A visual object stimulus database with standardized similarity information

Ellen M. Migo · Daniela Montaldi · Andrew R. Mayes

Published online: 6 October 2012
© Psychonomic Society, Inc. 2012

Abstract Although many visual stimulus databases exist, none has data on item similarity levels for multiple items of each kind of stimulus. We present such data for 50 sets of grayscale object photographs. Similarity measures between pictures in each set (e.g., 25 different buttons) were collected using a similarity-sorting method (Goldstone, Behavior Research Methods Instruments & Computers, 26(4):381–386, 1994). A validation experiment used data from 1 picture set and compared responses from standard pairwise measures. This showed close agreement. The similarity-sorting measures were then standardized across picture sets, using pairwise ratings. Finally, the standardized similarity distances were validated in a recognition memory experiment; false alarms increased when targets and foils were more similar. These data will facilitate memory and perception research that needs to make comparisons between stimuli with a range of known target–foil similarities.

Keywords Picture similarity · Recognition memory · Naturalistic pictures · Ratings · Norms · Target–foil similarity · Similarity sorting

Electronic supplementary material The online version of this article (doi:10.3758/s13428-012-0255-4) contains supplementary material, which is available to authorized users.

E. M. Migo
Department of Psychological Medicine, Institute of Psychiatry,
London, UK

D. Montaldi · A. R. Mayes
School of Psychological Sciences, University of Manchester,
Oxford Road,
Manchester M13 9PL, UK

E. M. Migo (✉)
King's College London,
Academic Unit of Neuropsychiatry,
3rd Floor Adamson Centre, South Wing Block D,
St Thomas's Hospital, Westminster Bridge Road,
London SE1 7EH, UK
e-mail: Ellen.Migo@kcl.ac.uk

Introduction

In cognitive psychology tests, stimulus selection is critical. Using pictures of objects, various standardized sets of stimuli have been published to include details such as visual complexity and familiarity. The line drawings with norms published by Snodgrass and Vanderwart (1980) have now been cited over 2,000 times, indicating their importance and continuing use. Various extensions to this database have been developed, including the addition of new pictures and updated information from different languages, cultures, or age groups (e.g., Berman, Friedman, Hamberger, & Snodgrass, 1989; Morrison, Chappell, & Ellis, 1997; Snodgrass & Yuditsky, 1996; Yoon et al., 2004). More recently, researchers have started to develop standardized databases of photographic stimuli that are more naturalistic (e.g., Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010; Viggiano, Vannucci, & Righi, 2004).

Our particular theoretical interest is in recognition memory research, where the similarity between targets and foils can differentially influence how well different kinds of memory, such as familiarity and recollection, support successful recognition memory (Migo, Montaldi, Norman, Quamme, & Mayes, 2009; Norman & O'Reilly, 2003). High target–foil similarity is important for at least one standardized neuropsychological test of memory: the Doors and People test (Baddeley, Emslie, & Nimmo-Smith, 1994). Quantifying the similarity between stimuli is increasingly important, so that different target–foil combinations can be properly matched or systematically varied.

Different types of stimuli have been developed and used for tasks requiring highly similar targets and foils, in perception and memory research. Here, we specifically mean experiments with identifiable but different target items, each with similar foils, as opposed to tests where all items to be remembered are similar or abstract (e.g., the kalidoscopes used in Voss & Paller, 2009). The kinds of stimuli include

complex scenes split into two (Dobbins, Kroll, & Liu, 1998; Tulving, 1981), color drawings of items (Jenison, Kirwan, Hopkins, Wixted, & Squire, 2010), line drawings of items or abstract patterns (Barens et al., 2005), pictures of scenes (Huebner & Gegenfurtner, 2012), and hand-drawn object silhouettes (Holdstock et al., 2002). These pictures were mainly chosen using experimenters' intuitive judgments; no direct measurements of the similarity between pictures were made. One study that used behavioral measures to demonstrate high target–foil similarity used drawn object silhouettes, as opposed to naturalistic pictures (Holdstock et al., 2002). The newly generated similar target–foil pairs were compared against each other in a pairwise discrimination task, which also included target–foil pairs from other experiments in the literature. The *similar* pairs were less accurately discriminated and had longer reaction times than did other stimuli, indicating that they were more similar than the stimuli standardly used in memory research. Another recent experiment used visually and conceptually similar foils, where visual similarity was measured by a computer algorithm based on color and spatial frequency (Huebner & Gegenfurtner, 2012). The pictures were in color and of scenes or objects embedded in scenes.

Although similarity data on specific stimuli provide useful information, having a greater variety of pictures and larger pictures sets would offer more choice to researchers wanting to select appropriate stimuli. For investigations of items and item memory, pictures of single items are preferable. Other databases of drawings of many different objects with many objects of each kind do not have this information either (e.g., Op De Beeck & Wagemans, 2001).

Despite the many databases with information about visual stimuli, we identified a need for a collection of photographs of different kinds of objects with properly measured levels of subjective perceptual similarity between pictures of each kind of object. Collecting similarity information can be very time consuming and tedious for participants if there are a lot of stimuli. It is usually done using pairwise similarity ratings or same/different judgments. To try to prevent participant fatigue, we used a similarity-sorting method developed by Goldstone (1994). Our overall aim was to develop sets of pictures for use in cognitive psychology research—specifically, to investigate theories of memory relating to the similarity between targets and foils. In order to do this, we needed reliable measures of picture similarity that were equivalent for all of the picture sets generated. Finally, we needed to demonstrate that these changes in picture similarity had a direct effect on memory performance, to show that the stimuli can be applied to cognitive research.

We collected 50 sets of photographs of different kinds of objects. The 50 sets comprised kinds of objects, where each set was relatively dissimilar to the others, with between 13 and 25 pictures per set. Participants sorted them on a

computer screen to efficiently collect similarity information between all pictures within a set in Experiment 1. We validated the results of this sorting procedure against the results for one picture set, using pairwise ratings, in Experiment 2. A separate rating experiment helped to standardize these dissimilarity values across sets in Experiment 3, and finally, the similarity information was validated in a recognition memory test in Experiment 4. These stimuli, although developed with recognition memory issues specifically in mind, should be relevant to a variety of psychological research areas where similarity is important, such as categorization, perception, and unaware forms of memory, such as priming.

Experiment 1: Obtaining rating information

Method

Participants

Twenty-five participants ($M[\text{age}] = 20.6$ years, 13 male) sorted 50 object picture sets so that similar pictures were placed closer together than dissimilar ones (see Goldstone, 1994, for method). Participants were paid for their time, and the study was approved by the Research Ethics Committee of the School of Psychological Sciences, University of Manchester.

Materials

The pictures of everyday items were taken with a digital camera and converted to grayscale. There were 50 sets of pictures, with between 13 and 25 pictures in each set ($M = 21.8$, median = 23). The pictures were all taken on a plain background, and this was removed manually for each picture, using the background eraser tool in PaintShopPro (1,095 pictures in total). One picture from each set is shown in Fig. 1, and an example of the full picture set for the Apple Set is shown in Fig. 2. All pictures are provided in the supplemental materials.

Procedure

Participants sorted pictures on a widescreen computer screen, using a mouse. On each trial, all the pictures from a picture set were presented, so that similarity ratings between all pictures were obtained at once. This sorting method has been shown to be as reliable as pairwise ratings and reaction times from same/different judgments (Goldstone, 1994). Each participant completed 50 trials, 1 for each picture set, in a fully randomized order, where the start position of each picture on the screen was also randomly assigned.

Fig. 1 An example picture from each of the 50 picture sets. Names used in the supplemental material lists are given, with full names/ descriptions underneath, where needed. The number of pictures in each set is also listed

Fig. 2 The full set of pictures in the Apple picture set

Participants were given instructions before starting, based on previous use of the method, using example screens before and after sorting taken from Goldstone (1994) and Busey and Tunnicliff (1999). They were asked to arrange the pictures so that the distance between them represented how similar they were to each other. Increasing distance indicated increasing dissimilarity. Participants could move pictures into any spatial arrangement of their choice and were under no time pressure to complete the sorting. They were encouraged to use the whole screen. Also, to reduce clumping of stimuli together, it was emphasized that the distance between pictures was a measure of how similar they were, so participants should use the distance as a continuous variable. The experiment was self-paced, and there was no time pressure. The experiment took approximately 100 min to complete. The stimuli were presented and responses collected using E-Prime 1.1, which measured the distance in pixels between all pictures following sorting. The computer was running Windows XP, with a screen resolution of $1,920 \times 1,200$ pixels.

Results and discussion

Results were averaged across all participants to give a matrix for each picture set of the dissimilarity distances (the term *dissimilarity* is used because a greater value indicates less similarity). The mean distance within a set ranged from 717 pixels (tomato) to 807 pixels (key), while the mean standard deviation for each full set of dissimilarity distances ranged from 314 pixels (hanger) to 412 pixels (keys). Supplementary materials provide the original output files for each participant and the mean dissimilarity distances for every pairing in every set and the standard deviations of these distances. In order to show the level of agreement between participants in a different way, the group multidimensional scaling (MDS) solution was obtained for each picture set, limited to two dimensions. This analysis was carried out using PROXSCAL as implemented in SPSS Version 20. This arrangement was compared against the final sorting positions for each picture set for each participant, using the

Procrustes function in MATLAB. This gives an overall measure of how similar each person's output is to the group MDS solution, on a scale from zero to one. The results are shown in Table 1. The most consistent results were seen with the holly picture set, as indicated by the lowest mean dissimilarity

Table 1 Mean similarity (standard deviation) between group multidimensional scaling solution and each participant's sorting plot, as measured by Procrustes analysis

Picture set	<i>N</i>	<i>d</i>	Picture set	<i>N</i>	<i>d</i>
Apple	24	.677 (.147)	Keys	15	.672 (.165)
Bag	25	.529 (.184)	Leaf	23	.590 (.103)
Biscuit	25	.683 (.138)	Lipstick	18	.617 (.192)
Bracelet	18	.409 (.273)	Man	25	.737 (.131)
Broc	25	.705 (.150)	Mushroom	25	.541 (.152)
Bun	23	.688 (.180)	Nut	25	.725 (.156)
Button	25	.685 (.162)	Onion	25	.784 (.166)
Cake	25	.638 (.172)	Paperclip	18	.441 (.139)
Candle	21	.482 (.173)	Pear	25	.423 (.240)
Car	24	.662 (.156)	Pen	23	.584 (.145)
Chilli	25	.427 (.166)	Plug	22	.673 (.148)
Chocolate	25	.595 (.215)	Present	25	.635 (.136)
Coins	15	.420 (.176)	Scissors	25	.576 (.203)
Daff	23	.539 (.214)	Screw	18	.536 (.209)
Feather	18	.588 (.189)	Shell	25	.556 (.158)
Flower	25	.557 (.189)	Stapler	16	.518 (.164)
Fork	19	.705 (.144)	Star	25	.617 (.223)
Fudge	20	.586 (.191)	Strawberry	21	.636 (.185)
Ginger	20	.664 (.134)	String	24	.776 (.136)
Glass	20	.529 (.209)	Sweet	25	.662 (.176)
Hand	22	.535 (.179)	Swirl	24	.752 (.140)
Hanger	15	.414 (.176)	Tomato	21	.688 (.164)
Holly	20	.398 (.178)	Toothbrush	13	.506 (.206)
Jigsaw	18	.601 (.278)	Umbrella	23	.491 (.214)
Key	24	.737 (.131)	Varnish	18	.644 (.148)

Note. *N* is the number of pictures per picture set.

measure (d) for the group. The highest dissimilarity is seen for the onion picture set, indicating that participants' plots varied the most, as compared with the group overall MDS solution.

Figure 3 gives some example sorting maps from participants and the group MDS solution for the Apple picture set. All maps are presented in common space coordinates. These maps indicate the different ways in which participants interpreted *similarity*. Procrustes analysis for all participants for the Apple picture set showed that Fig. 3b was the most different to the group map, while 3c was the closest. The participant who produced Fig. 3b appears to have sorted on the basis of which direction the apple pointed, whereas 3c seems to include dimensions on color and texture. Figure 3d is an example of a sorting map that has a dissimilarity to the group MDS map halfway between those shown in panels b and c. Although there is some evidence of clustering together of highly similar pictures, there were no participants who clumped items together and did not use the distance between pictures as a measure of similarity. The group MDS solution was best fit by four dimensions to the data, as indicated by the scree plot of how stress decreases with increasing dimensions. This showed a slight “elbow” in the graph at four dimensions. This demonstrates that the simple two-dimensional sorting method used here can, with a group, extract information across more than two dimensions.

Previous work using this similarity-sorting technique has demonstrated very high agreement with the more commonly used measures where pairs of pictures are used, such as similarity ratings or same/different judgments. Goldstone (1994) reported correlations of .85, .87, and .93 between response

times for *different* judgment and ratings, *different* response times and spatial sorting, and pairwise ratings and similarity sorting. However, the original use of the task used multiple presentations of pictures within a set of 64. Here, we have presented each set of pictures only once in order to obtain ratings from all 50 picture sets from each participant. This may be less reliable. In order to check that our measures from Experiment 1 were reliable, Experiment 2 collected pairwise rating data from a single picture set (apples) to compare with the dissimilarity matrix collected from the sorting method.

Experiment 2: Validating rating information

Method

Participants

Twenty-five new participants ($M[\text{age}] = 30.2$ years, 12 male) rated the similarity of every combination of pairs of pictures within the apple picture set. The experiment was approved by the Psychiatry, Nursing and Midwifery Research Ethics Subcommittee, King's College London.

Materials

The first picture set alphabetically was chosen to be used in this experiment; the apple picture set. All 24 pictures in this set are shown in Fig. 2.

Fig. 3 Example sorting maps. **a** Group MDS solution (two-dimensional solution shown). **b** Participant sorting map furthest from Group MDS solution. **c** Participant sorting map closest to Group MDS solution. **d**: Participant sorting map with a similarity to the Group MDS map halfway between those shown in panels b and c

Procedure

Participants rated the similarity of the apple pictures in pairs in an E-Prime 1.1 experiment. The experiment was carried out on a Dell laptop computer running Windows XP with a screen resolution of $1,280 \times 800$ pixels. Participants were shown the full picture set of pears before starting the experiment. This showed them an example of the full range of possible similarity within a picture set, to encourage them to use the entire scale, without preexposing them to the exact stimuli used in the experiment. Participants were asked to respond on a scale labeled from 0 to 20, where a higher number indicated higher similarity, using the mouse. Numbers were written on the scale at 0, 10, and 20, and a constant reminder was on the screen that 0 represented *not similar at all* and 20 represented *very similar*. The experiment recorded data on a percentage scale from 0 to 100 (marked as 20 on the screen). This variation meant that any value of similarity could be given, not just the 21 points on the scale from 0 to 20. This was more analogous to the similarity-sorting output measures than a limited point scale would be. Five example trials using pictures from the pear set were used to ensure that participants could use the scale properly. Comparing all 24 apple pictures against each other required 276 trials per participant, taking approximately 18 min. Participants had the opportunity to pause every 30 trials. The order of pairs was fully randomized, with the left/right position of the pictures counterbalanced across participants.

Results and discussion

Ratings from all participants were averaged together to give a mean similarity distance for each pair of pictures within the set. These scores were then converted to dissimilarity distances by reversing the scale to give a comparable output measure to the values from Experiment 1. The overall average dissimilarity was 53.47, ranging from 13.08 to 84.08, with a standard deviation of 16.88. This indicates

that participants did use the full range of the scale available. To investigate whether the two methods produce comparable results, following previous validation of the sorting method (Goldstone, 1994), we correlated the outputs for the apples from Experiments 1 and 2. The dissimilarity distances were significantly correlated, $r = .83$, $p < .001$, showing very high agreement between the two methods. This linear relationship is shown in Fig. 4. This significant correlation, accounting for 69 % of the variance of the data, validates that the values in Experiment 1 represent real similarity distances, as measured by alternative methods.

The time taken to complete the experiment (approximately 18 min), which included only 1 of the 50 available sets, again highlights the efficiency of the picture-sorting method. Participants also found arranging pictures on a screen more interesting, whereas many participants who completed the pairwise rating experiment here commented on finding the task very tedious. Using this pairwise method, at 18 min for each of our 50 picture sets, it would have taken each participant 15 h, which is not practical for many situations. The sorting method correlates very highly with standard pairwise data and is a much quicker way of obtaining the information.

In order to use stimuli from multiple picture sets in an experiment—for example, when pictures within the sets are used as target and foils in memory tests—we need to know whether any given dissimilarity distance is equivalent *across* the picture sets. If, for example, all the pictures of sweets are perceived as more similar to each other than all the pictures of toothbrushes (which seems plausible on inspection), then picking pictures on the basis of ratings from this experiment would not result in equivalent dissimilarities across both sets. Without further evidence comparing pictures across picture sets, it is not possible to know whether a large distance in one picture set indicates a level of stimuli dissimilarity equivalent to that for the same distance in another picture set. If the aim is to have a variety of different target items to study, each with foils of matching

Fig. 4 Scatterplot of dissimilarity distances from the sorting method (Experiment 1) against the pairwise ratings from Experiment 2

similarity, standardization is required. For use in certain cognitive tasks, a common scale across all picture sets is essential. In Experiment 3, we tried to adjust the output from Experiment 1 to provide this information. Here, we took an example pair from each stimulus set with matched similarity from the original rating experiment. From these, we obtained pairwise ratings and used them to recalculate the dissimilarity values to provide an equivalent scale. Where this information is not important or where stimuli within a single picture set are used, it may be more appropriate to use the outputs from Experiment 1 that have not been adjusted.

Experiment 3: Standardizing distances

Method

Participants

A novel group of 25 participants ($M[\text{age}] = 20.2$ years, 9 male) took part in this experiment, which was approved by the School of Psychological Sciences Research Ethics Committee, University of Manchester. Participants were paid for their time.

Materials

Using the output from Experiment 1, a sample picture pair was chosen from each picture set. Our aim was to match the pixel distance between the pairs of pictures, which was the direct measure of picture similarity, as closely as possible, and we chose an arbitrary level of 800. The mean value for the pairs was 799.9, with a standard deviation of 2.08 (range 793–806). In this way, the output dissimilarity distances from the pair from each picture set were matched as closely as possible before being standardized.

Procedure

Participants rated 50 pairs of pictures on a scale from one to nine, where a lower number indicated that the pictures were more similar. The stimuli were presented in E-Prime 1.1 in a fully randomized order, and the left–right position on the screen was counterbalanced between participants. Participants responded by pressing the appropriate number from one to nine on the keyboard. This experiment was presented on a Dell desktop computer, running Windows XP with an experimentally defined screen resolution of $1,024 \times 768$ pixels.

Results and discussion

These data were used to adjust the average distances generated in Experiment 1, so that dissimilarity values were

measured on an equivalent scale across data sets. For example, the mean similarity between the selected apple pictures (apple2 vs. apple19) was 3.90. Every value in the dissimilarity matrix for the apple pictures was multiplied by 3.90. This way, picture sets that were overall more dissimilar were given larger standardized dissimilarity distances. Tables were constructed giving standardized dissimilarity distances for all 50 picture sets, where values were rounded to the nearest ten. A difference of less than 10 on these scales is highly unlikely to be meaningful, and therefore, rounding was carried out to avoid overinterpreting very small differences between dissimilarities. These values are arbitrary and are interpretable only in relation to each other. All tables and stimuli are presented in the [Supplementary materials](#).

We next needed to validate that these standardized dissimilarity distances were meaningful and equivalent across the picture sets, especially since transforming the values may have added in extra error. In order to do this, we carried out a simple memory test in Experiment 4. In standard memory tests, it is clearly established that the perceptual similarity between a target and a foil affects performance, such that as target–foil similarity increases, so performance decreases (e.g., Tulving, 1981; Woodworth, 1938). If the ratings produced as an output from Experiment 3 really do provide scaled information on picture similarity, selecting targets and foils on this basis should produce a performance pattern where participants get worse as similarity increases.

Experiment 4: Validating dissimilarity scores in a recognition memory test

Method

Participants

Thirty-three new participants took part in a recognition memory test to validate the ratings by showing that memory accuracy decreased as target–foil similarity increased ($M[\text{age}] = 20.7$ years, 11 male). This experiment was approved by the School of Psychological Sciences Research Ethics Committee, University of Manchester. Participants were paid for their time.

Materials

Fifty pairs of pictures were chosen for the recognition memory study, one pair from each picture set; one member of each pair was used as a studied target, and the other member was used at test as a distractor of known similarity to the target. Therefore, 50 pictures were shown at study, and 100 pictures shown at test (the studied target and its similar foil). The range of dissimilarity in the adjusted matrices varied

from 1,000 to 6,000, at intervals of 1,000, along the arbitrary scale produced in Experiment 3. These six levels represented the majority of the range of dissimilarity found with the pictures (see Fig. 5). This resulted in either eight or nine pairs at each similarity level. Three versions of the experiment were created, using different picture pairs in each, to ensure that any performance effects were not due to any item effects.

Procedure

Participants were instructed that they were going to study a series of pictures and, at test, they would have to distinguish between very similar versions of each picture. Each picture was studied for 3 s, in a random order. There was a 1-min filled delay before the test phase, where mental arithmetic problems were used. This was a self-paced yes/no test where participants were asked to decide whether they had seen an item before. Participants were encouraged to answer quickly but to prioritize accuracy over speed. The experiment was presented in E-Prime 1.1, on a Dell desktop machine running Windows XP with an experimentally defined screen resolution of $1,024 \times 768$ pixels.

Results and discussion

Performance was indexed by a measure of hits minus false alarms (P_r) and is shown in Fig. 6, alongside the responses

Fig. 5 Example pairings at different target–foil similarity levels. The x -axis indicates dissimilarity, where higher values represent less similar pairings. Each value on the x -axis represents 1,000 arbitrary units on the scales produced from Experiment 3. T = target

for hits and false alarms separately. P_r is a preferable index to d' when using highly similar targets and foils in a yes/no recognition memory test (Migo et al., 2009), but below we also report statistical results using d' . To calculate hit and false alarm rates for the d' analysis, all data were systematically corrected for ceiling and floor effects as recommended by Snodgrass and Corwin (1988).

A repeated measures ANOVA showed that, as would be expected from Fig. 6a, there was an effect of target–foil similarity on task performance, where performance was worse with increasing target–foil similarity. This investigated the main effect of similarity (levels one to six; six levels) and experiment version (versions one to three; three levels). Combining the performance values across the three experiment versions, there was a significant main effect of similarity, $F(5, 150) = 23.95$, $MSE = 1.13$, $p < .0001$, $f = 0.89$, but the main effect of experiment version was not significant, $F(2, 30) = 0.074$, $MSE = 0.011$, $p = .929$, $f = 0.07$. Contrasts revealed a significant linear effect of similarity, $F(1, 30) = 109.71$, $MSE = 5.23$, $p < .001$, $f = 1.91$. The interaction between similarity and experiment version was nonsignificant, $F(10, 150) = 0.520$, $MSE = 0.025$, $p = .874$, $f = 0.18$. For data using d' , the same pattern emerged, with a significant effect of similarity, $F(5, 150) = 20.10$, $MSE = 10.04$, $p < .001$, $f = 0.82$, but not for experiment version, $F(2, 30) = 0.225$, $MSE = 0.34$, $p = .800$, $f = 0.12$. The linear contrast was significant $F(1, 30) = 85.471$, $MSE = 45.85$, $p < .001$, $f = 1.69$, but the interaction with experiment was not, $F(2, 30) = 0.66$, $MSE = 0.33$, $p = .760$, $f = 0.21$.

Inspection suggested that hit rates remained constant whereas false alarm rates increased with similarity (Fig. 6b). Separate repeated measures ANOVAs for hits and false alarm rates found a significant linear contrast for the false alarm rate, $F(1, 30) = 277.229$, $MSE = 6.173$, $p < .0001$, $f = 3.03$, but not the hit rate, $F(1, 30) = 2.335$, $MSE = 0.039$, $p = .137$, $f = 0.29$.

These measures of dissimilarity appear to have utility in cognitive experiments; as targets and foils decreased in similarity, so performance increased. There was no effect of experiment version, indicating that this is not an artificial effect driven by stimulus-specific effects. The linear pattern seen in false alarm rate indicates that the ratings of similarity in our database are sufficiently sensitive to show graded changes in performance.

General discussion

This article presents a series of experiments for collecting data about similarity information between pictures. The pictures were all grayscale photographs of everyday objects. Participants completed a similarity-sorting procedure to quickly and efficiently collect data from a large number of

Fig. 6 Performance on recognition memory task (stage 3) as measured by **a** P_r (hits minus false alarms) or **b** hit and false alarm rates. Error bars indicate standard error. The dissimilarity scale is in units of 1,000, as taken from the output tables from Experiment 3

stimuli. This would have been a very difficult and impracticably long task using standard pairwise methods. These dissimilarity distances were validated against the more commonly used pairwise rating method for an example picture set, where high agreement between methods was seen. A representative pair from each picture set was then used in a rating experiment with a separate group of participants in order to standardize the dissimilarity ratings across picture sets. Finally, a memory test validated these standardized outputs by showing that recognition memory was progressively worse as targets and foils were more similar. This effect was mainly mediated by increasing false alarms to more similar foils, which linearly increased with increasing similarity. Since there was no main effect of experiment version or any interactions including it, this validates, first, that the similarity ratings have meaning and, second, that they are equivalent and not specific to any particular pairings, at least in the context of those types of memory task. The values of similarity are sensitive enough to show scaled effects in memory

performance, and therefore, the picture similarity levels are at an appropriate level to affect behavior.

Interpretation of the distances obtained in this series of experiments depends on a number of assumptions—most important, that similarity/dissimilarity is linear and that similarity is symmetric (see Goldstone, 1994, for a discussion of potential limitations of the sorting procedure). Using a sorting method in two dimensions, on a computer screen that does not give a square sorting space, can also limit the ways in which participants can define similarity. As such, it could be argued that our data constitute a simplistic attempt to model similarity. However, the strong linear correlation obtained between sorting dissimilarity distances and the widely accepted method of pairwise ratings of similarity indicates that any potential problems with this method are unlikely to be too problematic, or at least no more problematic than with any other established method. Despite all these potential limitations, the results from the memory test in Experiment 4 show that, on the whole, the ratings that we

have work well and can be considered relative to each other. This use of the stimuli in a behavioral task demonstrates that meaningful information on relative similarity between pictures has been collected, at least for the purposes of cognitive tasks. In practical terms, obtaining this amount of information from every participant using traditional pairwise methods would have been far more laborious, and participant fatigue concerns would have been a very strong concern.

We did not ask participants to judge similarity on any particular dimension and, in fact, if asked, told them to decide how to judge similarity themselves (for Experiments 1–3). It is likely that judgments depend on multiple features of the stimuli, since our pictures vary in similarity in many dimensions, as is the case with naturalistic pictures. Different participants are likely to use overlapping but at least partially distinct criteria for rating similarity, and they may weight these criteria differently. This is shown to a large extent with the different sorting maps seen in Experiment 1 from different participants (and every participant's original dissimilarity matrix for each picture set is provided in the supplementary materials). Given this variation, Experiment 4 is an important test of whether these potentially idiosyncratic similarity ratings can be simplistically averaged together. The results showed that the ratings from one group of participants did result in predicted behavior differences in a different group. Although there will be individual differences in exactly how people rate similarity for these pictures, there is enough generality for the averaged ratings to be a very useful tool in experimental design.

For memory tasks with multiple foils, it is preferable for pictures to vary in different ways. This prevents one single feature being used to distinguish between a target and all the foils. Our stimulus categories also differ in how similar they are. In many experiments, it may be inappropriate to include pictures of a set of keys and pictures of single keys, or to include pictures of two types of leaves (i.e., holly and sycamore, labeled as *leaf*). Having 50 separate sets should provide sufficient options for many types of experiment.

We believe that these are psychologically meaningful measures of stimulus similarity, which may be more useful than measures computed mathematically from morphed pictures. This is an important step toward fully characterizing the properties of naturalistic picture stimuli. Having the opportunity to use multiple naturalistic stimuli where the subjective similarity between pictures can be systemically varied across trials and/or conditions may help us to understand the relationships between subjective similarity and other cognitive functions. This may be most relevant to experimental memory research, for which these stimuli were designed, but should also be useful for perception, categorization and other areas of cognitive research.

Acknowledgments We would like to thank Yu Li for his help with E-Prime, without which we could not have completed this project, and Rob Goldstone and Thomas Busey for providing their scripts and example stimuli. We also thank Richard Haworth for help with data collection and Lara Harris for her feedback on an earlier version of the manuscript. This work was supported by the Small Grants Scheme from the Experimental Psychology Society.

References

- Baddeley, A. D., Emslie, H., & Nimmo-Smith, I. (1994). *Doors and people: A test of visual and verbal recall and recognition*. Bury St. Edmunds, England: Thames Valley Test Co.
- Barens, M. D., Bussey, T. J., Lee, A. C. H., Rogers, T. T., Davies, R. R., Saksida, L. M., et al. (2005). Functional Specialization in the Human Medial Temporal Lobe. *The Journal of Neuroscience*, 25(44), 10239–10246.
- Berman, S., Friedman, D., Hamberger, M., & Snodgrass, J. (1989). Developmental picture norms: Relationships between name agreement, familiarity, and visual complexity for child and adult ratings of two sets of line drawings. *Behavior Research Methods*, 21(3), 371–382.
- 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.
- Busey, T. A., & Tunnichiff, J. L. (1999). Accounts of blending, distinctiveness, and typicality in the false recognition of faces. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(5), 1210–1235.
- Dobbins, I. G., Kroll, N. E. A., & Liu, Q. (1998). Confidence-accuracy inversions in scene recognition: A remember-know analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(5), 1306–1315.
- Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, 26(4), 381–386.
- Holdstock, J. S., Mayes, A. R., Roberts, N., Cezayirli, E., Isaac, C. L., O'Reilly, R. C., et al. (2002). Under what conditions is recognition spared relative to recall after selective hippocampal damage in humans? *Hippocampus*, 12(3), 341–351.
- Huebner, G. M., & Gegenfurtner, K. R. (2012). Conceptual and visual features contribute to visual memory for natural images. *PLoS One*, 7(6).
- Jenison, A., Kirwan, C. B., Hopkins, R. O., Wixted, J. T., & Squire, L. R. (2010). Recognition memory and the hippocampus: A test of the hippocampal contribution to recollection and familiarity. *Learning & Memory*, 17(1), 63–70.
- Migo, E. M., Montaldi, D., Norman, K. A., Quamme, J. R., & Mayes, A. R. (2009). The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Quarterly Journal of Experimental Psychology*, 62(6), 1198–1215.
- Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of Acquisition Norms for a Large Set of Object Names and Their Relation to Adult Estimates and Other Variables. *Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, 50(3), 528–559.
- Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, 110(4), 611–646.
- Op De Beeck, H., & Wagemans, J. (2001). Visual object categorisation at distinct levels of abstraction: A new stimulus set. *Perception*, 30(11), 1337–1361.

- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, 117(1), 34–50.
- 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.
- Snodgrass, J. G., & Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments, & Computers*, 28(4), 516–536.
- Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior*, 20(5), 479–496.
- Viggiano, M. P., Vannucci, M., & Righi, S. (2004). A new standardized set of ecological pictures for experimental and clinical research on visual object processing. *Cortex*, 40(3), 491–509.
- Voss, J. L., & Paller, K. A. (2009). An electrophysiological signature of unconscious recognition memory. *Nature Neuroscience*, 12(3), 349–355.
- Woodworth, R. S. (1938). *Experimental psychology*. New York: Holt.
- Yoon, C., Feinberg, F., Luo, T., Hedden, T., Gutchess, A., Chen, H.-Y., et al. (2004). A cross-culturally standardized set of pictures for younger and older adults: American and Chinese norms for name agreement, concept agreement, and familiarity. *Behavior Research Methods*, 36(4), 639–649.