COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 36 (2012) 1019–1050

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2012.01250.x

What Difference Reveals About Similarity

Eyal Sagi, Dedre Gentner, Andrew Lovett

Department of Psychology, Northwestern University

Received 31 December 2010; received in revised form 11 September 2011; accepted 12 September 2011

Abstract

Detecting that two images are different is faster for highly dissimilar images than for highly similar images. Paradoxically, we showed that the reverse occurs when people are asked to describe *how* two images differ—that is, to state a difference between two images. Following structure-mapping theory, we propose that this disassociation arises from the multistage nature of the comparison process. Detecting that two images are different can be done in the initial (local-matching) stage, but only for pairs with low overlap; thus, ''different'' responses are faster for low-similarity than for high-similarity pairs. In contrast, identifying a specific difference generally requires a full structural alignment of the two images, and this alignment process is faster for high-similarity pairs. We described four experiments that demonstrate this dissociation and show that the results can be simulated using the Structure-Mapping Engine. These results pose a significant challenge for nonstructural accounts of similarity comparison and suggest that structural alignment processes play a significant role in visual comparison.

Keywords: Structural alignment; Perceptual comparison; Same-different judgments; Structure-mapping; Alignable differences

Similarity plays an important role in cognitive science, both as an empirical phenomenon in its own right (e.g., Hahn, Chater, & Richardson, 2003; Markman & Gentner, 1993; Tversky, 1977) and as a component of other cognitive processes. Similarity processes have been implicated in categorization (e.g., Goldstone, 1994; Hampton, 1997; Medin & Schaffer, 1978; Nosofsky, 1984), induction (e.g., Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Sloman, 1993), memory retrieval (e.g., Gentner, Rattermann, & Forbus, 1993; Gillund & Shiffrin, 1984; Ross, Perkins, & Tenpenny, 1990), problem solving (Bassok, 1990; Gick & Holyoak, 1980, 1983; Novick, 1988; Ross, 1989), and decision-making (e.g., Medin, Goldstone, & Markman, 1995; Tversky & Kahneman, 1981). Accordingly, much attention has been devoted to developing psychological models of the processes underlying similarity judgments.

Correspondence should be sent to Eyal Sagi, Department of Psychology, Northwestern University, 2029 Sheridan Road, Evanston, IL 60208. E-mail: eyal@u.northwestern.edu

Yet, despite the centrality of similarity processing, there is no general agreement as to how to model it. One reason for this is the lack of agreement on how concepts are represented in the mind. For example, Tversky's (1977) classic contrast model assumes representations composed of independent features. Another prominent class of models assumes high-dimensional spatial representations (e.g., Shepard, 1974). A third class of models assumes structured representations and accords a central role to aligning these representations in similarity processing (Gentner, 1983; Gentner & Markman, 1997; Holyoak & Thagard, 1989; Hummel & Holyoak, 1997; Larkey & Love, 2003). In this paper, we examine the psychology of difference detection as a way of distinguishing among these competing models of similarity. We suggest that models that include structural alignment can best account for the phenomena.

We first draw predictions from structure-mapping theory (SMT), in which difference detection is an integral part of similarity processing. Then we compare its predictions with those made by other classes of models. According to SMT (Falkenhainer, Forbus, & Gentner, 1989; Gentner, 1983, 2003; Gentner & Markman, 1993, 1994, 1997; Markman & Gentner, 1993), comparing two things involves a process of structural alignment. The alignment of two representations goes beyond the identification of shared features; it also requires finding correspondences between the *relations* that connect the features. Because structure-mapping postulates that similarity involves an alignment of representational structure, it naturally predicts a psychological distinction between alignable differences and nonalignable differences. *Alignable differences* are differences that occupy corresponding positions in their respective relational structures; they emerge when the two representations have been aligned. For example, in Fig. 1, the black circle in A versus white center circle in B constitute an alignable difference. *Nonalignable differences* are differences that do not occupy corresponding roles (or between items that cannot be aligned). In Fig. 1, if we compare A and C, the lion in C is a nonalignable difference, as is the black center in A.

An important prediction of SMT—and one central to the logic of this paper—is that alignable differences are in general more salient than nonalignable differences.[1] This follows from the more general claim that comparing two things makes their common structure more salient. Indeed, phenomenologically, alignable differences often seem to pop out. For example, in Fig. 1, the alignable difference between the top two figures (A and B) stands out immediately, whereas the nonalignable differences between A and C do not. This is advantageous in that alignable differences (by definition) are more relevant to the common causal or perceptual structure that is the basis for the comparison. But it leads to the rather paradoxical prediction that it should be easier to notice differences for high-similarity than for low-similarity pairs. There are two reasons for this within SMT. First, high-similarity pairs are easier to align than low-similarity pairs (as amplified below); and once two representations are aligned, the alignable differences stand out. Second, high-similarity pairs have larger common systems than low-similarity pairs and thus more slots for alignable differences. Structure-mapping theory thus predicts that difference identification should be faster for high-similarity pairs than for low-similarity pairs. We test this claim in this paper.

While there has up to now been no evidence on the relative speed of detecting differences between alignable versus nonalignable pairs, there is evidence from both conceptual

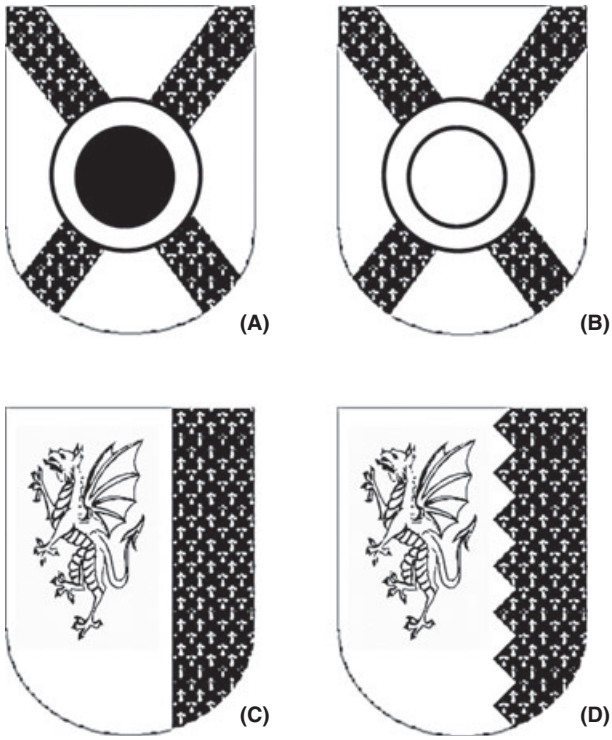


Fig. 1. Sample stimuli from Experiment 1. Images in the same row represent high-similarity pairs; images in the same column represent low-similarity pairs.

and perceptual comparisons for the two related claims—(1) that alignable differences are more salient than nonalignable differences; and (2) that differences are easier to detect for high-similarity pairs than for low-similarity pairs. Using a speeded-difference task, Gentner and Markman (1994) gave participants a page full of word pairs and asked them to find a difference between as many pairs as possible in a brief time period. Participants identified differences for many more high-similarity pairs than low-similarity pairs, and this surplus was chiefly made up of alignable differences. Applying this framework to perceptual comparison, Markman and Gentner (1996) gave participants image pairs and asked them to list either differences or commonalities. Again, participants listed more differences for highly similar images than for less similar ones, and again, this surplus was made up of alignable differences. Finally, Gentner and Gunn (2001) asked people to compare word pairs and write a commonality, and then gave them a speeded-difference task. As in the prior two studies, participants generated more differences (mostly alignable differences) for high-similarity than for low-similarity pairs. In addition, they generated more (alignable) differences for the previously compared pairs than for new pairs, showing the specific connection between alignment and difference-noticing. According to structure-mapping, the above findings stem from the related facts that (a) alignable differences are faster and easier to note than nonalignable differences (in general), and (b) high-similarity

pairs are easier to align than low-similarity pairs, as amplified below. Together these lead to the prediction that difference-identification will be faster for high-similarity than for low-similarity pairs.

On the face of it, this prediction seems at odds with a venerable body of research on same-different judgments. A well-established result is that the more similar two images are, the more difficult it is to identify that they are different (e.g., Farell, 1985; Goldstone & Medin, 1994; Luce, 1986; Posner & Mitchell, 1967; Tversky, 1969). That is, the more similar two things are, the more time people require to say ''different'' (and the more likely they are to erroneously identify the pair as ''same''). This result runs in the opposite direction from the prediction that people will be faster to identify differences in similar images than in dissimilar ones. For example, in Fig. 1, people should be faster to say ''different'' for pair AC than for pair AB; yet they should be faster to identify a specific difference for pair AB than for pair AC.

This disassociation poses problems for traditional accounts of similarity, as discussed below. But structure-mapping can resolve this apparent contradiction. As noted above, SMT readily predicts the greater ease of difference-identification for high-similarity than for low-similarity pairs. To draw predictions for the same-different task from structure-mapping, it is useful to review the alignment process as modeled by the Structure-Mapping Engine (SME) (Falkenhainer et al., 1989; Forbus, Gentner, & Law, 1995). Fig. 2 provides an overview of the three stages of this process (see also Gentner & Markman, 1997):

1. All possible local identity matches between elements in the two representations are made in parallel,[2] regardless of whether they are mutually consistent.
2. The local matches are coalesced into a set of structurally consistent connected structures (called *kernels*).

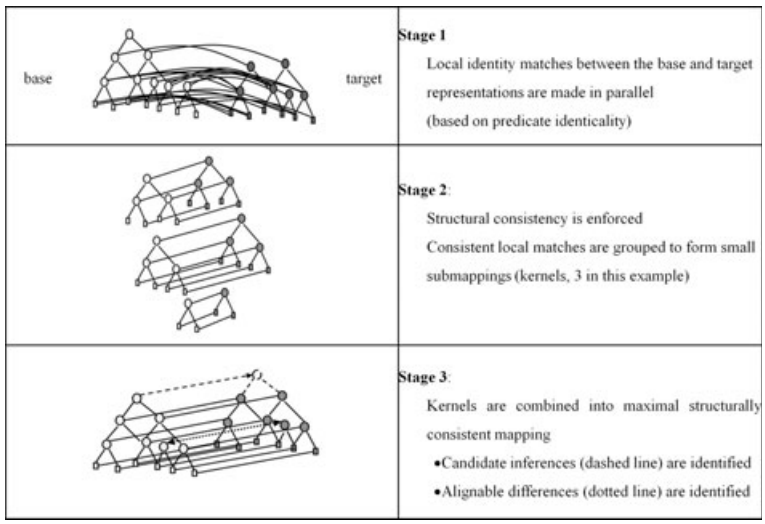


Fig. 2. Diagram showing the three stages of structural alignment in SME.

3. The kernels are merged, beginning with the largest and deepest kernel[3] and adding others that are structurally consistent with it. This results in one or a few large, structurally consistent *global mappings* between the representations. The global mapping reveals both commonalities and alignable differences between the two representations. (At this stage, candidate inferences may also be projected from one representation to the other, but this will not concern us here.)

The reason that high-similarity pairs are faster to align than low-similarity pairs (Gentner & Kurtz, 2006) is that in high-similarity pairs, the initial matches are largely mutually consistent: the object-property matches support the relational matches, and the relational matches are consistent with each other. Because most of the matches are compatible, this typically results in a dominant large, well-structured kernel, which is much larger (in terms of its structural evaluation) than the next-largest kernel. This means that the greedy merge process only needs to run once.[4] For low-similarity pairs, there are typically many small kernels, and the final merge step may require comparing two or more different orders of merging. Thus, the merge stage takes longer for low-similarity than for high-similarity pairs (assuming equal-sized representations). To use an analogy, a literal similarity match is like a wide highway with no branchpoints, whereas a low-similarity match is a set of faint criss-crossing paths.[5]

Given this alignment process, the prediction for the difference-identification task is straightforward. Because alignable differences emerge only from the global mapping, the process must proceed to the end. This means that whatever speeds up alignment will also speed difference-identification. Thus, difference-identification should be faster for high-similarity than for low-similarity pairs. To make predictions for the same-different task, consider first the high-similarity case. For high-similarity pairs, the same alignment process just described takes place; once the global mapping is made, a difference will be apparent and the ''different'' response can be made. But in the low-similarity case, a full alignment is not needed. For highly dissimilar pairs, there will be very few initial local matches (relative to the size of the representations). Since the possibility that two images are identical can be ruled out without proceeding further (Markman & Gentner, 2005), a quick ''different'' response can be made in the first stage. So in the same-different task, low-similarity pairs will yield faster ''different'' responses than high-similarity pairs, because only the latter require a full alignment process.

Structure-mapping theory therefore predicts that the two tasks will show opposite patterns with respect to similarity. For the same-different task, which permits a shortcut for very dissimilar images, participants should be faster to respond ''different'' for dissimilar than for similar pairs. Thus, in Fig. 1, they should be faster to say ''different'' for pair A and C than for the highly similar pair A and B. In contrast, for the difference-identification task, which always requires aligning the images, participants should be faster to respond to similar than to dissimilar pairs. Because pair A and B share a common organizing structure (as well as many features), the alignment process should be easy and fast, causing the alignable difference between them (the color of the central circle) to leap out. It should take longer to identify a difference between A and C (the two leftmost images), which have only minimal alignable structure.

If this predicted reversal is obtained—that is, if participants are faster to recognize that two images are different for low-similarity pairs but faster to identify a difference for high-similarity pairs—it will pose significant challenges for two prominent classes of models of perceptual similarity. Both feature-intersection models and mental distance models account very naturally for the finding that people are faster and more accurate to say ''different'' for very different pairs than for rather similar pairs. However, they also most naturally predict the same pattern for difference-identification. The greater the number of differences between two objects, the easier it should be *both* to detect that they are different and to identify a specific difference between them.

In feature-intersection models (e.g., Tversky, 1977; see Navarro & Lee, 2004; for a comparison of several different feature-intersection models) objects are represented by sets of independent features. Similarity between objects is increased by shared features and decreased by distinctive features. The reverse applies when computing a difference judgment; two objects are more different the greater their number of distinctive features. Thus, the greater the number of distinctive features, the easier it should be to detect that two objects are different, and the easier it should be to identify a distinctive feature. For instance, in Fig. 1, it should be easier both to distinguish A from C (because they are different in several features) than A from B (because they vary on only a single feature); and the large number of distinctive features should also make it easier to identify differences between A and C than between A and B.

In mental distance models (e.g., Nosofsky, 1984; Shepard, 1974; Shoben, 1983) similarity is modeled as the inverse of the distance between points within a multi-dimensional mental space. Relative positions within this space can then be used to measure how different the two objects are from one another. The farther apart two points are within the space, the easier it should be both to detect *that* they are different and to find specific differences in dimensional values between them. Thus, people should be faster for the dissimilar pair A and C than for the similar pair A and B for both tasks.

1. Plan of experiments

The prior research has used very different materials and response measures for the two tasks. In the present studies, we equated the tasks as far as possible. We used the same materials—pairs of images—for both tasks, and the same dependent measure, response times. Experiments 1 and 2 test whether the predicted disassociation between same-different responding and difference-identification occurs in human perceptual comparison. In Experiment 1, we gave participants pairs of structured images and asked them either to perform a same-different judgment or to identify a particular difference. In Experiment 2, we generalized the findings to more complex and naturalistic images. Experiment 3 tests an alternative account of the findings, and Experiment 4 tests a further prediction from structure-mapping theory. With these results in hand, we then present a computer simulation of both tasks, using the SME in combination with a sketching system that permits automatic encoding of perceptual materials.

2. Experiment 1

In Experiment 1, we designed a highly controlled set of materials in which similar pairs differ on a single, salient, feature while dissimilar pairs differ on a multitude of salient features (see Fig. 1). One group of participants carried out a same-different task (S/D) and the other performed a difference-identification task, in which they were to type out a specific difference between the two images.

Structure-mapping theory predicts that the two tasks will show opposite patterns with respect to similarity. For the S/D task, participants need only complete the first stage of the alignment process to determine that the dissimilar images are different, and should therefore be faster to make a ''different'' judgment for these than for similar ones. However, because the difference-identification task requires aligning the images, participants should be faster to respond to a similar pair than to a dissimilar one. A third prediction is that response times should be longer for the difference-identification task than for the S/D task. This third prediction is less telling than the other two, because (a) it does not differentiate structure-mapping from other models of comparison and (b) the result could simply stem from the fact that the difference-identification task requires verbalization. Nonetheless, failure to find this pattern would be problematic for our account.

2.1. Method

2.1.1. Participants

Forty-four undergraduate students at Northwestern University were randomly assigned to the two conditions: 24 to the S/D condition and 20 to the difference-identification condition.

2.1.2. Materials

The materials were 60 images designed in the likeness of heraldic shields. Forty of the images were pairs of highly similar and alignable images. In these high-similarity pairs, the two images differed in a single design element (e.g., the crest, a central component, etc.). These 20 pairs were then combined into groups of two pairs, such that the images of one pair would be highly dissimilar to the images of the other pair (see Fig. 1). An independent group of 14 raters provided similarity ratings on these image pairs. All participants rated the high-similarity image pairs (M = .82) as more similar than the low similar image pairs (M = .38). For each group, half the participants viewed the two high-similarity pairs (e.g., A&B, C&D) and the other half viewed the two low-similarity pairs (e.g., A&C, B&D). The remaining 20 images were used to create 20 pairs of identical images (''same'' pairs). Each participant saw 10 high-similarity pairs and 10 low-similarity pairs, as well as the 20 ''same'' pairs.

Finally, 10 further pairs (5 identical and 5 non-identical) consisting of arrangements of geometrical forms were used for training.

2.1.3. Procedure

The experiment was presented by computer. Participants read the instructions, completed a training phase, and then went on to the main task, presented in two blocks of equal length.

For both groups, the presentation of each pair was preceded by a half-second fixation period during which a crosshair appeared at the center of the screen. In the S/D task, participants received 20 pairs in each block (half same and half different). For each image pair participants performed a same-different judgment (i.e., identical or non-identical) by pressing the left or right control key (with left-right assignment counterbalanced). The time between the onset of presentation and the response was recorded.

For the difference-identification task, participants received only the 20 different pairs (10 in each block). For each image pair, participants pressed the space key after identifying a difference. The image pair then disappeared and the participant typed in the response. The time between the onset of presentation of an image pair and participants' pressing the space bar was recorded.

2.2. Results and discussion

Only correct ''different'' responses were used in the analysis. This excluded approximately 6.5% of the ''different'' responses. The mean results are shown in Fig. 3. As predicted by structure-mapping theory, the two tasks showed opposite patterns: participants were faster to make a ''different'' judgment for low-similarity pairs than for high-similarity pairs but were slower to identify a difference between low-similarity pairs than between high-similarity pairs.

The median response time for each of the two types of experimental pairs was calculated for each participant and the results were analyzed using a repeated-measures ANOVA of Task(between-s) × Similarity(within-s).[6] This analysis confirmed the predicted interaction between task (S/D judgment vs. difference-identification) and similarity (high-similarity vs. low-similarity), $F(1,42) = 16.41$, $MS_e = 1.07$, $p < .001$, $\eta^2 = .25$.

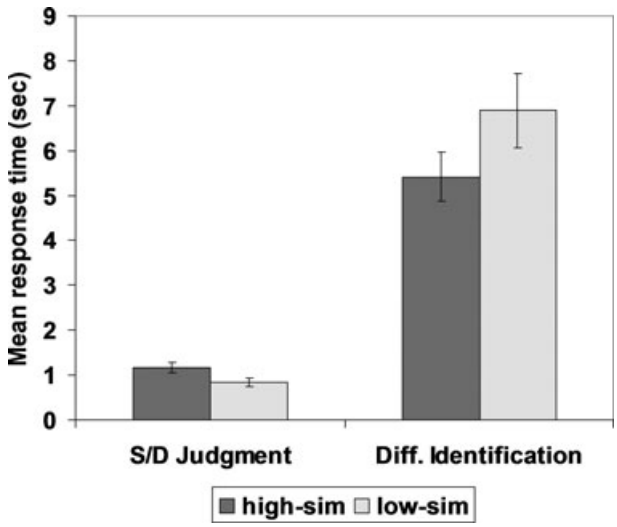
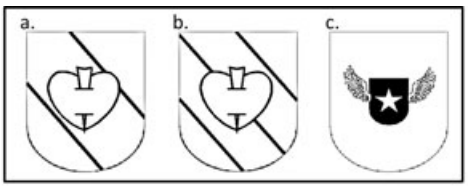


Fig. 3. Results of Experiment 1 (error bars represent the standard error of the mean).

Additionally, we found the predicted main effect for task ($F(1,42) = 78.85$, $MS_e = 7.34$, $p < .001$, $\eta^2 = .78$; $M_{\text{same-different}} = 1.00$, $M_{\text{difference-identification}} = 6.15$). There was also a main effect of similarity ($F(1,42) = 7.07$, $MS_e = 1.07$, $p < .05$, $\eta^2 = .10$). Planned comparisons revealed that the observed performance differences in response time across similarity levels were reliable for both tasks (one-tailed paired-samples t-tests; S/D judgments, $t(23) = 11.48$, $p < .001$, $d = 2.41$; difference-identification, $t(19) = 3.02$, $p < .01$, $d = .68$). An item ANOVA also showed an interaction between task and similarity ($F(1,76) = 16.07$, $MS_e = 1.45$, $p < .001$, $\eta^2 = .016$) and a main effect of task ($F(1,76) = 431.80$, $MS_e = 1.45$, $p < .001$, $\eta^2 = .30$) and similarity ($F(1,76) = 6.99$, $MS_e = 1.45$, $p < .05$, $\eta^2 = .005$).

2.2.1. Difference listings

Our prediction of faster response times for high-similarity pairs in the difference-identification task is based on the idea that (a) it is faster and easier to align high-similarity pairs than low-similarity pairs; and (b) alignable differences emerge naturally when a pair is aligned (Markman & Gentner, 1993). This reasoning also predicts that participants should provide more alignable differences for high-similarity than for low-similarity pairs. To test this, we examined all the differences participants noted for the 20 pairs of similar images and 20 pairs of dissimilar images—a total of 400 differences (18 of which were non-usable) (see Fig. 4). Following Markman and Gentner's (1993) coding system, we coded a difference as alignable if (a) it mentioned contrasting (aligned) properties of the two images (e.g., ''one has black center and the other has a white center'') or (b) it included an explicit comparative construction (e.g., ''the left image has a darker diagonal''). All other differences were considered nonalignable differences, including simple negation of one item's property



Differences	a&b	b&c
Alignable	•cut into thirds vs cut into fourths •three diagonal lines on left, two on right •There is an extra stripe on the left •3 diagonal strips in background in left vs 2 on right •one has three stripes instead of 2 •Left has 3 diagonal lines, right has 2 •the second has one last stripe in the back •one has three lines and the other has two lines •One has three lines and the other only has two •The left image has another diagonal	•wings vs lines •diff emblem in center of sheild •1 is more professional •left has pierced heart and stripes, right has winged shield •The image in the middle of one is much smaller than the other •crest on left has star, crest on right a heart
Nonalignable		•the shield on the left doesn't have a heart being stabbed •Object on left is striped •one has wings •The left image has three lines running across it

Fig. 4. Sample differences listed in Experiment 1 for alignable pair and nonalignable pair.

as applied to the other (e.g., ''one has a dragon, the other does not''). As expected, participants gave more alignable differences for the similar image pairs (67%) than the dissimilar ones (56%), $\chi^2(1) = 5.39, p < .05$.

The results bear out the predictions of structure-mapping theory: Participants were faster to distinguish two images when they were dissimilar, but slower to identify a specific difference between them. Also as predicted, participants listed more alignable differences for similar than for dissimilar pairs.

3. Experiment 2

The results so far are consistent with the predictions of structure-mapping theory. Identifying specific differences was fastest for high-similarity (highly alignable) pairs, even though detecting that a pair was different was faster for low-similarity pairs. At this point one might be tempted to grant structure-mapping the laurels, since it readily predicts the task disassociation that feature models and multidimensional spatial models cannot. However, before drawing such a conclusion, we need to ask whether the results will generalize from the rather artificial materials of Experiment 1 to more naturalistic materials. A particular concern is the fact that the high-similarity pairs differed in only one feature, whereas the low-similarity pairs differed in many features. Defenders of feature-intersection models could argue that people were slow to generate a difference for the low-similarity pairs not because they were hard to align, but because of the need to select from many possible differences. A similar argument applies for the multidimensional space models. In other words, the high-similarity advantage resides in decision processes, not in alignment processes.

Experiment 2 aims to replicate the previous results using more naturalistic stimuli. We used sketches of plants taken from the Dover series (Harter, 2008), which are more complex and variable than the materials of Experiment 1. These have the key advantage that the high-similarity pairs (and the low-similarity pairs) differ in several features. Another advantage in terms of generalizing the phenomenon is that their encoding is more likely to draw on real-world knowledge; for instance, the identification of an image as a flower contributes to the identification of its parts as petals, whereas the same parts might be identified as leaves in an image of a bush.

In addition, we equated the time allotted for examining the pairs between the two tasks. Rather than having the pair stay on the screen until the participant responded, we presented the image pairs on the screen for 1,500 ms. Participants could respond at any time after the pair was presented. For both groups, response time was measured from stimulus onset until the first key press.

3.1. Method

3.1.1. Participants

Eighty undergraduate students at Northwestern University were randomly assigned to the two conditions: 40 to the S/D condition and 40 to the difference-identification condition.

3.1.2. Materials

The materials were 60 detailed drawing of plants, organized into sets of four as for the previous experiments. Forty drawings were used for the experimental stimuli. As shown in Fig. 5, each drawing belonged to both a high-similarity pair and a low-similarity pair. Similarity ratings were collected as in Experiment 1; all 14 independent raters rated the high-similarity image pairs (M = .27) as more similar than the low-similarity pairs (M = .54). Participants saw each drawing only once (in either a high-similarity pair or a low-similarity pair). Twenty additional drawings were used to create 20 ''same'' pairs. The pattern of presentation for the drawings followed that of Experiment 1. Most notably, each participant saw 10 high-similarity pairs and 10 low-similarity pairs.

3.1.3. Procedure

The procedure for Experiment 2 was similar to that for Experiment 1. However, images were displayed for a fixed period of 1500 ms after which they disappeared. Participants in the S/D condition were presented with a blank screen until they made their decision, while participants in the difference-identification condition were presented with a prompt asking them to type a difference. Response time was measured from the onset of the presentation of the images. For participants in the S/D condition, response time was measured from the onset of presentation of the images until they made their choice by pressing the appropriate key. For participants in the difference-identification condition, response time was measured from the onset of presentation of the images until the first key press of their response.

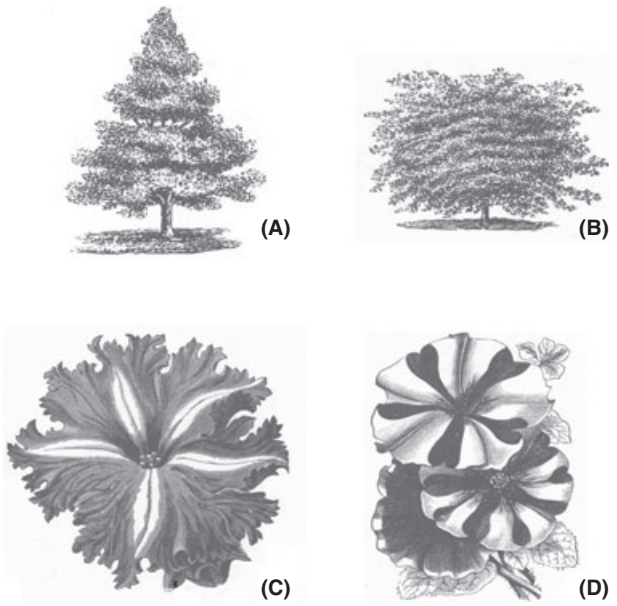


Fig. 5. Sample stimuli from Experiment 2. Images in the same row represent high-similarity pairs; images in the same column represent low-similarity pairs.

3.2. Results and discussion

Only correct responses were included, resulting in the exclusion of 4.0% of the responses in the S/D task. Again, both predictions were upheld: (1) high-similarity pairs were faster than low-similarity pairs in the difference-identification task; and (2) low-similarity pairs were faster than high-similarity pairs in the same-different judgment task (see Fig. 6). As before, the median response times for participants by condition were analyzed using a repeated-measures ANOVA of Task (between-s) × Similarity(within-s). As predicted, there was a main effect of task ($F(1,78) = 29.47$, $MS_e = 1.41$, $p < .001$, $\eta^2 = .27$; $M_{\text{same-different}} = .84$, $M_{\text{difference-identification}} = 1.86$) and an interaction between task and similarity, $F(1,78) = 17.3$, $MS_e = .3$, $p < .001$, $\eta^2 = .17$. There was also a main effect of similarity ($F(1,78) = 8.33$, $MS_e = .3$, $p < .01$, $\eta^2 = .08$).

Planned comparisons revealed that the observed performance differences in response time across similarity levels were reliable for both tasks (one-tailed paired-samples *t*-tests: S/D judgments, $t(39) = 3$, $p < .01$, $d = .49$; difference-identification, $t(39) = 3.6$, $p < .01$, $d = .58$).

An item ANOVA also showed a reliable interaction ($F(1,76) = 18.07$, $MS_e = .064$, $p < .001$, $\eta^2 = .072$) and main effects of task ($F(1,76) = 152.55$, $MS_e = .064$, $p < .001$, $\eta^2 = .60$) and similarity ($F(1,76) = 6.13$, $MS_e = .064$, $p < .05$, $\eta^2 = .024$).

Experiment 2 replicated the results of the previous experiments. Participants find it easy to identify differences between two highly alignable images, but difficult to decide that these two images differ. Furthermore, this result does not appear to depend on the amount of time participants spend looking at the image pairs, but rather depends on whether a pair of images is alignable.

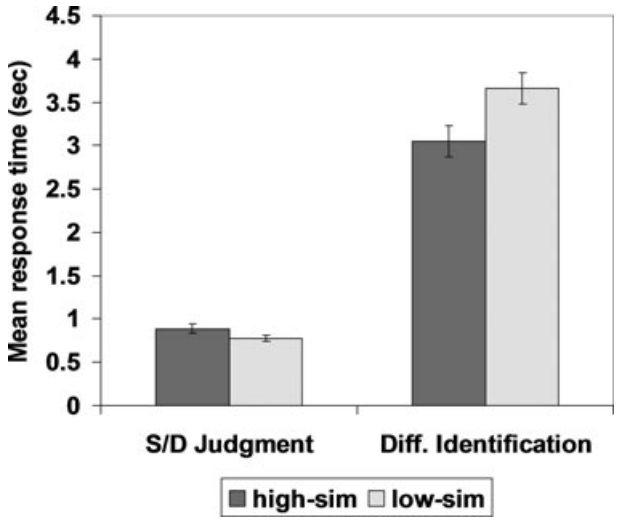


Fig. 6. Results of Experiment 2 (error bars represent the standard error of the mean).

Importantly, both the high-similarity and low-similarity pairs differed in several ways. Thus, these results argue against a possible alternative interpretation of the results of the first two studies. In Experiment 1, the high-similarity pairs differed in only one feature, while the low-similarity pairs differed in many. Thus, the finding of the predicted disassociation—specifically, the longer time for low-similarity than for high-similarity in the difference-identification task—could have resulted from the need to select among candidate differences for the low-similarity, but not the high-similarity, items. The finding that the same pattern holds even when multiple differences exist for both kinds of pairs is evidence against this interpretation. Of course, it could still be argued that there are more potential differences for the low-similarity items than for the high-similarity items, and that this difference accounts for the difference in response time, depending on one's assumptions about how selection time varies with number of candidates. Therefore, in Experiment 3, we adopted a different technique to try to rule out the selection argument.

4. Experiment 3

The results so far are consistent with SMT's prediction that difference-identification should be faster for high-similarity than for low-similarity pairs. But as just discussed, the response time for low-similarity pairs could be inflated by the need to select from among many potential differences. To rule out an explanation based on greater selection time, in Experiment 3 we adopted a precuing method[7] in which participants were told *which* difference to look for in advance for each pair. The images were arrays made up of six simple shapes, each a different color (see Fig. 7). Prior to seeing a pair of images, participants were shown a black shape (say, a square). Then the pair of images was shown. The task was simply to respond whether the two squares were the same color, by pressing the ''same'' key or the ''different'' key. Half the pairs were highly similar in their spatial structure (and hence alignable) and the other half were dissimilar in spatial structure (and hence nonalignable). In each pair, all shapes other than the target shape were identical to each other (in color as well as in shape).

This design removes the selection problem: Participants are told exactly *which* difference to look for and in any case only one shape will differ in color in the ''different'' pairs. Thus, the response time for low-similarity pairs cannot be inflated by the need to select from several potential differences. Nonetheless, SMT predicts faster responding for high-similarity pairs, because they are easier to align. When the overall arrays are aligned, the differently colored targets will constitute an alignable difference and will pop out to participants. This prediction is particularly interesting because aligning the arrays is purely optional here; the task only requires attending to the precued shape.

To summarize, if we find the same pattern of responding as in Experiments 1 and 2—namely, longer response times for low-similarity than for high-similarity pairs—then this will be strong evidence for structural alignment processes in perceptual comparison. If, on the other hand, the results of Experiments 1 and 2 are due to differential selection time, then no such difference should be observed in Experiment 3.

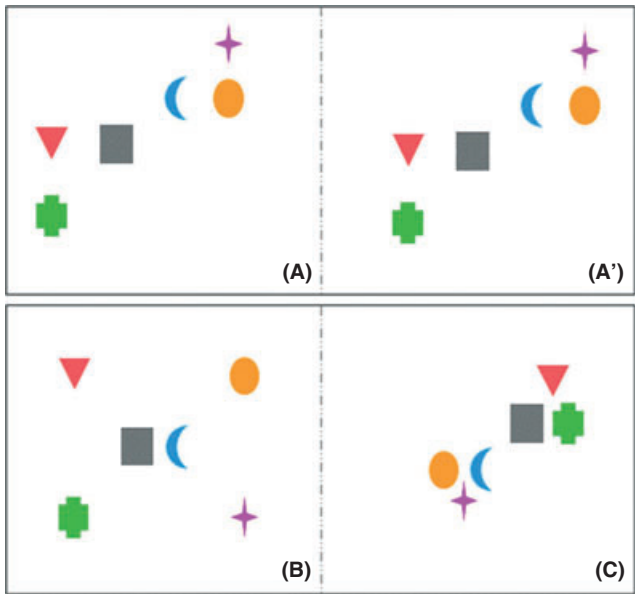


Fig. 7. Sample stimuli from Experiment 3. Each of the geometric shapes (e.g., square, triangle, etc.) was displayed using a different color. The target shape for this sample is a square. A–C represent the three images that comprise a set. In this case A is used in the structurally alignable condition (A–A′) and B and C are used in the non-structurally aligned condition.

4.1. Method

4.1.1. Participants

Sixty undergraduate students at Northwestern University participated. The design was within-subject (except for counterbalancing), with two factors: Alignability (High/Low) × Color of Target Shapes (Same/Different).

4.1.2. Materials

The materials were 48 images, each composed of six geometrical shapes surrounded by a frame. Within each image, all six shapes were of different colors. These images were divided into 16 groups of three (triads) (see Fig. 7). Within each triad all of the images used the same six shapes. For each triad, one of the constituent shapes was selected as the *target shape*. In the ''same'' condition, the target shapes were the same color; in the ''different'' condition, the target shapes differed in color. For each participant, half the triads were assigned to the ''same'' condition and the other half to the ''different'' condition.

For each participant, each triad was used to generate two image pairs. One of the image pairs (the alignable pair) was constructed by using the same original image for both of the images in the pair; thus, these images were identical except for the target shape, which could be same or different in color. The second image pair (the nonalignable pair) consisted

of the other two images. There were therefore three possible ways in which the two image pairs could be constructed from each triad. This was counterbalanced across participants.

4.1.3. Procedure

The experiment was presented by computer. Participants read the instructions, completed a training phase, and were then presented with the experimental pairs. These pairs were presented in two blocks of equal length.

On each trial, participants were first presented with a black geometric shape (the precue) that appeared in the center of the screen for three seconds. This was the target shape that participants needed to respond to. For example, if the precue was a square, then the task was to respond ''same'' if the two squares had the same color, and ''different'' if they did not. The precue was followed by the presentation of the two images, on the left and right side of the screen (with left and right randomly assigned). To avoid having the images at the same height (which might inflate the alignment results), each image's height on the screen was also determined randomly. Each of the distinct geometric shapes to be presented was randomly assigned a color from the palette. Within each pair, corresponding shapes (other than the target shapes) were always the same color.

For each image pair, participants pressed ''same'' if the precued shape had the same color in both images, and ''different'' if the precued shape had a different color. The ''same'' and ''different'' keys were the left and right control keys (with left-right assignment counterbalanced). The time between the onset of presentation and the response was recorded.

4.2. Results and discussion

Four participants (6.67%) were dropped because they provided fewer than three correct responses in one or more of the conditions. For the remaining participants, only correct responses were used in the analysis. This excluded approximately 12% of the responses. For each condition, the median RT scores for each participant were computed and used in the analysis below. The means of these medians are shown in Fig. 8.

As predicted, participants were faster to make both ''same'' and ''different'' judgments for components of alignable (structurally identical) pairs than for components of nonalignable (structurally different) pairs. The results were analyzed using a repeated-measures ANOVA of Structural Similarity (alignable vs. nonalignable) × Response Type (''same'' or ''different''). As predicted, there was a reliable main effect of alignability ($F(1,55) = 26.24$, $MS_e = 0.13$, $p < .001$, $\eta^2 = .023$). There was no reliable effect of response type ($F(1,55) = .076$, $MS_e = 0.18$, *n.s.*, $\eta^2 = 0$) nor interaction ($F(1,55) = .77$, $MS_e = 0.043$, *n.s.*, $\eta^2 = 0$).

Planned comparisons revealed that the observed performance differences in response time across structural similarity levels were reliable for both ''same'' and ''different'' responses (one-tailed paired-samples *t*-tests: ''same,'' $t(55) = 3.78$, $p < .001$, $d = .46$; ''different,'' $t(55) = 5.15, p < .001, d = .82$).

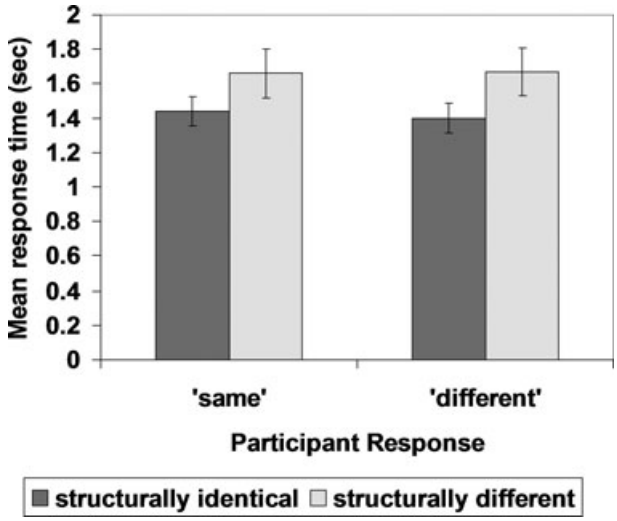


Fig. 8. Results of Experiment 3 (error bars represent the standard error of the mean).

An item ANOVA showed a reliable main effect of alignability ($F(1,46) = 8.72$, $MS_e = 0.056$, $p < .01$, $\eta^2 = .16$) but no reliable effect of response type ($F(1,46) = 1.86$, $MS_e = 0.026$, *n.s.*, $\eta^2 = .04$) nor interaction ($F(1,46) = .008$, $MS_e = .025$, *n.s.*, $\eta^2 = 0$).

The results of Experiment 3 again bear out the prediction that it is easier to identify specific differences between images that are structurally alignable than between images that are nonalignable. More important, it shows that the positive effect of alignability on the speed of identifying a difference does not depend on a selection effect whereby the difference-identification response time for low-similarity pairs is elevated by having to choose among several possible differences. Even when participants know exactly which difference to focus on, they are faster to identify this difference when the overall arrays are readily alignable.

5. Experiment 4

Experiments 1–2 demonstrated a dissociation in the effects of similarity on two seemingly related tasks: detecting *that* two figures differ and identifying *how* they differ. Experiment 3 showed that this dissociation cannot be attributed to a post-comparison difference in the selection time required for the difference-identification task. Therefore, it seems that this dissociation arises during the comparison process itself. These results are consistent with structure-mapping theory and with its multistage process model, SME. In the difference-identification task, participants are faster to respond to a similar pair than to a dissimilar one because similar pairs are more easily aligned. In the S/D task, participants are faster to make a ''different'' judgment for dissimilar pairs than for similar pairs because dissimilar pairs can be rejected in the first stage.

In Experiment 4, we test more specific predictions of the process model—specifically, predictions concerning the fine structure of similarity. Our studies so far have considered only the overall similarity between two images. In Experiment 4, we distinguish object similarity from relational similarity, based on prior evidence for their differential contributions to mapping tasks (Gentner & Kurtz, 2006; Krawczyk, Holyoak, & Hummel, 2004; Markman & Gentner, 1993). *Object similarity* refers to the number of matching objects (or more precisely, the number of matching object attributes) in the two images being compared. *Relational similarity* refers to the degree of relational overlap in the images, that is, whether the images contain the same set of spatial relations between objects. In the previous experiments, the number of object matches and the relational overlap of the images were strongly correlated. High-similarity pairs were alike both in their relational structure (hence, alignability) and in their object features, and low-similarity pairs were low in both. While this correlation is fairly typical in real-world experience (as well as in prior research on S/D judgments), we must go beyond it to test our model's predictions fully.

We can now sharpen the predictions for the two tasks. Given pairs that vary orthogonally in object similarity and relational similarity (see Fig. 9), our model predicts that both object

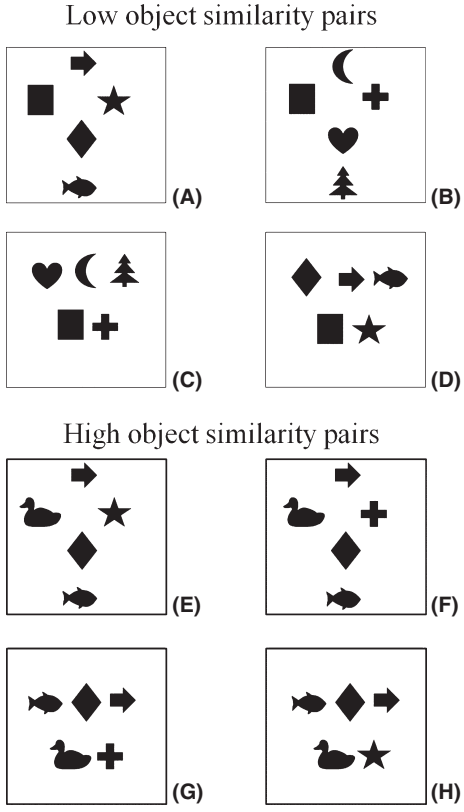


Fig. 9. Sample stimuli from Experiment 4. Within each set, images in the same row represent high relational similarity pairs; images in the same column represent low relational similarity pairs.

similarity and relational similarity will affect the S/D task. That is, people can reject pairs in the initial stage of matching either when there are few local matches between object attributes or when there is little relational overlap. In contrast, the difference-identification task will be affected only by relational similarity, and in the opposite direction. This is because fast responses depend on aligning the two images to identify an alignable difference between them; and alignment depends critically on relational similarity. In sum, the SME process model predicts that high relational similarity (high alignability) will result in slower S/D responding and faster difference-identification responding. High object similarity will result in slower S/D responding and will not influence the speed of difference-identification.

5.1. Method

5.1.1. Participants

Fifty-three undergraduate students at Northwestern University participated, 20 in the S/D condition and 33 in the difference identification condition.

5.1.2. Materials

The materials were 60 images, each composed of five distinct objects (silhouettes) surrounded by a frame. Forty of the images (20 pairs) were designed such that in both images the spatial organization of the objects was highly similar (e.g., the rows in Fig. 9). In half of these pairs (the *high object similarity pairs*), four of the five objects were shared between the two images, while in the other half (*low object similarity pairs*), only one of the five objects was shared. The 20 pairs were then combined into groups of two pairs that differed in their spatial organization[8] but included the same objects (e.g., Fig. 9, A–B/C–D). This allowed us to create all four kinds of pairs: pairs with high relational similarity (high-alignable pairs) and low object similarity (e.g., A–B and C–D); pairs with low relational similarity (low-alignable pairs) and high object similarity (A–D and B–C); pairs high in both relational and object similarity (E–F and G–H) and pairs low in both (A–C and B–D). The remaining 20 images were used to create 20 pairs of identical images (''same'' pairs).

The 20 pairs were then combined into groups of two pairs that differed in their spatial organization but included the same objects. In half of these double pairs (*low object similarity pairs*), the images comprising a pair differed on four of their five constituent objects (e.g., Fig. 9, A–B/C–D). In the other half, the images in a pair differed on only one of their five constituent objects (e.g., E–F/G–H). This allowed us to create the four kinds of pairs defined by crossing relational similarity with object similarity: high relational similarity/high object similarity (e.g., E–F and G–H), high relational similarity/low object similarity (A–B and C–D), low relational similarity/high object similarity (E-G and F-H), and low relational similarity/low object similarity) (A–C and B–D). The remaining 20 images were used to create 20 pairs of identical images (''same'' pairs).

Each participant saw 20 ''different'' pairs: five from each of the four experimental conditions (*high relational similarity/high object similarity*, *high relational similarity/low object similarity*, *low relational similarity/high object similarity*, *and low relational similarity/low object similarity*). In addition, participants in the S/D condition were also given the 20

''same'' pairs. Finally, ten pairs (five identical, five non-identical) consisting of arrangements of geometrical forms were used for training.

5.1.3. Procedure

The experiment was presented by computer. After completing a training phase, participants received the experimental pairs in two blocks of equal length. Each pair was preceded by a half-second fixation period during which a crosshair appeared at the center of the screen. The pair remained on the screen for 3000 ms.

In the S/D condition, participants judged whether the pair was identical or non-identical by pressing the left or right control key (counterbalanced). In the difference-identification condition, participants typed in a difference between the two images. When the participant responded (by making a S/D judgment or by starting to type a difference) or 3000 ms elapsed, the presented pair disappeared from the screen. In the difference-identification condition, participants were then presented with a screen where they typed (or continued typing) the difference they had identified. (Participants in the difference-identification task were free to start typing at any time after the pair was presented; whenever they started, they saw a screen displaying what they had typed.) As in Experiment 2, for both tasks, the time between the onset of presentation of the pair and the response was recorded.

5.2. Results and discussion

Only correct ''different'' responses were used in the S/D analysis. This excluded approximately 9% of the ''different'' responses. Trials in which participants viewed different image pairs but responded ''same'' were also removed (approximately 17% of the responses to different image pairs). The median response time for each condition was then computed for each participant and each item. These medians provided the data for the statistical analysis; their condition means are shown in Fig. 10.

As predicted, the two tasks showed different response patterns. In the S/D task, participants were faster to say ''different'' for pairs with low object similarity than for pairs with high object similarity, and for pairs with low relational similarity (different spatial array) than for pairs with high relational similarity. In contrast, participants in the difference-identification condition were faster to identify a difference for pairs with high relational similarity than for those with low relational similarity. Their performance showed no effect of object similarity. As in the prior studies, S/D judgments were much faster than difference-identification (which took more than twice as long).

Repeated-measures ANOVAS of Object Similarity × Relational Similarity for each task bore out these patterns. There was a significant effect of relational similarity in both tasks (though in opposite directions). (Same-different: $F(1, 19) = 36.32$, $MS_e = .057$, $p < .01$, $\eta^2 = .23$; Difference-identification: $F(1, 32) = 7.18$, $MS_e = .25$, $p < .05$, $\eta^2 = .028$.) However, object similarity only affected performance in the S/D task. (Same-different: $F(1, 19) = 48.07$, $MS_e = .022$, $p < .01$, $\eta^2 = 0.12$; Difference-identification: $F(1, 32) = .12$, $MS_e = 0.17$, *n.s.*, $\eta^2 = 0$.) Likewise, the two variables showed a statistically significant interaction for the S/D task, but not for the difference-identification task. (Same-different:

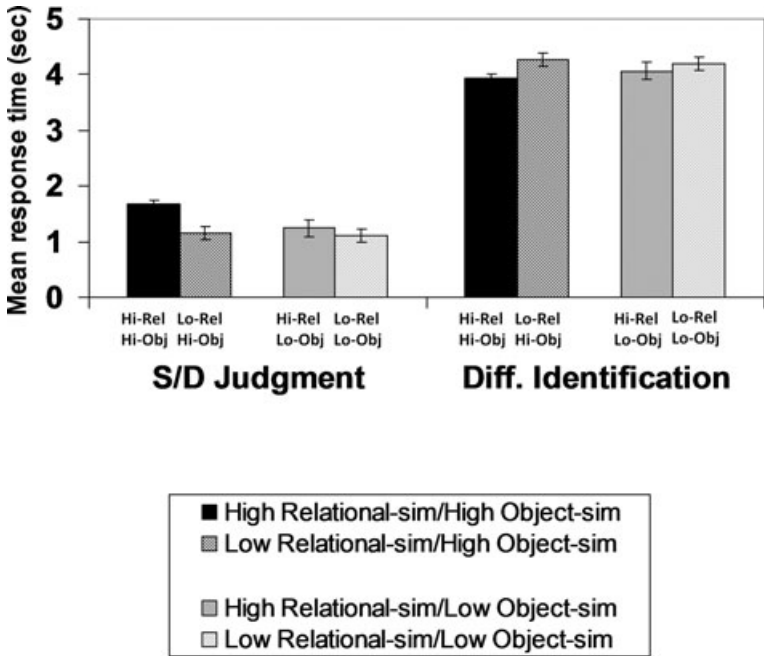


Fig. 10. Results of Experiment 4 (error bars represent the standard error of the mean).

$F(1, 19) = 20.84$, $MS_e = .037$, $p < .01$, $\eta^2 = 085$; Difference-identification: $F(1, 32) = 2.5$, $MS_e = .14$, *n.s.*, $\eta^2 = .005$.)

Item ANOVAs for the two tasks showed similar patterns. There was a main effect of relational similarity (again in opposite directions) on both tasks (Same-different: $F(1, 36) = 26.22$, $MS_e = .023$, $p < .001$, $\eta^2 = .21$; Difference-identification: $F(1, 36) = 19.29$, $MS_e = .023$, $p < .001$, $\eta^2 = .34$); and a main effect of object similarity only on the S/D task (Same-different: $F(1, 36) = 21.68$, $MS_e = .023$, $p < .001$, $\eta^2 = .42$; Difference-identification: $F(1, 36) = .10$, $MS_e = .023$, *n.s.*, $\eta^2 = .002$). As in the subject analysis, the interaction was significant only for participants in the S/D task (Same-different: $F(1, 36) = 18.96$, $MS_e = .023$, $p < .001$, $\eta^2 = .18$; Difference-identification: $F(1, 36) = .77$, $MS_e = .023$, *n.s.*, $\eta^2 = .014$).

5.2.1. Alignability ratings for differences produced

Finally, two raters blind to the hypothesis and similarity condition rated whether the differences identified by participants were alignable, using the same guidelines as for Experiment 1 (following Markman & Gentner, 1993). The two raters agreed on 84% of the differences and only the differences on which the two raters agreed were used in the analysis below.

As predicted, participants were more likely to produce alignable differences for pairs with high relational similarity (47% of the time) than for pairs with low relational similarity (15% of the time), $\chi^2(1) = 51.12$, $p < .001$. For example, when comparing the high-alignable image pair E-F in Fig. 9 and 12 out of 17 participants (70%) produced an alignable

difference that contrasted the star in one image with the plus sign in the other (e.g., ''Right had star instead of a plus sign''). In contrast, when comparing the low-alignable pair E-G, only 4 out of 14 participants (29%) produced an alignable difference; not surprisingly, none of these identified the star in one image and the plus sign in the other. This pattern is similar to the pattern found in Experiment 1, as well as to prior findings with both pairs of concepts (Gentner & Gunn, 2001; Gentner & Markman, 1994; Markman & Gentner, 1993) and pairs of images (Markman & Gentner, 1996), in which participants were more likely to produce alignable differences for highly alignable pairs[9]

5.2.2. Summary

These findings further bear out the predicted task dissociation whereby similarity facilitates fast responding in the difference-identification task and hampers it in the S/D task. They also carry the findings to a new level of specificity, by delineating distinct effects of object similarity and relational similarity. For the S/D task, high similarity—both object similarity and relational similarity—is associated with slow responding, consistent with the claim that pairs with very few early local matches can be rejected immediately. For the difference-identification task, high relational similarity (but not high object similarity) is associated with fast responding, consistent with the claim that alignable differences pop out to people only after the alignment process is complete.

6. Computational model

We now present a computational simulation of the two tasks, using SME to carry out the comparison process. We focus on Experiment 4, which offers the most detailed set of findings. We also simulated the results of Experiment 1 (see Appendix 1); however, this simulation is less informative than that of Experiment 4, because in Experiment 1 relational similarity and object similarity were varied together, while in Experiment 4 they were varied orthogonally.

A key goal in our work on similarity and analogy is to escape the need for hand-coding the input. The use of hand-coded representations of the input, as is common practice in cognitive simulations, allows the researcher to tailor the representations to fit the program's capabilities (Gentner & Forbus, 2011; Hofstadter & Mitchell, 1994), with a concomitant loss of credibility in the findings. To escape hand-coded representations, the input to the simulation was created by an automatic encoding system for perceptual images called CogSketch (Forbus, Usher, Lovett, Lockwood, & Wetzel, 2008; Lovett, Gentner, & Forbus, 2006).

CogSketch is a sketch understanding system that automatically encodes representations of two-dimensional images. It can generate representations for sketches drawn on a tablet by a user or (as in the present case) for images imported from PowerPoint. The representations produced by CogSketch include both object descriptions and relations between objects. These include positional relations, which describe one object's location relative to another, and topological relations, which describe cases where one object intersects another or is

located inside another. The representations used in this simulation also include configural groupings of objects (e.g., Love, Rouder, & Wisniewski, 1999; Navon, 1977). Configural groupings, computed based on colinearity and proximity, include rows of three objects or pairs of adjacent objects. These groups provide a rough approximation of the salient configurations within the images.

6.1. Model implementation

The comparison process uses the SME (Falkenhainer et al., 1989; Forbus & Oblinger, 1990). Recall that SME implements Gentner's (1983, 1989) structure-mapping theory as a three-step, incremental process (see Fig. 2). First, it identifies all local identity matches between expressions in the two representations. This is done without regard for consistency; there are typically many mutually inconsistent matches. In stage 2, structural consistency is enforced, and the local matches separate into internally consistent clusters (kernels). In stage 3, the kernels are merged into a large, structurally consistent global mapping. This global mapping—which constitutes the aligned structure between the two analogs—gives rise to alignable differences (differences that occupy the same structural role in the two analogs). When these processes run to completion, the model can both determine whether the two items are the same or different and name a specific difference between them. However, for very dissimilar stimuli, the model can abort the alignment process early, based on finding relatively few initial local matches between the two items. This permits fast ''different'' responses for very low-similarity pairs.

We simulated the two tasks by running SME on the materials used with human participants. The questions of interest are (a) Do the similarity measures generated by SME correlate appropriately with human response times on the tasks? and (b) Does SME show the same dissociation between the tasks as was found in the human data? For the difference-identification task, according to the theory, a full alignment is required. We can gauge the alignability of a pair simply by taking SME's standard structural evaluation score (a measure of similarity that takes into account structural overlap as well as featural matches) for the global mapping.[10] Recall that high-similarity pairs are characterized by high structural evaluations and by fast alignment times (because they generally require only a single greedy merge pass). Thus, we should find that high structural evaluation scores predict fast difference-identification responses. For the S/D task, the dominant factor in predicting timing is whether the alignment process can be aborted early, permitting very fast ''different'' responses. For this, we needed to devise a measure of initial local matches, as described below. Thus, the prediction is that for low-similarity pairs, there will be many fewer initial matches than for high-similarity pairs, permitting early termination and fast ''different'' responses for low-similarity pairs.

6.2. Simulation

We evaluated the model by running it directly on the materials used with human participants, using the encodings automatically generated by CogSketch.

6.2.1. Procedure

We ran the model on the same 40 pairs as were used in Experiment 4[11] The images were imported directly into CogSketch as bitmaps from the same PowerPoint stimuli given to the participants. CogSketch automatically constructed an object representation for each PowerPoint shape[12] and encoded spatial relations (such as left of) and configural groupings applying to two or three shapes (such as row).

We considered two measures generated by SME for each pairing. Both measures were normalized based on the sum of the sizes of the two representations:

Local object matches score: The number of local identity matches found by SME between object attributes in the first stage of SME. This is a derived measure in which consider only the contribution of objects to similarity in SME's first stage.[13]

Mapping score: The structural evaluation score for global mappings computed by SME between the complete representations. This score reflects the overall alignment of the two images, which is chiefly determined by the degree of relational similarity.

The key question is whether SME's patterns will match those of the human participants. Recall that in Experiment 4, our participants showed distinct patterns of similarity effects for the two tasks. For the S/D task, low similarity—either object similarity or relational similarity—is associated with fast responding, consistent with the claim that fast ''different'' responses are possible whenever the first stage of SME's comparison process reveals very few initial matches. For the difference-identification task, high relational similarity (but not high object similarity) is associated with fast responding. For this task, alignability is the key factor, because alignable differences leap out only after the alignment process is complete.

Thus, for the simulation to match human performance the results should be as follows:

(1) For the S/D task, both the local object matches score and the mapping score should correlate positively with human response times, because fast ''different'' responses can be made when either object similarity or relational similarity is low; (2) for the difference-identification task, SME's mapping score—which reflects the degree to which the two images are aligned—should correlate negatively with human response times (that is, high scores should lead to fast response times). However, the object matches score should not correlate with performance, reflecting the prediction that object similarity will (in general) not play a role in difference identification.

6.2.2. Results

Fig. 11 provides scatterplots that plot the predictions of the model against the mean response times from Experiment 4. We consider each of these predictions in turn. First, for the S/D task, both mapping score and local object matches show a strong positive correlation with response times (mapping score: $r = .63$, $p < .01$ local object matches: $r = .59$, $p < .01$). This fits with the pattern that humans require more time to say ''different'' for more similar pairs (both at the object level and at the relational level).

For the difference-identification task, as predicted, mapping score shows a strong negative correlation with response time: that is, a high structural alignment score is associated with fast response times ($r = -.54$, $p < .01$). A further prediction is that we should see no

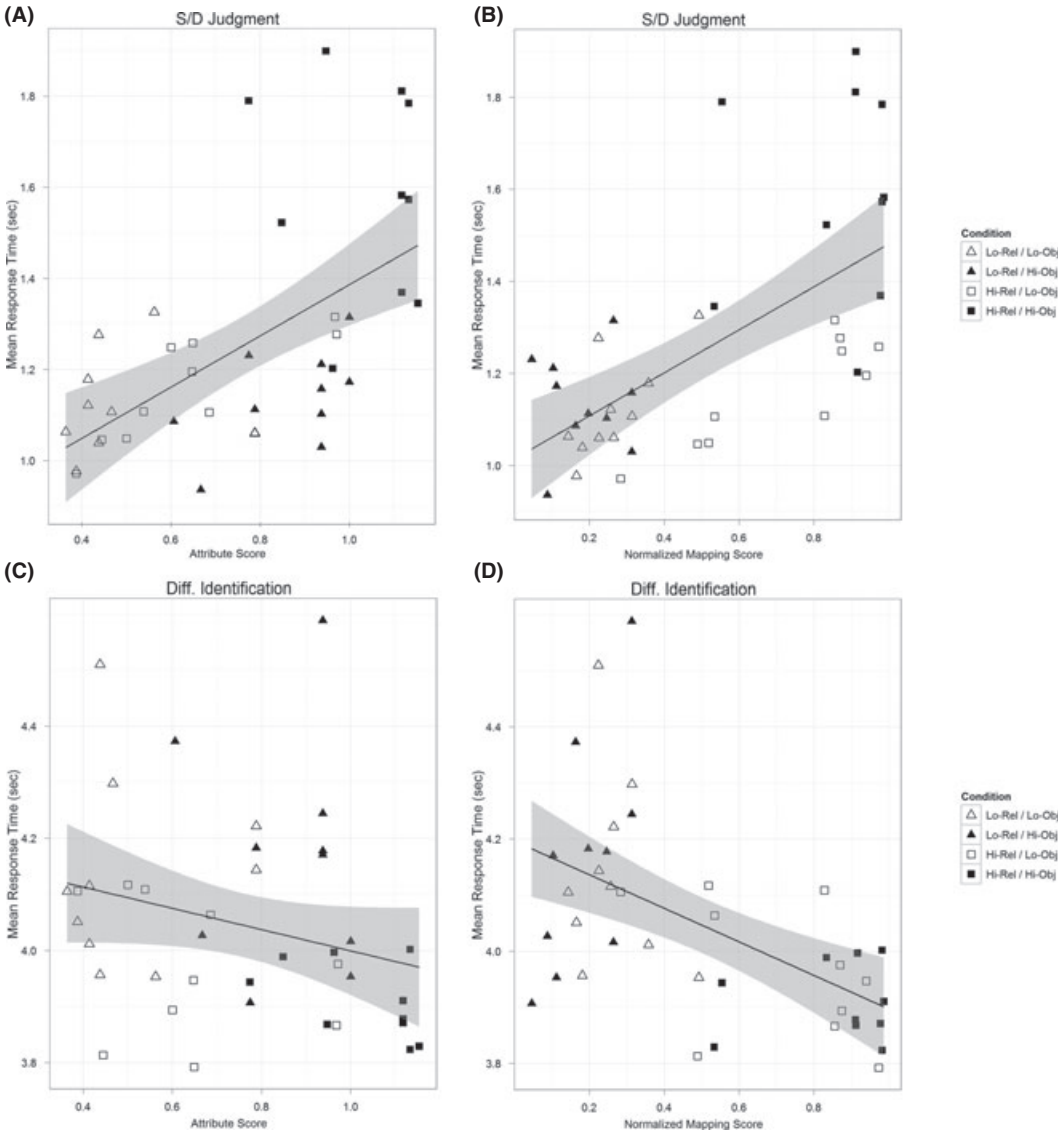


Fig. 11. Scatterplots of the response times in Experiment 4 plotted against the predictions of the model. Results for the S/D task are shown in the top line (A and B), and for the difference-identification task in the bottom line (C and D). (A) and (C) plot the results against SME's local object matches score, (B) and (D) plot the results against SME's mapping score. The line in each plot shows the fitted regression line and the gray area around it represents the 95% confidence interval.

correlation between the local object matches score and the difference-identification response time. This correlation is nonsignificant ($r = -.27$, $p = .09$), producing the expected dissociation.[14] This fits with the pattern that noticing a specific difference in general requires aligning the pair.

These results demonstrate that SME generates measures that correlate with human performance on both similarity tasks. Furthermore, its measure of object similarity shows the expected dissociation between the tasks: It correlates with performance on the S/D task, in which object similarity matters, but not on the difference-identification task, in which object similarity is far less important.

6.2.3. Further analyses

In the above analyses, we used SME's mapping score as a measure of relational similarity in the S/D task. However, this is an approximate measure, because the full structural evaluation score is only computed at the end of the alignment and includes credit for depth of structure as well as for sheer number of matches. To better approximate the number of local relational matches in the initial stages of SME's operation, we devised another measure. Based on prior research suggesting that relational configurations—such as rows and columns of objects—are encoded and matched early in perceptual similarity tasks (Love et al., 1999), we computed the number of local matches between configural groupings of objects. This configural matches score is the number of local matches at the first stage of SME between configurations such as horizontal rows of objects or vertical pairs[15] of objects.

Our prediction was that the configural matches score would correlate positively with response time in the S/D task, because quick ''different'' responses are possible only when the number of initial matches is low. This prediction was borne out ($r = .62$, $p < .01$). Importantly, the configural matches score also correlated negatively with response times in the difference-identification task ($r = -.57$, $p < .01$), consistent with the idea that configural matches contribute to relational similarity.

6.2.4. Summary

SME captures the pattern shown by human participants, including the distinct effects of object similarity and relational similarity. For the S/D task, the number of local matches in the first stage (both relational and object attribute matches) correlates negatively with participants' response times. This is consistent with the finding that high similarity—both object similarity and relational similarity—is associated with relatively slow S/D responding, because pairs with many early local matches cannot be rejected in the first stage. For the difference-identification task, high relational similarity (but not high object similarity) is associated with fast responding, consistent with the claim that alignable differences pop out to people only after the alignment process is complete.

7. General discussion

We tested the predictions of structure-mapping theory, and, more specifically, of the model of comparison processing embodied in SME. In this model, comparison processing is accomplished by a three-stage process of (a) identifying local matches; (b) sorting the initial set of matches into structurally consistent kernels; and (c) combining the kernels into one or more large global mappings. This model makes three predictions: (1) saying ''different'' in a same-different task

should be fastest for low-similarity pairs; (2) identifying a specific difference should be fastest for high-similarity pairs; and (3) the same-different task should require far less time than the difference-identification task. Thus, the two highly related difference-processing tasks examined here should show radically different patterns of responding. While non-structural theories of similarity, such as feature-set-intersection and mental distance models, can predict patterns (1) and (3), they do not predict pattern (2) and its disassociation with (1).

In Experiment 1, we found evidence for all three predictions, using pairs of simple heraldic shield images. The pairs in Experiment 1 were constructed to differ in only one feature for the high-similarity pairs, and in many features for the low-similarity pairs. In Experiment 2, we generalized the findings to a more natural situation by using complex drawings of plants, for which both high- and low-similarity pairs had many differences. Again, all three predictions were borne out—evidence for the robustness of the phenomenon. However, nonstructural theories of similarity could still be maintained, by arguing that the disassociation between patterns (1) and (2) is not inherent in the comparison process, but rather stems from post-comparison choice processes. That is, it could be that the greater time to identify a difference for low-similarity pairs simply results from the difficulty of choosing which of the many potential differences to name. Experiment 3 was designed to rule out this possibility. In that study, participants were precued as to which difference to report. Even though participants knew in advance which shape to report on (same-color or different-color), they still showed an advantage for alignable over nonalignable pairs. This difference cannot be accounted for by post-comparison selection time, but it is predicted by the structure-mapping process model. Because detecting a specific difference is fastest when the difference appears as an alignable difference, people will be faster in this task with pairs that are easily aligned than with those that are not.

In Experiment 4, we went beyond the simple dichotomy between high- and low-similarity and tested our predictions at a more fine-grained level. Structure-mapping makes a distinction between relational similarity—similarity between the patterns of relations in the two items—and object similarity—similarity in the elements within the two items. This distinction is important because it is relational similarity that determines whether a pair can be aligned. In Experiment 4, we independently varied object similarity and relational similarity, and found the predicted pattern. First, as in the prior three studies, the two tasks showed a reverse relation with similarity, with S/D responding slower for high similarity and difference identification faster for high similarity. More tellingly, we also found the predicted effects of specific kinds of similarity. For the S/D task, low similarity—either object similarity and/or relational similarity—led to fast responding, consistent with the claim that pairs with very few early local matches can be rejected immediately. For the difference-identification task, only high relational similarity led to fast responding, consistent with the claim that a full alignment is required in order for alignable differences to pop out. Thus, we must add a fourth pattern to the three noted above: Pattern (4) is that difference identification is preferentially sensitive to relational similarity and not object similarity.

We verified that SME's three-stage comparison process successfully simulates both Experiment 1 (in Appendix 1) and Experiment 4. In both simulations, SME captured the pattern of relatively rapid S/D responding that is fastest (at detecting difference) for highly

dissimilar pairs; and relatively slower difference-identification responding that is fastest for highly similar pairs. SME also captured the finding from Experiment 4, that S/D responding is sensitive to any kind of local similarity—whether at the object level or at the relational level—while difference identification is sensitive only to relational similarity. Importantly, the representations used by SME were automatically generated by an independent sketch-understanding system, Forbus et al.'s (2008) CogSketch; thus, the results do not arise from tailoring the representations to fit our predictions.

These findings of a task disassociation are difficult to reconcile with nonstructural accounts of representation and comparison, such as feature-set-intersection models and mental distance models. Both of these predict a positive relation between the two tasks. In featural models, the fewer the differences that exist between two objects, the harder it should be *both* to detect that they are different and to identify a specific difference between them. Likewise, in mental distance models, the fewer the dimensions of difference, and the smaller the distance along a given dimension, the harder both tasks should be. Nonstructural models also have no way to capture our fourth pattern that the speed of difference identification depends specifically on relational similarity. Given the centrality of comparison processes in human cognition, these findings add to the case for structured models of representation (Gentner & Markman, 1995; Holyoak & Hummel, 2000; Jones & Love, 2007; Markman & Dietrich, 2000).

In sum, our findings suggest that S/D judgments are qualitatively different from the identification of differences. More specifically, the alignability of the images plays a large role in the identification of differences, but not in S/D judgments. This dissociation between the tasks is best explained by positing structural comparison processes, as in structure-mapping theory.

8. Conclusions

Similarity comparison is fundamental to human cognition and perception. It is central in recognition and categorization, in decision making, and in learning and transfer. While the importance of finding commonalities is widely recognized in conceptual structure (e.g., Goldstone, 1994; Murphy, 2002; Smith & Medin, 1984), the role of differences—particularly alignable differences—is also crucial to an understanding of category structure (Markman & Wisniewski, 1997). Further, individual differences in propensity to attend to alignable differences may signal the degree to which people attend to conceptual structure rather than simply attending to associative strength (Golonka & Estes, 2009). The present results support the idea that human comparison processes are multistage computations that operate over complex structured representations.

Notes

1. There are, of course, limits on this prediction. For example, if we compare *a small white pebble* with *a small black pebble plus an elephant*, the first difference we're

likely to notice is elephant/no elephant—a nonalignable difference. But assuming roughly comparable intrinsic salience, alignable differences will stand out.

2. In the current implementation of SME, this process is serial (but very fast). However, we posit that it may proceed in parallel in the brain.
3. If the two largest kernels are similar in size, SME computes the global match twice—once starting from the largest kernel and one starting from the second-largest kernel (see Forbus et al., 1994, for details).
4. When the largest kernel is sufficiently greater in its structural evaluation than the next-largest, it is not necessary to compute a second merge.
5. The prediction that a larger match will be faster than a smaller match results from the key starting assumption of structured representations; it is impossible to make this prediction if one assumes independent-feature representations.
6. We chose to use the per-participant/condition median rather than the mean to reduce the effect of reaction time outliers. For a discussion of this and other methods for controlling outliers in reaction time data, see Ratcliff (1993).
7. We thank Mark Beeman and Steve Franconeri for suggesting this method.
8. We assume that the relational structure that participants notice and use in these images will include configural patterns such as *parallel rows of objects* as well as detailed relations between pairs of objects, such as *Left-of(duck, star).*
9. Considering only pairs with high relational similarity, a higher percentage of alignable differences was produced for pairs that were also high in object similarity than for pairs that were low in object similarity (62% vs. 32%, $\chi^2(1) = 20.70$, $p < .001$). This is consistent with prior findings that literal similar matches are easier to align than are purely relational matches (Gentner & Kurtz, 2006).
10. This metric takes into account not only the size of the global mapping but also its depth.
11. We also simulated the results of Experiment 1 (Appendix 1). SME's measures correlate as predicted with human response times. However, because relational similarity and object similarity were conflated in Experiment 1 (that is, there were just two conditions—high and low similarity), the measures generated by SME for the two tasks were highly correlated.
12. Of the ten constituent object shapes used in the psychological experiment, two were made up of multiple shapes in PowerPoint. These were simplified in PowerPoint so that CogSketch would build only one shape for each of them. In addition, two of the 40 images used required slight touching up in PowerPoint.
13. The motivation for considering only the initial object-property matches is that evidence from other studies indicates that in encoding the items, object properties are encoded before relations (Lovett, Gentner, Forbus, & Sagi, 2009).
14. Because this correlation is marginally significant, we investigated further and found that this correlation is due to variance that is almost entirely shared with the mapping score. This is not the case for the correlation between local matches and the results of the S/D task.

15. A vertical (or horizontal) pair, such as *Horizontal-pair (DS)* differs from a binary relation between two objects, such as *Left-of (duck, star)* in that it does not specify the order of its arguments; instead, the arguments have separate *part-of* relations with the group, as in *Part-of (duck, DS).*
16. To ensure that these changes did not alter the pattern of results, we replicated Experiment 1 with a new group of human subjects using these simplified materials and found the same pattern of results.

Acknowledgments

This research was supported by Office of Naval Research grant N00014-02-1-0078 and by NSF SLC grant SBE-0541957 awarded to the Spatial Intelligence and Learning Center (SILC). We thank Steve Franconeri for valuable advice throughout the project. We also thank Satoru Suzuki and Mark Beeman for helpful suggestions, and Kathleen Braun for her inestimable help with the research throughout the project.

References

Bassok, M. (1990). Transfer of domain-specific problem-solving procedures. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *16* (3), 522–533.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*, 1–63.

Farell, B. (1985). Same-different judgments: A review of current controversies in perceptual comparisons. *Psychological Bulletin*, *98*, 419–456.

Forbus, K., Ferguson, R., & Gentner, D. (1994). Incrementalstructure-mapping. In A. Ram & K. Elselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 313–318). Hillsdale, NJ: Erlbaum.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19* (2), 141–205.

Forbus, K., & Oblinger, D. (1990). Making SME Greedy and Pragmatic. In: Proceedings of the 12th Annual Meeting of the Cognitive Science Society.

Forbus, K., Usher, J., Lovett, A., Lockwood, K., & Wetzel, J. (2008). CogSketch: Open-Domain Sketch Understanding for Cognitive Science Research and for Education. In: Proceedings of the Fifth Eurographics Workshop on SBIM.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199–241). London: Cambridge University Press.

Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and cognition* (pp. 195–235). Cambridge, MA: MIT Press.

Gentner, D., & Forbus, K. (2011). Computational models of analogy. *WIREs Cognitive Science*, *2*, 266–276.

Gentner, D., & Gunn, V. (2001). Structural alignment facilitates the noticing of differences. *Memory and Cognition*, *29* (4), 565–577.

Gentner, D., & Kurtz, K. (2006). Relations, objects, and the composition of analogies. *Cognitive Science*, *30*, 609–642.

Gentner, D., & Markman, A. B. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431–467.

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science*, *5* (3), 152–158.

Gentner, D., & Markman, A. B. (1995). Similarity is like analogy: Structural alignment in comparison. In C. Cacciari (Ed.), *Similarity in language, thought and perception* (pp. 111–147). Brussels: Brepols.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*, 45–56.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, *25* (4), 524–575.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91* (1), 1–67.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52* (2), 125–157.

Goldstone, R. L., & Medin, D. L. (1994). Time course of comparison. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20* (1), 29–50.

Golonka, S., & Estes, Z. (2009). Thematic relations affect similarity via commonalities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1454–1464.

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, *87* (1), 1–32.

Hampton, J. A. (1997). Conceptual combination: Conjunction and negation of natural concepts. *Memory and Cognition*, *25*, 888–909.

Harter, J. (Ed.) (2008) *Plants: 2,400 copyright-free illustrations of flowers, trees, fruits and vegetables (Dover Pictorial Archive Series)*. Mineola, NY: Dover Publications.

Hofstadter, D. R., & Mitchell, M. (1994). The Copycat Project: A model of mental fluidity and analogy-making. In K. J. Holyoak & J. A. Barnden, (Eds.), *Advances in connectionist and neural computation theory, vol. 2. Analogical connections* (pp. 31–112). Norwood, NJ: Ablex.

Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 229–264). Hillsdale, NJ: Erlbaum.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13* (3), 295–355.

Hummel, J. E., & Holyoak, K. J. (1997). LISA: A computational model of analogical inference and schema induction. *Psychological Review*, *104*, 427–466.

Jones, M., & Love, B. C. (2007). Beyond common features: The role of roles in determining similarity. *Cognitive Psychology*, *55*, 196–231.

Krawczyk, D., Holyoak, K., & Hummel, J. (2004). Structural constraints and object similarity in analogical mapping and inference. *Thinking and Reasoning*, *10*, 85–104.

Larkey, L. B., & Love, B. C. (2003). CAB: Connectionist analogy builder. *Cognitive Science*, *27* (5), 781–794.

Love, B. C., Rouder, J. N., & Wisniewski, E. J. (1999). A structural account of global and local processing. *Cognitive Psychology*, *38*, 291–316.

Lovett, A., Gentner, D., & Forbus, K. (2006). Simulating time-course phenomena in perceptual similarity via incremental encoding. In R. Sun & N. Miyake (Eds.), *Proceedings of the Twenty-eighth Annual Meeting of the Cognitive Science Society* (pp. 1723–1728). Mahwah, NJ: Erlbaum.

Lovett, A., Gentner, D., Forbus, K., & Sagi, E. (2009). Using analogical mapping to simulate time-course phenomena in perceptual similarity. *Cognitive Systems Research*, *10*, 216–228.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Markman, A. B., & Dietrich, E. (2000). In defense of representation. *Cognitive Psychology*, *40* (2), 138–171.
Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, *32*, 517–535.
Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory & Cognition*, *24* (2), 235–249.
Markman, A. B., & Gentner, D. (2005). Nonintentional similarity processing. In R. Hassin, J. A. Bargh, & J. S. Uleman (Eds.), *The new unconscious* (pp. 107–137). New York: Oxford University Press.
Markman, A. B., & Wisniewski, E. J. (1997). Similar and different: The differentiation of basic level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 54–70.
Medin, D. L., Goldstone, R. L., & Markman, A. B. (1995). Comparison and choice: Relations between similarity processes and decision processes. *Psychonomic Bulletin and Review*, *2*, 1–1.
Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85* (3), 207–238.
Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press.
Navarro, D. J., & Lee, M. D. (2004). Common and distinctive features in stimulus representation: A modified version of the contrast model. *Psychonomic Bulletin & Review*, *11* (6), 961–974.
Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9* (3), 353–383.
Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10* (1), 104–114.
Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14* (3), 510–520.
Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97* (2), 185–200.
Posner, M. I., & Mitchell, R. F. (1967). Chronometric analysis of classification. *Psychological Review*, *74* (5), 392–409.
Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114* (3), 510–532.
Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15* (3), 456–468.
Ross, B. H., Perkins, S. J., & Tenpenny, P. L. (1990). Reminding-based category learning. *Cognitive Psychology*, *22*, 460–492.
Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, *39*, 373–421.
Shoben, E. J. (1983). Applications of multidimensional scaling in cognitive psychology. *Applied Psychological Measurement*, *7*, 473–490.
Sloman, S. (1993). Do simple associations lead to systematic reasoning? *Behavioral and Brain Sciences*, *16*, 471–471.
Smith, E. E., & Medin, D. L. (1984). *Categories and concepts*. Cambridge, MA: Harvard University Press.
Tversky, B. (1969). Pictorial and verbal encoding in a short-term memory task. *Perception & Psychophysics*, *6*, 225–233.
Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.
Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Appendix 1

Simulation of Experiment 1

Experiment 1 used complex heraldic imagery. Because object similarity and relational similarity covaried for this stimulus set, the pattern of human results shows a simple

disassociation: (1) faster ''different'' responses in the S/D task for low-similarity pairs; and (2) faster difference identification for high-similarity pairs.

Simulation

To facilitate the computer simulation of this experiment, we first simplified the stimuli. The simplification involved replacing all patterns with solid colors and replacing objects with geometric shapes (e.g., the dragon in Fig. 1, images C & D, was replaced with a lightning bolt)[16].

We ran SME on the simplified materials using same 40 pairings as were used in Experiment 1: 20 high-similarity and 20 low-similarity pairs of images in the style of heraldic shields. The images were imported directly into CogSketch from PowerPoint. The measures used were taken from SME's normal operation (as in the Experiment 4 simulation). These were:

Local object matches: The number of local matches found by SME between object attributes only. This derived measure represents the object similarity of the stimuli.

Mapping score: The structural evaluation score for global mappings computed by SME between the complete representations. This measure represents the alignability of the stimuli and is a good approximation of relational similarity.

The predictions from structure-mapping theory are (1) For the S/D task, both low local object matches and low mapping score should correlate with human performance, reflecting the claim that any type of low similarity should allow people to quickly determine that the images are different. (2) For the difference-identification task, SME's mapping score should correlate with human performance, since structure-mapping predicts that stimuli that share more structure can be aligned more easily.

Results and discussion

We consider each of the predictions in turn. First, both the local object matches score and the mapping score correlate positively with S/D response times (local object matches: $r = .39$, $p < .05$; mapping score: $r = .55$, $p < .01$). This fits with the pattern that humans require more time to say ''different'' for more similar pairs (both at the object level and at the relational level).

Second, mapping score correlates negatively with response time in the difference-identification ($r = -.43$, $p < .01$), matching the faster performance for high-similarity pairs. However, because similarity and alignability co-varied to a large degree in Experiment 1 ($r = .78$, $p < .01$), the local object matches score also correlates negatively with difference-identification ($r = -.39$, $p < .05$).

Overall, the results of the simulation are as predicted for this simple stimulus set: High values of either the mapping score or the local object matches score are positively correlated with human response times in the S/D task, and negatively correlated with them in the difference identification task. In Experiment 4, we distinguish relational similarity from object similarity so as to test more fine-grained predictions of the model.