. https://doi.org/10.1037//0033-295X.100.2.254

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology*, *32*(1), 49–96. https://doi.org/10.1006/cogp.1997.0645

Migo, E. M., Montaldi, D., & Mayes, A. R. (2013). A visual object stimulus database with standardized similarity information. *Behavior Research Methods*, *45*, 344–354. https://doi.org/10.3758/s13428-012-0255-4

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, *4*, 128. https://doi.org/10.3389/fpsyg.2013.00128

Nakatsuji, N., Ihara, H., Seno, T., & Ito, H. (2016). Visualizing similarity of appearance by arrangement of cards. *Frontiers in Psychology*, *7*, 698. https://doi.org/10.3389/fpsyg.2016.00698

Navarro, D. J., & Lee, M. D. (2004). Common and distinctive features in stimulus representation: A modified version of the contrast model. *Psychonomic Bulletin & Review*, *11*(6), 961–974. https://doi.org/10.3758/bf03196728

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54–65. https://doi.org/10.1037/0278-7393.14.1.54

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*(1), 25–53. https://doi.org/10.1146/annurev.psych.43.1.25

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018). Toward the development of a feature-space representation for a complex, natural-category domain. *Behavior Research Methods*, *50*, 530–556. https://doi.org/10.3758/s13428-017-0884-8

Okada, K., & Lee, M. D. (2016). A Bayesian approach to modeling group and individual differences in multidimensional scaling. *Journal of Mathematical Psychology*, *70*, 35–44. https://doi.org/10.1016/j.jmp.2015.12.005

Pruzansky, S., Tversky, A., & Carroll, J. D. (1982). Spatial versus tree representations of proximity data. *Psychometrika*, *47*(1), 3–24. https://doi.org/10.1007/BF02293848

Richie, R., White, B., Bhatia, S., & Hout, M. C. (2020). The spatial arrangement method of measuring similarity can capture high-dimensional, semantic structures. *Behavior Research Methods*, *52*, 1906–1928. https://doi.org/10.3758/s13428-020-01362-y

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, *42*(3), 319–345. https://doi.org/10.1007/BF02293654

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime User's Guide*. Psychology Software Tools Inc.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Shepard, R. N. (2004). How a cognitive psychologist came to seek universal laws. *Psychonomic Bulletin & Review*, *11*, 1–23. https://doi.org/10.3758/bf03206455

Shoben, E. J. (1983). Applications of multidimensional scaling in cognitive psychology. *Applied Psychological Measurement*, *7*(4), 473–490. https://doi.org/10.1177/014662168300700406

Spence, I., & Domoney, D. W. (1974). Single subject incomplete designs for nonmetric multidimensional scaling. *Psychometrika*, *39*(4), 469–490. https://doi.org/10.1007/BF02291669

Sturidsson, K., Långström, N., Grann, M., Sjöstedt, G., Åsgård, U., & Aghede, E. M. (2006). Using multidimensional scaling for the analysis of sexual offence behaviour: A replication and some cautionary notes. *Psychology,* *Crime & Law*, *12*(3), 221–230. https://doi.org/10.1080/10683160500126227

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

Tversky, A. (1977). Features of similarity. *Psychological Review*, *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. *Psychological Review*, *93*(1), 3–22. https://doi.org/10.1037//0033-295X.93.1.3

Tsogo, L., Masson, M., & Bardot, A. (2000). Multidimensional scaling methods for many object-sets: A review. *Multivariate Behavioral Research*, *35*(3), 307–319. https://doi.org/10.1207/S15327906MBR3503_02

Verbeemen, T., Vanpaemel, W., Pattyn, S., Storms, G., & Verguts, T. (2007). Beyond exemplars and prototypes as memory representations of natural concepts: A clustering approach. *Journal of Memory and Language*, *56*(4), 537–554. https://doi.org/10.1016/j.jml.2006.09.006

Verheyen, S., Ameel, E., & Storms, G. (2007). Determining the dimensionality in spatial respresentations of semantic concepts. *Behavior Research Methods*, *39*(3), 427–438. https://doi.org/10.1037/e527352012-176 https://doi.org/10.3758/BF03193012

Verheyen, S., Droeshout, E., & Storms, G. (2019). Age-related degree and criteria differences in semantic categorization. *Journal of Cognition*, *2*(1), 17. https://doi.org/10.5334/joc.74

Verheyen, S., & Peterson, M. (2020). Can we use conceptual spaces to model moral principles? *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-020-00495-5

Verheyen, S., & Storms, G. (2011). Towards a categorization-based model of similarity. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 614–619). Cognitive Science Society.

Verheyen, S., & Storms, G. (2020). Whether the pairwise rating method and the spatial arrangement method yield comparable dimensionalities depends on the dimensionality choice procedure. Manuscript submitted for publication.

Verheyen, S., Voorspoels, W., Vanpaemel, W., & Storms, G. (2016). Caveats for the spatial arrangement method: Comment on Hout, Goldinger, and Ferguson (2013). *Journal of Experimental Psychology. General*, *145*(3), 376–382. https://doi.org/10.1037/a0039758

Verheyen, S., White, A., & Egré, P. (2019). Revealing criterial vagueness in inconsistencies. *Open Mind: Discoveries in Cognitive Science*, *3*, 41–51. https://doi.org/10.1162/opmi_a_00025

Voorspoels, W., Storms, G., Longenecker, J., Verheyen, S., Weinberger, D. R., & Elvevåg, B. (2014). Deriving semantic structure from category fluency: Clustering techniques and their pitfalls. *Cortex*, *55*, 130–147. https://doi.org/10.1016/j.cortex.2013.09.006

White, A., Voorspoels, W., Storms, G., & Verheyen, S. (2014). Problems of reliability and validity with similarity derived from category fluency. *Psychiatry Research*, *220*(3), 1125–1130. https://doi.org/10.1016/j.psychres.2014.10.001

Young, F. W., & Cliff, N. (1972). Interactive scaling with individual subjects. *Psychometrika*, *37*(4), 385–415. https://doi.org/10.1007/BF02291217

Appendix A

Table A1. *Overview of the exemplars per category in decreasing order of familiarity.*

exemplar id	*Sports*	*vegetables*	*vehicles*	*birds*
1	cycling	tomato	car	pigeon
2	billiards	potato	train	rooster
3	badminton	carrot	bicycle	chicken
4	soccer	bell pepper	truck	duck
5	swimming	cucumber	plane	blackbird
6	tennis	lettuce	metro	swan
7	running	onion	bus	sparrow
8	chess	cauliflower	jeep	peacock
9	volleyball	broccoli	tram	magpie
10	judo	mushroom	dirt bike	gull
11	basketball	zucchini	tractor	stork
12	table tennis	spinach	van	penguin
13	hiking	green beans	motorcycle	pheasant
14	long jump	parsley	taxi	crow
15	squash	chicory	helicopter	ostrich
16	horse riding	corn	air balloon	tit
17			caravan	parrot
18			skateboard	parakeet
19			sled	flamingo
20			boat	owl
21			scooter	turkey
22			go cart	canary
23			rocket	robin
24			submarine	swallow
25				woodpecker
26				eagle
27				heron
28				falcon
29				toucan
30				vulture
31				cuckoo
32				pelican

Table A2. *Steiner system used in Study 3 for multi-arrangement SpAM of sports (n = 8). Each column corresponds to a trial, while the entries indicate the stimuli that are presented on each trial.*

Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
1	1	2	4	2	1
2	5	3	6	3	4
3	6	5	7	6	5
4	7	8	8	7	8

Table A3. *Steiner system used in Study 3 for multi-arrangement SpAM of vegetables (n = 16). Each column corresponds to a trial, while the entries indicate the stimuli that are presented on each trial.*

Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
1	1	1	2	2	6
2	6	9	3	3	7
3	7	10	4	4	8
4	8	11	5	5	12
5	9	12	9	12	13
6	10	13	10	13	14
7	11	14	11	14	15
8	16	15	16	15	16

Table A4. *Steiner system used in Study 3 for multi-arrangement SpAM of vehicles (n = 24). Each column corresponds to a trial, while the entries indicate the stimuli that are presented on each trial.*

Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
1	1	1	2	2	8
2	8	13	3	3	9
3	9	14	4	4	10
4	10	15	5	5	11
5	11	16	6	6	12
6	12	17	7	7	18
7	13	18	13	18	19
8	14	19	14	19	20
9	15	20	15	20	21
10	16	21	16	21	22
11	17	22	17	22	23
12	24	23	24	23	24

Table A5. *Steiner system used in Study 3 for multi-arrangement SpAM of birds (n = 32). Each column corresponds to a trial, while the entries indicate the stimuli that are presented on each trial.*

Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
1	1	1	2	2	10
2	10	17	3	3	11
3	11	18	4	4	12
4	12	19	5	5	13
5	13	20	6	6	14
6	14	21	7	7	15
7	15	22	8	8	16
8	16	23	9	9	24
9	17	24	17	24	25
10	18	25	18	25	26
11	19	26	19	26	27
12	20	27	20	27	28
13	21	28	21	28	29
14	22	29	22	29	30
15	23	30	23	30	31
16	32	31	32	31	32