



Visual and semantic similarity norms for a photographic image stimulus set containing recognizable objects, animals and scenes

Zhuohan Jiang^{1,2} · D. Merika W. Sanders^{3,4} · Rosemary A. Cowell³

Accepted: 21 October 2021
© The Psychonomic Society, Inc. 2022

Abstract

We collected visual and semantic similarity norms for a set of photographic images comprising 120 recognizable objects/animals and 120 indoor/outdoor scenes. Human observers rated the similarity of pairs of images within four categories of stimuli—inanimate objects, animals, indoor scenes and outdoor scenes—via Amazon's Mechanical Turk. We performed multidimensional scaling (MDS) on the collected similarity ratings to visualize the perceived similarity for each image category, for both visual and semantic ratings. The MDS solutions revealed the expected similarity relationships between images within each category, along with intuitively sensible differences between visual and semantic similarity relationships for each category. Stress tests performed on the MDS solutions indicated that the MDS analyses captured meaningful levels of variance in the similarity data. These stimuli, associated norms and naming data are made available to all researchers, and should provide a useful resource for researchers of vision, memory and conceptual knowledge wishing to run experiments using well-parameterized stimulus sets.

Keywords Stimulus norms · Visual similarity · Semantic similarity · Database · Multidimensional scaling

Introduction

Visual perception allows humans to understand their immediate environment, navigate the world safely and build conceptual knowledge. Arguably the most complex and important feat of visual perception is the recognition of objects and scenes. Whether attempting to find something in a crowded closet or trying to orient oneself in an unfamiliar forest, the ability to perceive and interpret a wide variety of real-world

objects and scenes is crucial for completing daily tasks and for survival. In keeping with the central role of object and scene perception in human existence, there is an abundance of psychological research that uses object and scene stimuli to investigate the perceptual and cognitive processes underlying these abilities (Ashby, Prinzmetal, Ivry, & Maddox, 1996; Duncan & Humphreys, 1989; Jiang, Lee, Asaad, & Remington, 2015; Tresch, Sinnamon, & Seamon, 1993).

For lower-level visual perception, studies frequently involve precise and quantitative manipulation of the degree of similarity between different visual stimuli. Much has been learned about how individual neurons or even whole cortical regions represent simple stimulus properties, by presenting stimuli that vary systematically in edge orientation (e.g., Hubel & Wiesel, 1962, 1968), color (e.g., Brouwer & Heeger, 2009; Dow & Gouras, 1973; Solomon & Lennie, 2007), or motion direction (e.g., Geisler, Albrecht, Crane, & Stern, 2001; Rodman & Albright, 1989). Phenomena such as attention and perceptual learning have also been fruitfully investigated by exploiting well-defined stimulus-similarity continua (e.g., Treisman, 1991; Yang & Maunsell, 2004). Such research has been possible because low-level visual image properties like edge orientation, color and motion are easy to define and quantify.

Zhuohan Jiang and D. Merika W. Sanders contributed equally to this work.

✉ Rosemary A. Cowell
rcowell@umass.edu

- ¹ Neuroscience Program, Smith College, Northampton, MA, USA
- ² Integrated Program in Neuroscience, McGill University, Montreal, Quebec, Canada
- ³ Department of Psychological and Brain Sciences, University of Massachusetts Amherst, 135 Hicks Way, Amherst, MA 01003, USA
- ⁴ Department of Psychology, Harvard University, Cambridge, MA, USA

Analogously to lower-level vision research, our understanding of high-level vision can be greatly advanced by using stimuli that vary systematically along quantifiable dimensions relating to high-level properties of objects and scenes. However, this is difficult to do in practice, because it is much harder to define and measure these high-level properties. We cannot simply decompose objects and scenes into their low-level image features and assume that these features capture the critical essence of the representations driving object and scene recognition, much less that they reliably capture the conceptual content of the stimuli (Martin, Douglas, Newsome, Man, & Barense, 2018; McClelland & Rogers, 2003; but see Greene & Oliva, 2009; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Oliva & Torralba, 2001 for evidence that scene recognition may in part depend on summary statistics). Moreover, when a research question requires knowledge of the perceptual similarity relationships between stimuli, mathematical measures of stimulus properties have been found not to predict human perceptions of similarity in a straightforward way, for both low-level and high-level stimulus properties (Busey, 1998; Cheung, 2016; Hopper, Finklea, Winkelman, & Huber, 2014; Larkey & Markman, 2005; Li, Liang, Lee, & Barense, 2020; Robertson, 1977; Schurgin, Wixted, & Brady, 2020).

A solution often adopted by researchers wishing to understand the perceptual similarity evoked by low-level stimulus properties is to gather empirical ratings of subjective similarity. In such studies, participants typically use a numerical scale to report how similar they perceive two items to be, repeatedly for many pairs of items (e.g., Busey, 1998; Caramazza, Hersh, & Torgerson, 1976; Hopper et al., 2014; Li et al., 2020). Some studies employ alternative methods such as requiring participants to “drag and drop” stimuli into screen locations that reflect perceived similarity relationships (Kriegeskorte & Mur, 2012). Most studies of perceived similarity allow the authors to use some variant of multidimensional scaling (MDS) to construct a perceptual similarity map of the evaluated stimulus set, in which items are laid out in a two-dimensional (2D) array with locations that best reflect the perceived “distances” between them.

Many existing stimulus databases provide color images of objects and scenes, for example, the Massive Memory Picture Database (Brady, Konkle, Alvarez, & Oliva, 2008; Konkle, Brady, Alvarez, & Oliva, 2010), the Amsterdam Library of Object Images (Geusebroek, Burghouts, & Smeulders, 2005), the revised Snodgrass and Vanderwart object pictorial set (Rossion & Pourtois, 2004; Snodgrass & Vanderwart, 1980), the Bank of Standardized Stimuli (Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010), the SUN database of scene images (Xiao, Ehinger, Hays, Torralba, & Oliva, 2014; Xiao, Hays, Ehinger, Oliva, & Torralba, 2010), and the Nature Scene

Collection (Geisler & Perry, 2011). However, we are aware of no publicly available scene image database that is associated with published empirical similarity ratings, and only a few such object image sets (Hout, Goldinger, & Brady, 2014; Migo, Montaldi, & Mayes, 2013). Importantly, the object similarity studies we found report perceptual similarity for object images within the same basic-level category (e.g., the similarity relationships between a set of butterfly images) but not between objects with different basic-level identities (e.g., perceived similarity between a tiger and a lion, or a tiger and a whale). Yet, the perceived similarity between images from different basic-level categories of objects or scenes may be an important research design consideration. As one example, memory researchers may wish to systematically investigate the effects of stimulus-relatedness on memory for visual images. More generally, many research questions within visual cognition may require experimental designs that group items into sets that minimize visual or semantic interference either within or between sets; for this, stimulus-relatedness must be known.

The goal of the present study was to provide object and scene image databases that are accompanied by naming data, empirical similarity ratings, both visual and semantic, and similarity maps constructed from those ratings. We collated four stimulus sets: inanimate objects, animals, indoor scenes and outdoor scenes, each containing 60 images (240 images in total). All stimulus images were color photographs. We collected ratings of visual and semantic similarity within each image category and performed MDS to analyze the similarity ratings data. MDS is an exploratory data analysis technique that allows spatial visualization of the similarity relationships between the items in a set (Mugavin, 2008); it has been used extensively in studies of perception (Jaworska & Chupetlovska-Anastasova, 2009). The similarity ratings are used to construct a matrix from which a spatial map of the relationships between all items is derived, usually in two or three dimensions. A short distance between two items in the map means that the two items are similar, whereas a long distance means that they are dissimilar. We used MDS to visualize the relationships between the images in each stimulus set, for both visual and semantic perceived similarity. We make the stimuli (along with source attributions where available), naming data, ratings data and similarity maps available at <https://osf.io/smk25/>, so that researchers may use these stimuli in any experiments for which stimulus-relatedness information is useful.

Methods

Participants

A total of **765** participants were recruited via Amazon’s Mechanical Turk (MTurk). Previous research has shown

Table 1 Summary of similarity rating participant sample size

Image category	Similarity rating	Recruited	Eliminated	Analyzed	Mean <i>n</i> ratings
Animal	Visual	81	11	70	10
	Semantic	81	11	70	10
Inanimate object	Visual	87	17	70	10
	Semantic	87	17	70	10
Indoor scene	Visual	115	45	70	10
	Semantic	97	26	71	10.1
Outdoor scene	Visual	105	35	70	10
	Semantic	112	37	75	10.7

Participant sample sizes by category and rating type. Values in “Recruited,” “Eliminated” and “Analyzed” columns are the number of participants that reached each stage of data collection. “Mean *n* ratings” reflects the average number of participants, across HITs, who rated any given image pair within an image category and similarity rating type combination. For example, on average, **10** participants rated each of the image pairs within the animal-visual group. A very small number of collected ratings were “mis-clicks” with the mouse, which resulted in no recorded numerical value (110 out of a total of 141,600 ratings, or 0.078% of the collected data).

that MTurk workers are more representative of adults in the United States than in-person recruited participants (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011), and in visual cognition tasks, the data collected from MTurk have been shown to closely match data collected in a lab environment (Brady & Alvarez, 2011; Brady, Shafer-Skelton, & Alvarez, 2017; Brady & Tenenbaum, 2013). All participants were located in the United States and were over 18 years old. For each of the four image categories (i.e., animals, inanimate objects, indoor scenes, outdoor scenes), we created all possible image pairs (1770 pairs) and randomized those pairs into seven groups (see Procedures). Each group of images was bundled into a single Human Intelligence Task (HIT) on MTurk. For each HIT, ratings were collected from at least **10** participants (see Table 1). A participant was allowed to complete ratings for more than one HIT, with the constraint that no two HITs on the same stimulus set could be performed by the same participant. All participants provided informed consent before starting the experiment and were compensated \$2 for their participation. A total of 199 participants were excluded from the data analysis because they failed to pass the attention test that was embedded as part of the experimental design (see Procedures) or because they entered the same response for all questions (a response that “passed” the attention tests, but was inappropriate for most other questions). Previous work indicates that an exclusion rate of 26% is not unusual for MTurk studies employing similar screening methods (Thomas & Clifford, 2017).

To collect naming data, an additional 14 participants were recruited from the University of Massachusetts-Amherst community and the community of authors D.M.W.S. and R.A.C. and were compensated \$15 per hour of participation.

Stimulus materials

The stimulus image set contained 240 unique images in four categories: 60 inanimate objects, 60 animals, 60 indoor scenes and 60 outdoor scenes. Inanimate objects (e.g., printer) and animals (e.g., zebra) were color photographs taken mostly from the online databases of Konkle and Oliva (<https://konklab.fas.harvard.edu/#>; Konkle & Caramazza, 2013; Konkle & Oliva, 2012). We replaced any images that were deemed too low-resolution with higher-resolution photographs of the same object or animal taken from an internet search. All objects and animals were placed on white backgrounds and resized to a standard size, slightly smaller than the size of the background (600×600 pixels), to eliminate any large differences in image size. We aimed for a relatively heterogeneous set of animals and set of inanimate objects (e.g., including both large and small, with a range of animal taxonomic orders and a range of object functions), to span a range of stimuli that are visually and semantically distinguishable for the purposes of memory studies.

Scene images, both indoor (e.g., living room) and outdoor (e.g., volcano) were also color photographs. Most of the scene pictures were taken from Ross, Sadil, Wilson and Cowell (2018), but several images from this original set were removed and replaced with another item in cases where an image was deemed too semantically related to another image in the set (e.g., tennis court and volleyball court). The final sets of scene images (both indoor and outdoor) contained no image pairs with overlapping context (e.g., there were no two beaches or two tennis courts). All scene images were then resized to a standard size (600×800 pixels) to eliminate any large differences in image size. Again, for each category of scenes we aimed to find a relatively heterogeneous set of

A

How **visually** similar are the following two pictures? Please only make judgements based on how much the pictures **look alike**.

Please rate the images on a scale of 1 to 5. A rating of 1 means the pictures are not similar at all and a rating of 5 means the pictures are highly similar.



- ☐ 1-not similar at all
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5-highly similar

B

How **semantically** similar are the following two pictures? Please only make judgements based on how much the pictures **have to do with each other**. Pictures are judged to be related when the featured items (a) are used together or are mutually complementary (like hammer and nail), (b) serve the same purpose (glass and cup), (c) often occur in the same situation or environment (lancet and stethoscope), or (d) are exemplars of the same basic category (like a pickup truck and a limousine).

Please rate the images on a scale of 1 to 5. A rating of 1 means the pictures are not similar at all and a rating of 5 means the pictures are highly similar.



- ☐ 1-not similar at all
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5-highly similar

C

How **semantically** similar are the following two pictures? Please only make judgements based on how much the pictures **have to do with each other**. Pictures are judged to be related when the featured items (a) serve the same purpose (like baseball fields), or (b) are exemplars of the same basic category (like beaches).

Please rate the images on a scale of 1 to 5. A rating of 1 means the pictures are not similar at all and a rating of 5 means the pictures are highly similar.



- ☐ 1-not similar at all
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5-highly similar

Fig. 1 MTurk similarity rating trial examples. Participants were asked to rate either the visual (**a**) or semantic (**b, c**) similarity between two images presented on the screen. Instructions for rating the visual similarity were identical for all four image categories; however, instruc-

tions for rating the semantic similarity differed by the examples given in the second sentence for animals/objects (**b**) versus indoor/outdoor scenes (**c**).

exemplars that would be reasonably distinguishable from each other and elicit distinct, unique names.

Procedures

For similarity ratings, for each category of 60 images (animals, inanimate objects, indoor scenes and outdoor scenes), there were 1770 possible pairs of images (“60 choose 2” combinations, in mathematical terms). The 1770 pairs were divided into seven groups: six groups of 253 trials, plus one group of 252 trials. The 1770 image pairs were assigned randomly to the seven groups, to avoid any systematic ordering such as, for example, all image pairs that contain a printer being assigned to Group 1 within inanimate objects. An additional 10 attention-check trials were embedded within each image-pair group to track whether participants were paying attention across the trials. Attention-check trials presented two identical images, such that only a “highly similar” response (“5”) would be appropriate. The 10 attention-check trials were randomly interleaved between the regular trials in each image-pair group, yielding a total of 263 (or 262) trials within each image-pair group. Each set of 263 (or 262) trials comprising a group of image pairs and 10 attention-check trials was described as a “task” or a “HIT” on Amazon MTurk. Each HIT had a 20-minute time limit. A very small number of HITs were collected after Amazon MTurk instituted a change to the maximum allowed length of a HIT, such that these HITs were constrained to contain fewer trials than in our original design. To collect this last batch of data, we conducted more HITs with fewer trials per HIT, and this is reflected in the indoor semantic and outdoor semantic rows of Table 1.

Within each HIT, participants saw a series of trials in which they were asked to rate the similarity between two pictures presented on the screen, based on either semantic or visual properties of the images (Fig. 1). For visual similarity ratings, participants were asked “How visually similar are the following two pictures? Please only make judgments based on how much the pictures look alike.”¹ For semantic similarity ratings, participants were asked “How semantically similar are the following two pictures? Please only

make judgments based on how much the pictures have to do with each other.” Ratings were made using a numerical scale from 1, not similar at all, to 5, highly similar. Participants responded by checking the box of the corresponding value. After a box was checked, the next trial was presented. After finishing all trials in the HIT, participants submitted their ratings by clicking the submit button.

Each worker was assigned a unique ID number by Amazon MTurk, which enabled restrictions on how participants were allowed to complete multiple HITs. There were four stimulus sets (animals, inanimate objects, indoor scenes and outdoor scenes) and two rating types (visual, semantic). No participant was allowed to rate a given pair of images twice, meaning participants (1) could not complete a particular HIT more than once (say, the HIT containing image pairs #1-253 for “inanimate object-visual” ratings) and (2) could not complete both kinds of ratings for a particular set of image pairs (say, the HIT containing image pairs #1-253 for “inanimate object-visual” ratings and the HIT containing the same pairs for “inanimate object-semantic” ratings). An additional constraint (imposed incidentally for convenience) is that participants who had completed ratings of one type (either visual or semantic) for a set of image pairs in a given category were unable to complete ratings of the other type for any set of image pairs in that category. Outside of these constraints, participants were allowed to complete multiple HITs, always providing ratings for stimulus pairings that they had not previously been exposed to.

We collected at least 10 ratings per image pair (Table 1). To assess whether a dataset of size 10 ratings is adequate to provide a stable MDS solution, we conducted simulation analyses. In brief, we sampled subsets of the full dataset that steadily increased in size and asked how much the addition of one extra rating changed the MDS solution at each dataset size (see [Supplementary Information](#)). The simulation results for the

Table 2 Similarity rating descriptive statistics

Image category	Similarity rating	Mean ratings	Standard deviation
Animal	Visual	1.53	0.88
	Semantic	1.76	1.06
Inanimate object	Visual	1.27	0.63
	Semantic	1.37	0.80
Indoor scene	Visual	1.60	0.92
	Semantic	1.62	1.09
Outdoor scene	Visual	1.47	0.81
	Semantic	1.63	1.07

¹ We did not explicitly instruct participants to exclude any prior visual knowledge of the real-world objects that the pictures depict, such as size. This was intentional, because we assume that in most studies employing the images, no such instructions to exclude real-world visual knowledge will be issued. Thus, we expect visual ratings to reflect an “organic” mixture of prior visual knowledge (e.g., a hippopotamus usually looks much larger than a crab) and current visual input (e.g., in which the hippopotamus and crab pictures project the same size image onto the retina), which will most accurately index the similarity of visual representations formed in response to these images in later studies.

Table 3 Inter-rater reliability as measured by Kendall's coefficient of concordance (W)

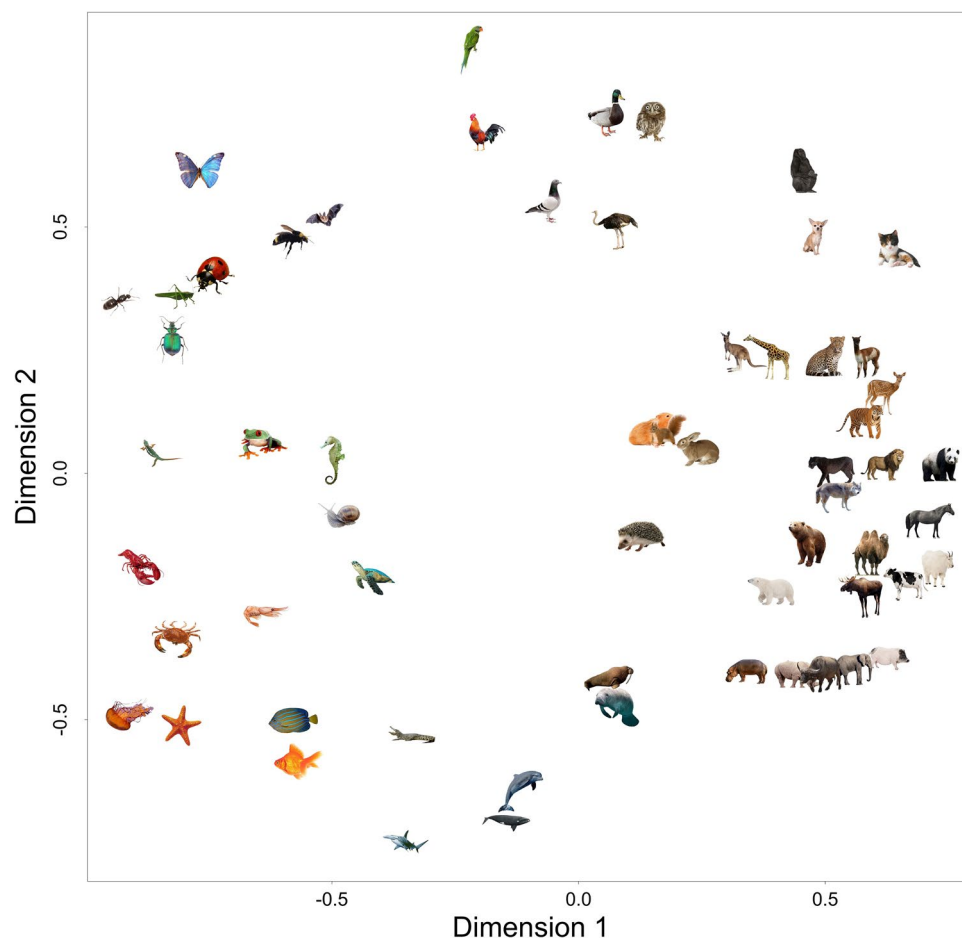
Image category	Similarity rating	HIT 1	HIT 2	HIT 3	HIT 4	HIT 5	HIT 6	HIT 7	Mean
Animal	Visual	.46	.46	.39	.39	.41	.47	.44	.43
	Semantic	.51	.51	.42	.49	.40	.56	.38	.46
Inanimate object	Visual	.39	.22	.33	.31	.31	.36	.27	.38
	Semantic	.28	.18	.30	.26	.31	.27	.30	.27
Indoor scene	Visual	.35	.27	.27	.37	.25	.33	.23	.30
	Semantic	.30	.16	.21	.35	.36	.22	.29	.27
Outdoor scene	Visual	.44	.30	.31	.41	.32	.28	.26	.33
	Semantic	.39	.33	.34	.30	.34	.27	.24	.32

Kendall's coefficient of concordance (W) by image category, similarity rating and HIT. HITs 1 through 6 had 10 raters and 253 image pairs; HIT 7 had 10 raters but only 252 image pairs. The mean value in the last column reflects Kendall's W for each category and rating type, averaged across the seven HITs.

2D case are presented in Supplementary Figure 1, and they imply that the MDS solutions would not be greatly modified by continuing to collect more ratings beyond 10 per image pair (other dimensionalities are also presented

in the Open Science Framework (OSF) project at <https://osf.io/smk25/>).

Following the online collection of similarity ratings, we obtained naming data from 14 in-person

**Fig. 2** Visualization of the 2D MDS solution for animal-visual similarity ratings

participants. Each of the 240 images was presented across eight blocks of 30 trials. Blocks alternated between two broad image categories—animal/inanimate objects and indoor/outdoor scenes—and the category of the initial block was counterbalanced across participants. We informed participants that they would be naming animate and inanimate objects, and indoor and outdoor scenes, and we requested that they provide specific and identifiable labels, avoiding more generic identifiers such as “wild animal” or “store.” Image presentation was self-paced, such that the next image did not appear until participants submitted their response for the current image. Formal or quantitative analyses of these data were not of primary interest, but the raw data and “counts” of supplied names for each image have been made available at the OSF repository. Not only do these data provide normative names for all of the stimuli, but the “spread” of names for each image offers an additional qualitative assessment of whether it was consistently identified.

Statistical analysis

Multidimensional scaling (MDS) Data analyses were performed using R (R Core Team, 2016) and the *smacof* package for multidimensional scaling (de Leeuw & Mair, 2009). For all stimulus sets (inanimate objects, animals, indoor scenes, outdoor scenes), we used the similarity ratings to create distance matrices for visual and semantic ratings separately, yielding eight matrices (four stimulus sets by two rating types). The similarity ratings were transformed into distance ratings by subtracting the original pairwise similarity ratings from 5, which yielded perceptual distance values ranging from 0 to 4 (e.g., a similarity level of 1 becomes a “perceptual distance” of 4; a similarity level of 5 becomes a “perceptual distance” of 0). Next, we averaged these perceptual distance values across participants and used the resulting average distance matrix for each category-rating combination as the input for non-metric MDS analysis (Kruskal & Wish, 1978).

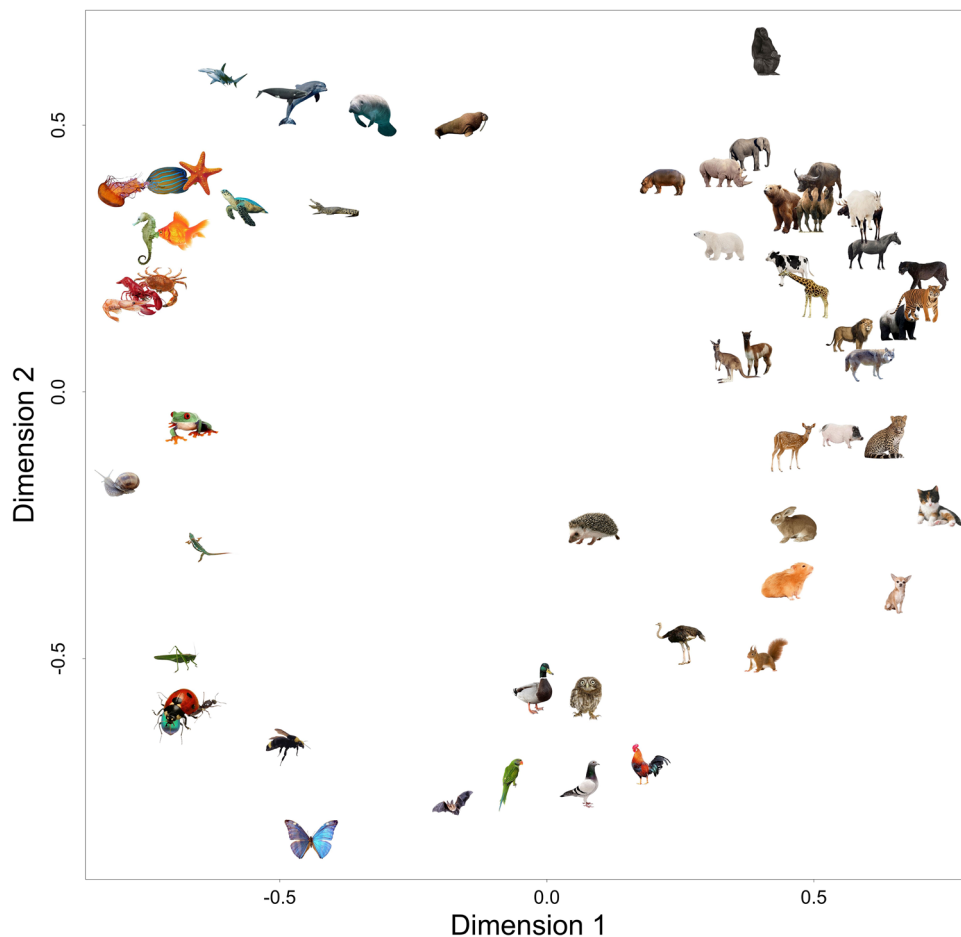


Fig. 3 Visualization of the 2D MDS solution for animal-semantic similarity ratings

Dimensionality of the MDS space and scree plots of stress level

Because MDS involves a reduction in dimensionality of the original data, it can be helpful to find the optimal dimensionality for representing the data. The optimal dimensionality of an MDS solution is usually determined by a scree plot, which displays the stress level of an MDS solution at each level of dimensionality assayed. Stress is the value that is optimized in finding the MDS solution, and it provides a measure of the mismatch between the Euclidean distances between each pair of stimuli in the MDS solution and the corresponding distances in the empirical data (Kruskal & Wish, 1978). A higher stress value indicates a poorer fit between the distances in the empirical dataset and the distances in the MDS solution, and stress is invariably higher for MDS solutions of lower dimensionality. However, in choosing the dimensionality there is a trade-off: although stress decreases as the number of dimensions increases, it is harder to interpret the data visualization offered by MDS in higher dimensions.

For each of the eight category-rating combinations (e.g., inanimate objects-visual), we calculated six different MDS solutions with dimensionalities from 2 to 7. We created scree plots showing the normalized stress level (Kruskal & Wish, 1978; Takane, Young, & de Leeuw, 1977) at each dimensionality of MDS solution, for each category-rating combination. To conveniently visualize stimulus-relatedness, we constructed maps based upon the 2D MDS solutions for each category-rating combination.

Monte Carlo simulations to determine the validity of MDS solutions

Although stress level provides a measure of how well an MDS solution fits the raw distance data, low stress level per se is not a guarantee that the MDS solution has found a meaningful interpretation of the similarity relationships in the empirical data. Therefore, we performed a Monte Carlo simulation, sampling random similarity ratings, to test how the stress levels of MDS solutions generated from random ratings compare

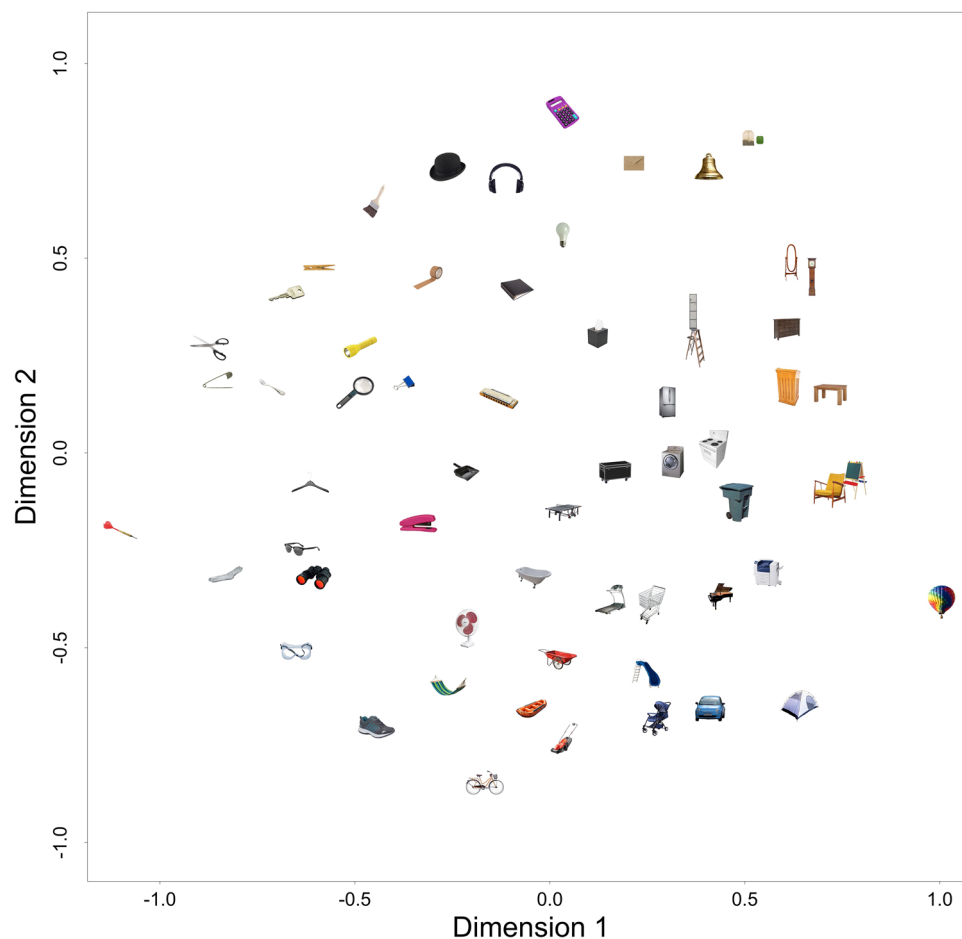


Fig. 4 Visualization of the 2D MDS solution for inanimate object-visual similarity ratings

to the stress levels we observe in the empirical MDS solutions (Hout et al., 2014). If the MDS solution for the empirical data has identified meaningful similarity relationships, then the empirical stress levels will fall in the extreme left-hand tail of the null distribution generated by the Monte Carlo simulations. To generate a Monte Carlo null distribution with characteristics that resembled the empirical distribution, but in which meaningful relationships were obliterated, ratings were randomly sampled with replacement from the pool of 141,600 empirical ratings (this total reflects 10 ratings of 1770 image pairs in each of the eight category-rating combinations). For each simulation, we sampled 10 ratings for each image pair (to simulate 10 participants) and assigned them randomly to the image pair labels. The 10 ratings per image pair were then averaged together before being entered into the MDS analysis, and the stress level of the resulting MDS solution was computed. This whole process was carried out 1000 times to construct a density plot of stress levels from

random data that contain no meaningful similarity relationships.

Results

Similarity rating descriptive statistics

The mean and standard deviation of participant similarity ratings within each condition is provided in Table 2. Participants tended to use the lower end of the rating range, corresponding to a judgment of "dissimilar." This is perhaps unsurprising since all image pairs (except on attention check trials) comprised two items from different basic-level categories (e.g., a hippo and a crab, rather than two crabs). However, violin plots of all similarity ratings shown in Supplementary Figure 3 demonstrate that participants did use the full range of ratings across all image pairs.

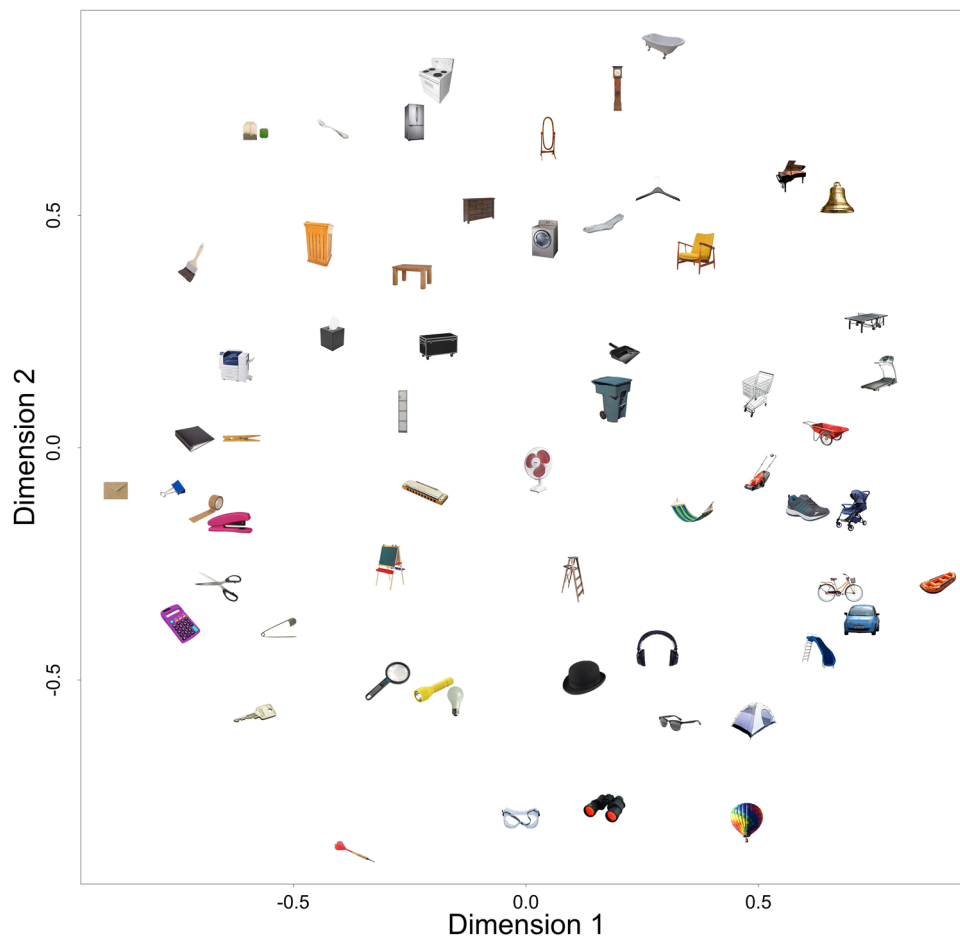


Fig. 5 Visualization of the 2D MDS solution for inanimate object-semantic similarity ratings

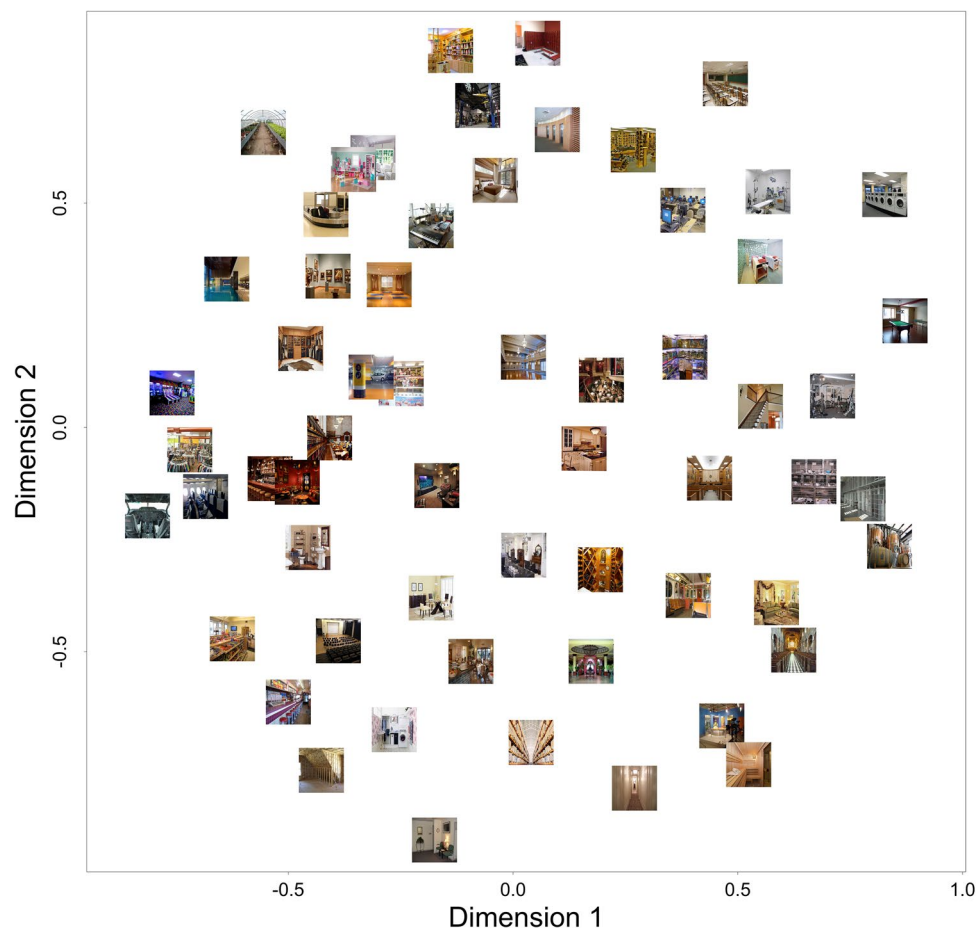


Fig. 6 Visualization of the 2D MDS solution for indoor scene-visual similarity ratings

Additionally, we assessed inter-rater reliability via Kendall's coefficient of concordance (W) across image category, similarity rating and HIT (Table 3). Kendall's W averaged over all 56 groups was .34 ($SD = .09$) and ranged between .16 and .56. Chi-square tests revealed that inter-rater reliability was statistically significant across all groups ($p < .001$), indicating that participants provided similar ratings when comparing the same image pair.

Multidimensional scaling solutions We performed MDS with dimensionalities 2 through 7 for all eight category-rating combinations (e.g., animals-visual, indoor-semantic, etc.). The full MDS solutions at all six dimensionalities for all eight category-rating combinations are provided at <https://osf.io/smk25/>. To visualize the results of MDS, we plotted the 2D solutions for all eight category-rating combinations, in Figs. 2, 3, 4, 5, 6, 7, 8 and 9. Each plot shows all 60 stimuli in the set (e.g., all animals in Fig. 2) located in a 2D map such that the physical distances between

images on the plot reflect the distances between images in the 2D MDS solution derived from the similarity ratings data. Inspection of two 2D solutions derived from the same set of images via different ratings—for example, animals rated visually and animals rated semantically in Figs. 2 and 3, respectively—reveals that participants were following instructions, and that the MDS solution identified meaningful relationships. For instance, the seahorse was rated as being close to the snail, frog and lizard in the visual ratings MDS solution, whereas it was grouped with other sea animals, such as the starfish and crab, in the semantic ratings MDS solution.

Scree plots of stress levels

Figure 10 displays scree plots showing the normalized stress level (Kruskal & Wish, 1978; Takane et al., 1977) at each dimensionality of MDS solution for each category-rating combination. The scree plots indicate, as expected, that

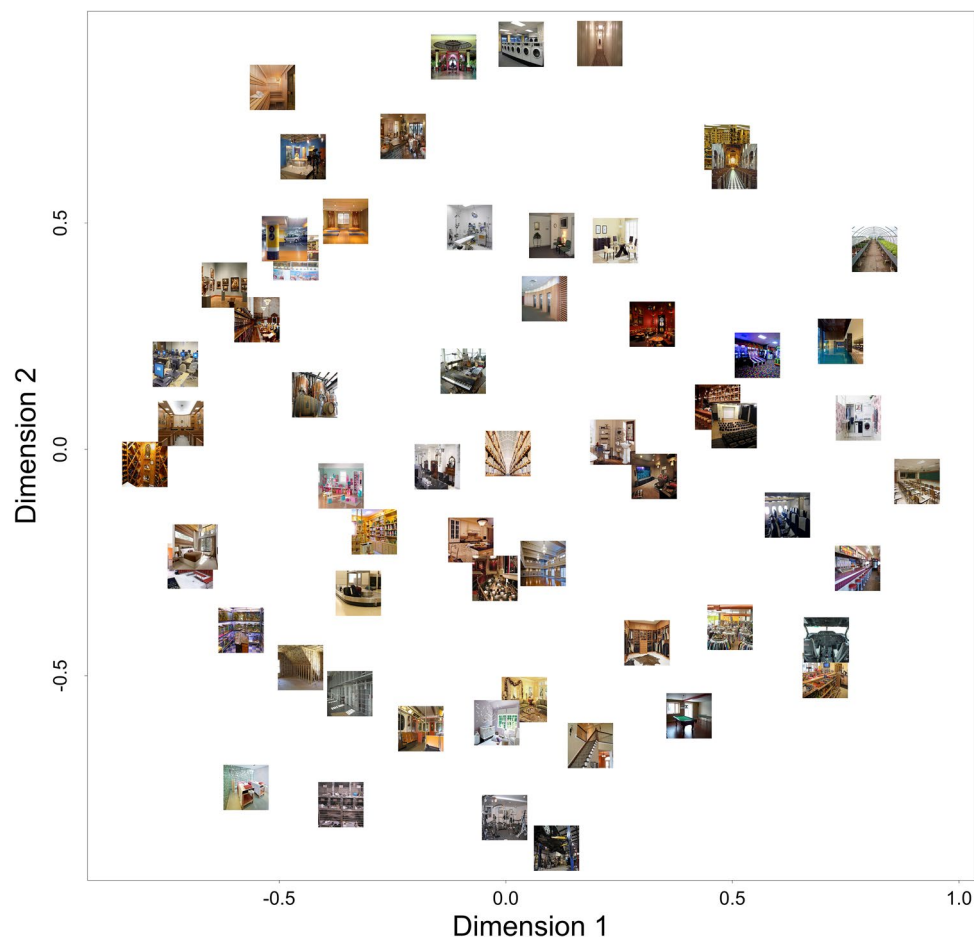


Fig. 7 Visualization of the 2D MDS solution for indoor scene-semantic similarity ratings

greater dimensionality leads to lower stress. However, for all category-rating combinations, even the 2D MDS solution has a stress level well below 0.376² (Table 4), indicating a good fit between the proximities between images in the MDS solution and the rating distances in the empirical data (Sturrock & Rocha, 2000). This suggests that the 2D visualizations for all stimulus image categories may be sufficient to represent the similarity relationships between the images. However, we nonetheless make the MDS solutions for all dimensionalities available at <https://osf.io/smk25/>.

² This cutoff reference for stress was obtained from Table 2 of Sturrock & Rocha (2000). The authors propose that random similarity matrices, which have no structure to the relationship between items, should produce a “worst-case stress value when scaled” (p. 51). Accordingly, the first percentile of a distribution of non-metric MDS stress values generated from 587,200 random similarity matrices can be used as an upper limit for stress values generated by *nonrandom, structured* matrices with the same number of items and scaled in the same number of dimensions. In the present case of 60 items and two dimensions, that first percentile cutoff value is 0.376.

Monte Carlo simulations to assess MDS solutions

We created MDS plots from randomly generated similarity ratings to produce a null distribution of stress levels for the scenario where there is no meaningful similarity information in the data. Figure 11 shows, for the case of 2D MDS solutions, the distribution of stress levels associated with MDS solutions derived from randomly sampled similarity ratings, generated via bootstrapping. In this figure, the highest stress level from among all eight category-rating conditions (namely, indoor scenes-semantic ratings, with a stress level of 0.110 for the 2D MDS solution), is displayed as a vertical red line. As seen in the figure, the stress level in our empirical MDS solution is far below the lower tail of the null distribution that was generated by sampling similarity ratings randomly. This indicates that the empirical MDS solution explains much more of the variance in the images' similarity relationships than the MDS solutions applied to randomly sampled image-pair ratings, suggesting that the empirical

MDS solution captures meaningful similarity structure not present in the randomly sampled data. The analogous simulations for MDS solutions in 3D to 7D are shown in Supplementary Figure 2.

Discussion

The goal of this study was to provide visual image stimulus sets with accompanying naming data, and quantitative similarity metrics for both visual and semantic judgments. We provide these stimulus sets and metrics for use by psychologists and cognitive scientists running empirical studies involving objects and scenes.

The similarity data were collected through MTurk via pairwise ratings, in which participants judged the similarity of two simultaneously presented stimuli. Alternative means of collecting similarity data include the spatial arrangement method (SpAM; Goldstone, 1994), which includes single- and multi-arrangement methods,

inverse MDS (Kriegeskorte & Mur, 2012), free-sorting (Coxon, 1999) and others. One disadvantage of the pairwise method is that the number of pairings increases rapidly (quadratically) with the number of stimuli in the set. In a lab environment, it is difficult to obtain enough participants, or enough ratings per participant, to collect sufficient data for all pairwise ratings. Alternative methods often require fewer trials than pairwise ratings, but each has its own disadvantages, such as precluding the discovery of similarity structures in more than two dimensions for the single-arrangement method, or limiting the data to binary similarity measures (same versus different category) for free-sorting (Kriegeskorte & Mur, 2012). Further, because each of our stimulus categories contained 60 stimuli, this would yield an unreasonable number of images for the single-arrangement SpAM method (spatially arranging by comparing each image with 59 others simultaneously). However, 60 items yield only 1770 pairs, which is a manageable number for pairwise ratings when many participants can be recruited via MTurk.

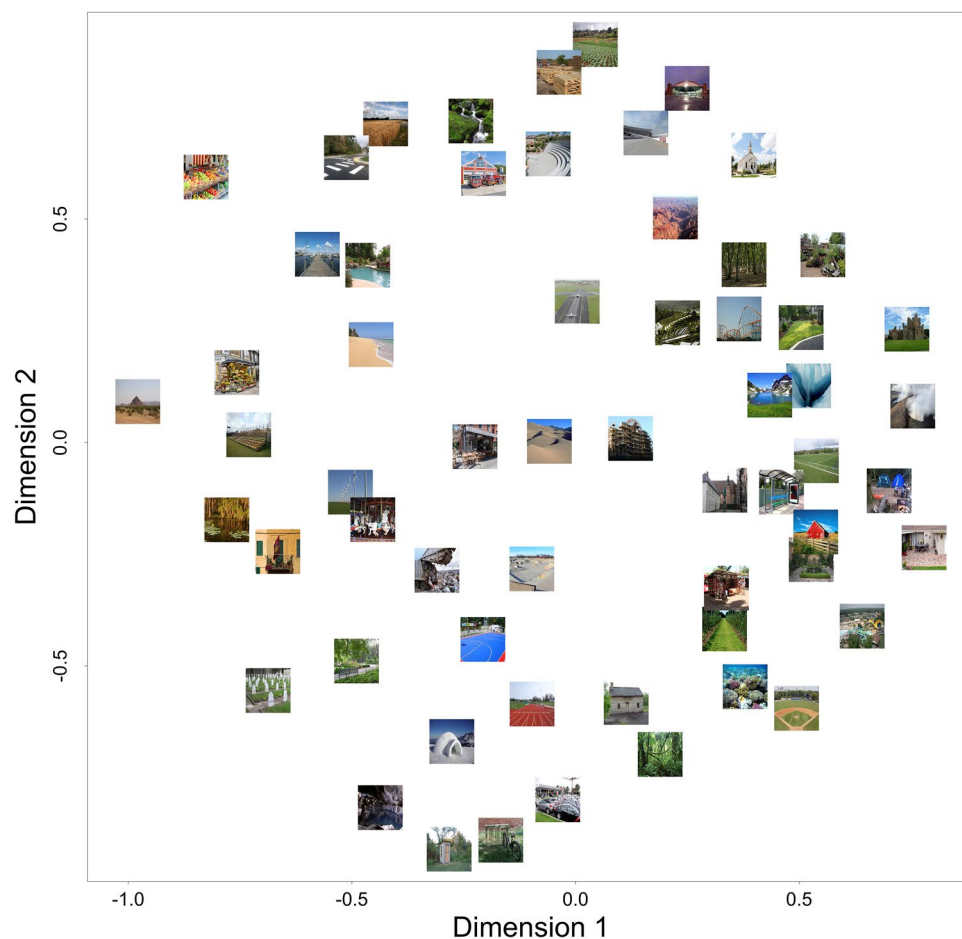


Fig. 8 Visualization of the 2D MDS solution for outdoor scene-visual similarity ratings

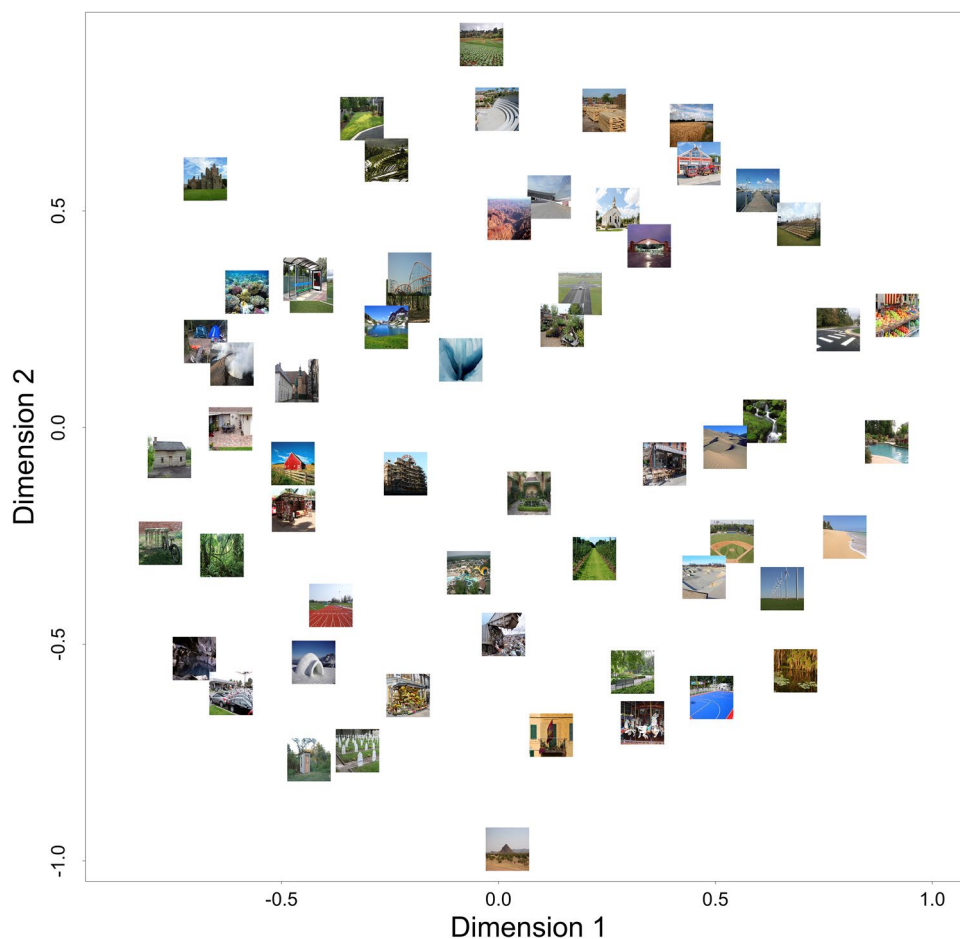
Table 4 Summary of stress levels for MDS solutions in two dimensions

Category, rating combination	Stress level
Animal, visual	0.057
Animal, semantic	0.052
Inanimate object, visual	0.082
Inanimate object, semantic	0.089
Indoor scene, visual	0.109
Indoor scene, semantic	0.110
Outdoor scene, visual	0.089
Outdoor scene, semantic	0.093
Average	0.085

Stress levels reflect the degree of correspondence between ideal item distances (produced by the MDS solution) and actual item distances (in the empirical data). A stress level at or below 0.376 indicates an excellent degree of correspondence.

We validated the MDS solutions derived from the empirical similarity ratings in two quantitative ways. First,

we verified that stress levels for all MDS solutions fell into a range typically taken to indicate good agreement between similarity relationships in the data and relationships in the MDS solution. Second, we ran Monte Carlo simulations and revealed that the distances between items in our MDS solutions corresponded to the actual distances between items in our empirical ratings data to a greater degree than would be expected from a random sampling of similarity ratings. Finally, we also validated the MDS solutions qualitatively, by visual inspection. Inspection of the 2D MDS solutions in Figs. 2 through 9 reveals that the perceptual and conceptual similarity spaces are intuitive and sensible. Further, the differences between the perceptual (visual) and conceptual (semantic) similarity maps for a given stimulus category revealed discrepancies that were entirely expected. For example, the inanimate objects were clustered in the visual condition according to color or global form, but in the semantic condition according to their function or the environment in which they are typically found (Figs. 4 and 5).

**Fig. 9** Visualization of the 2D MDS solution for outdoor scene-semantic similarity ratings

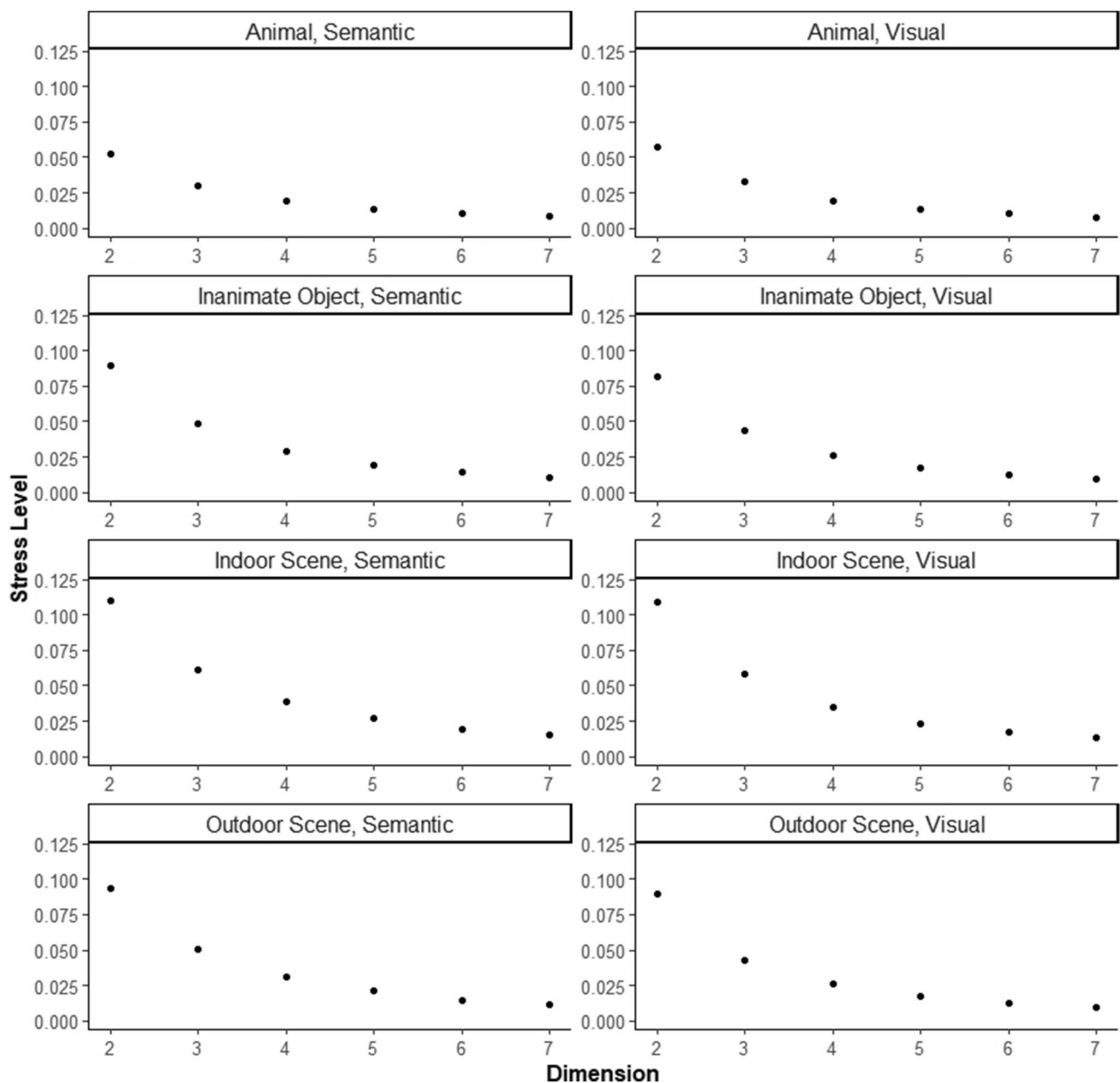


Fig. 10 Scree plots for each category-rating combination. For all eight combinations of image category (animal, inanimate object, indoor scene, outdoor scene) and similarity rating type (visual, semantic), the level of stress (y-axis) decreased as the dimensionality

of the MDS solution (x-axis) increased. However, even at two dimensions, all MDS solutions have stress levels below 0.12, indicating that they fit the empirical data well.

We also examined whether the dataset size of 10 ratings was sufficient to attain a stable MDS solution that would not change substantially with the collection of further ratings (see [Supplementary Information](#)). For this, we conducted “subsampling” analyses in which we sampled subsets of the full dataset that systematically increased in size. Our approach was to measure

the distance between pairs of MDS solutions that were created from pairs of datasets that differed in size by one rating (i.e., n ratings versus $[n+1]$ ratings). Our goal was to identify the size of dataset at which incorporating an additional rating into the dataset produced a negligible change in the MDS solution and/or “diminishing returns” in terms of eliminating any residual distance between the

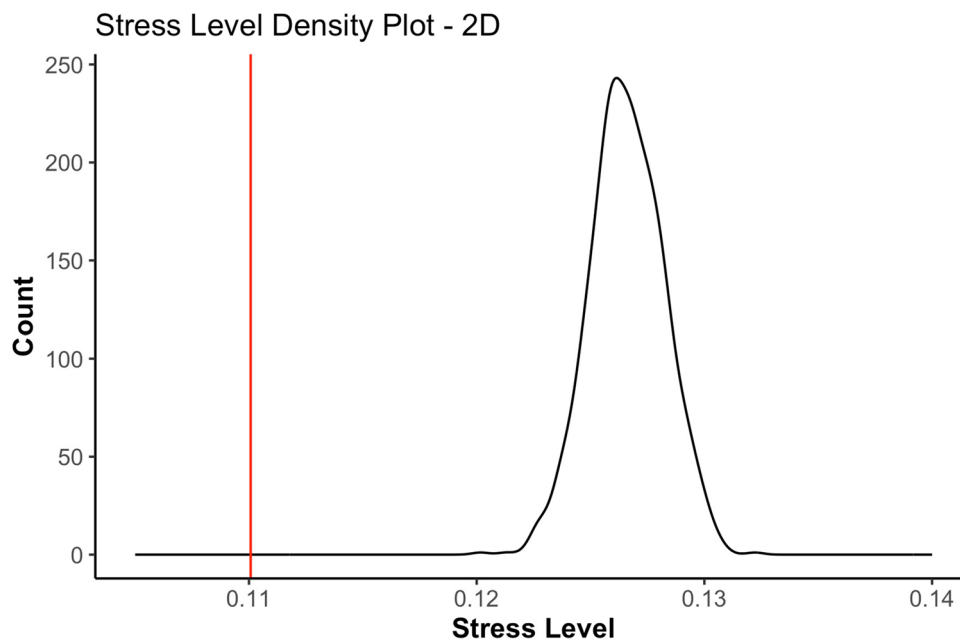


Fig. 11 Comparison between the distribution generated by Monte Carlo simulation of 2D MDS solutions and the highest stress level from any 2D MDS solution for the empirical data. The red vertical line denotes the highest empirical stress level from the 2D MDS solutions for all of our category-rating combinations: the indoor scenes-semantic rating (0.110). The line falls far below the lower-tail of the bootstrapped distribution, which was generated by repeatedly sam-

pling similarity ratings randomly from the rating data from all participants, deriving the 2D MDS solution and calculating the stress. The placement of the red vertical line indicates that the empirical MDS solution was successful in identifying meaningful similarity relationships from the data. Equivalent figures for three to seven dimensions yield the same conclusion and are shown in the [Supplementary Information](#).

MDS solutions derived from datasets that differ by one rating. As seen in Supplementary Figure 1 (the results for 2D MDS solutions) and on the OSF repository (results for 3D through 7D solutions), the extent to which the MDS solution is altered by adding an additional rating stabilizes at a low asymptote by the time the dataset reaches a size of 10 ratings, for all category-rating combinations in the 2D through 5D MDS solutions. We suggest that researchers wishing to use MDS solutions with dimensionalities 6D or 7D consult the [Supplementary Information](#) and the “Subsampling Analysis” figures at <https://osf.io/smk25/> to determine the stability of the MDS solution for the dataset of interest at the desired dimensionality. We reiterate that inter-rater reliability was high for all category-rating combinations and that, for all MDS dimensionalities, stress levels were significantly below those of MDS solutions for randomly sampled data. These additional “subsampling” analyses simply provide a guide as to which dimensionality of MDS solution may be optimally stable for a given dataset.

We make all stimuli, naming data, ratings data and the full MDS solutions for both visual and semantic ratings available to other researchers, at the following OSF website:

<https://osf.io/smk25/>. We hope that this provides a valuable set of visual stimuli for experiments involving objects and scenes in which the similarity between different instances is critical to the experimental design.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13428-021-01732-0>.

Funding This research was supported by National Science Foundation Award #1554871 to R.A.C. and by National Institutes of Health Award #1RF1MH114277-01 to R.A.C.

Declarations

Conflict of interest The authors have no known conflict of interest to disclose.

References

- Ashby, F. G., Prinzmetal, W., Ivry, R., & Maddox, W. T. (1996). A formal theory of feature binding in object perception. *Psychological Review*, 103(1), 165–192. <https://doi.org/10.1037/0033-295X.103.1.165>
- Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com’s

- Mechanical Turk. *Political Analysis*, 20(3), 351–368. <https://doi.org/10.1093/pan/mpm057>
- Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, 22(3), 384–392. <https://doi.org/10.1177/0956797610397956>
- Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38), 14325–14329. Retrieved from <https://doi.org/10.1073/pnas.0803390105>
- Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, 43(6), 1160–1176. Retrieved from <https://doi.org/10.1037/xhp0000399>
- Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, 120(1), 85–109. <https://doi.org/10.1037/a0030779>
- Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE*, 5(5), e10773. <https://doi.org/10.1371/journal.pone.0010773>
- Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *The Journal of Neuroscience*, 29(44), 13992–14003. <https://doi.org/10.1523/JNEUROSCI.3577-09.2009>
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high quality data? *Perspectives on Psychological Science*, 6(1), 3–5. <https://doi.org/10.1177/1745691610393980>
- Busey, T. A. (1998). Physical and psychological representations of faces: Evidence from morphing. *Psychological Science*, 9(6), 476–484. Retrieved from <https://doi.org/10.1111/1467-9280.00088>
- Caramazza, A., Hersh, H., & Torgerson, W. S. (1976). Subjective structures and operations in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 15(1), 103–117. [https://doi.org/10.1016/S0022-5371\(76\)90011-6](https://doi.org/10.1016/S0022-5371(76)90011-6)
- Cheung, V. (2016). Uniform Color Spaces. In J. Chen, W. Cranton, & M. Fihn (Eds.), *Handbook of Visual Display Technology* (pp. 187–196). Cham, Switzerland: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-14346-0_14
- Coxon, A. P. M. (1999). *Sorting Data: Collection and Analysis*. Thousand Oaks, CA: SAGE Publications, Inc.
- de Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, 31(3), 1–30.
- Dow, B. M., & Gouras, P. (1973). Color and spatial specificity of single cortex units in Rhesus monkey foveal striate cortex. *Journal of Neurophysiology*, 36(1), 79–100. Retrieved from <https://doi.org/10.1152/jn.1973.36.1.79>
- Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, 96(3), 433–458. Retrieved from <https://doi.org/10.1037/0033-295X.96.3.433>
- Geisler, W. S., Albrecht, D. G., Crane, A. M., & Stern, L. (2001). Motion direction signals in the primary visual cortex of cat and monkey. *Visual Neuroscience*, 18(4), 501–516.
- Geisler, W. S., & Perry, J. S. (2011). Statistics for optimal point prediction in natural images. *Journal of Vision*, 11(12), 1–17. <https://doi.org/10.1167/11.12.14.Introduction>
- Geusebroek, J.-M., Burghouts, G. J., & Smeulders, A. W. M. (2005). The Amsterdam Library of Object Images. *International Journal of Computer Vision*, 61(1), 103–112. Retrieved from <https://doi.org/10.1023/B:VISI.0000042993.50813.60%0A>
- Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, 26(4), 381–386. Retrieved from <https://doi.org/10.3758/BF03204653>
- Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, 58(2), 137–176. <https://doi.org/10.1016/j.cogpsych.2008.06.001>
- Hopper, W. J., Finklea, K. M., Winkelman, P., & Huber, D. E. (2014). Measuring sexual dimorphism with a race-gender face space. *Journal of Experimental Psychology: Human Perception and Performance*, 40(5), 1779–1788. Retrieved from <https://doi.org/10.1037/a0037743>
- Hout, M. C., Goldinger, S. D., & Brady, K. J. (2014). MM-MDS: A Multidimensional scaling database with similarity ratings for 240 object categories from the Massive Memory Picture Database. *PLoS ONE*, 9(11), e112644. <https://doi.org/10.1371/journal.pone.0112644>
- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1), 106–154. Retrieved from <https://doi.org/10.1113/jphysiol.1962.sp006837>
- Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1), 215–243. Retrieved from <https://doi.org/10.1113/jphysiol.1968.sp008455>
- Jaworska, N., & Chupetlovska-Anastasova, A. (2009). A review of multidimensional scaling (MDS) and its utility in various psychological domains. *Tutorials in Quantitative Methods for Psychology*, 5(1), 1–10. Retrieved from <https://doi.org/10.20982/tqmp.05.1.p001>
- Jiang, Y. V., Lee, H. J., Asaad, A., & Remington, R. (2015). Similarity effects in visual working memory. *Psychonomic Bulletin & Review*, 23(2), 476–482. <https://doi.org/10.3758/s13423-015-0905-5>
- Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, 47(26), 3286–3297. <https://doi.org/10.1016/j.visres.2007.09.013>
- Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, 139(3), 558–578. <https://doi.org/10.1037/a0019165>
- Konkle, T., & Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25), 10235–10242. <https://doi.org/10.1523/JNEUROSCI.0983-13.2013>
- Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, 74(6), 1114–1124. <https://doi.org/10.1016/j.neuron.2012.04.036>
- Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3, 1–13. <https://doi.org/10.3389/fpsyg.2012.00245>
- Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling: Volume 11 of Quantitative Applications in the Social Sciences*. SAGE Publications, Inc. Retrieved from <https://doi.org/10.4135/9781412985130>
- Larkey, L. B., & Markman, A. B. (2005). Processes of similarity judgment. *Cognitive Science*, 29, 1061–1076. Retrieved from https://doi.org/10.1207/s15516709cog0000_30
- Li, A. Y., Liang, J. C., Lee, A. C. H., & Barense, M. D. (2020). The validated circular shape space: Quantifying the visual similarity of shape. *Journal of Experimental Psychology: General*, 149(5), 949–966. Retrieved from <https://doi.org/10.1037/xge0000693>

- Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L. Y., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *ELife*, 7, 1–29. <https://doi.org/10.7554/elife.31873>
- McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4(4), 310–322. <https://doi.org/10.1038/nrn1076>
- Migo, E. M., Montaldi, D., & Mayes, A. R. (2013). A visual object stimulus database with standardized similarity information. *Behavior Research Methods*, 45(2), 344–354. <https://doi.org/10.3758/s13428-012-0255-4>
- Mugavin, M. E. (2008). Multidimensional scaling: A brief overview. *Nursing Research*, 57(1), 64–68. Retrieved from <https://doi.org/10.1097/01.NNR.0000280659.88760.7c>
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3), 145–175. Retrieved from <https://doi.org/10.1023/A:1011139631724>
- R Core Team. (2016). R: A language and environment for statistical computing. : R Foundation for Statistical Computing. Retrieved from <https://www.r-project.org/>
- Robertson, A. R. (1977). The CIE 1976 color-difference formulae. *Color Research & Application*, 2(1), 7–11. Retrieved from <https://doi.org/10.1002/j.1520-6378.1977.tb00104.x>
- Rodman, H. R., & Albright, T. D. (1989). Single-unit analysis of pattern-motion selective properties in the middle temporal visual area (MT). *Experimental Brain Research*, 75(1), 53–64.
- Ross, D. A., Sadil, P., Wilson, D. M., & Cowell, R. A. (2018). Hippocampal engagement during recall depends on memory content. *Cerebral Cortex*, 28(8), 2685–2698. <https://doi.org/10.1093/cercor/bhx147>
- Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, 33(2), 217–236. <https://doi.org/10.1068/p5117>
- Schurigin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*. <https://doi.org/10.1038/s41562-020-00938-0>
- Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2), 174–215. Retrieved from <https://doi.org/10.1037/0278-7393.6.2.174%0A>
- Solomon, S. G., & Lennie, P. (2007). The machinery of colour vision. *Nature Reviews Neuroscience*, 8(4), 276–286. <https://doi.org/10.1038/nrn2094>
- Sturrock, K., & Rocha, J. (2000). A multidimensional scaling stress evaluation table. *Field Methods*, 12(1), 49–60. <https://doi.org/10.1177/1525822X0001200104>
- Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42(1), 7–67. Retrieved from <https://doi.org/10.1007/BF02293745>
- Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, 77, 184–197. <https://doi.org/10.1016/j.chb.2017.08.038>
- Treisman, A. (1991). Search, similarity, and integration of features between and within dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 17(3), 652–676. Retrieved from <https://doi.org/10.1037/0096-1523.17.3.652>
- Tresch, M. C., Sinnamon, H. M., & Seamon, J. G. (1993). Double dissociation of spatial and object visual memory: Evidence from selective interference in intact human subjects. *Neuropsychologia*, 31(3), 211–219. [https://doi.org/10.1016/0028-3932\(93\)90085-E](https://doi.org/10.1016/0028-3932(93)90085-E)
- Xiao, J., Ehinger, K. A., Hays, J., Torralba, A., & Oliva, A. (2014). SUN Database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119(1), 3–22. <https://doi.org/10.1007/s11263-014-0748-y>
- Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN Database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3485–3492. Retrieved from <https://doi.org/10.1109/CVPR.2010.5539970>
- Yang, T., & Maunsell, J. H. R. (2004). The effect of perceptual learning on neuronal responses in monkey visual area V4. *Journal of Neuroscience*, 24(7), 1617–1626. <https://doi.org/10.1523/JNEUROSCI.4442-03.2004>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The data and materials for the experiment are available at <https://osf.io/smk25/>, and it was not preregistered.