Memory & Cognition
1997, 25 (2), 237-255

Similarity in context

ROBERT L. GOLDSTONE
Indiana University, Bloomington, Indiana

DOUGLAS L. MEDIN
Northwestern University, Evanston, Illinois

and

JAMIN HALBERSTADT
Indiana University, Bloomington, Indiana

Similarity comparisons are highly sensitive to judgment context. Three experiments explore context effects that occur within a single comparison rather than across several trials. Experiment 1 shows reliable intransitivities in which a target is judged to be more similar to stimulus A than to stimulus B, more similar to B than to stimulus C, and more similar to C than to A. Experiment 2 explores the locus of Tversky's (1977) diagnosticity effect in which the relative similarity of two alternatives to a target is influenced by a third alternative. Experiment 3 demonstrates a new violation of choice independence which is explained by object dimensions' becoming foregrounded or backgrounded, depending upon the set of displayed objects. The observed violations of common assumptions to many models of similarity and choice can be accommodated in terms of a dynamic property-weighting process based on the variability and diagnosticity of dimensions.

Assessments of similarity seem to be important for a variety of cognitive acts, ranging from problem solving to categorization to memory retrieval. If William James (1890/1950) was correct in stating that "this sense of Sameness is the very keel and backbone of our thinking" (p. 459), we might expect similarity judgments to be stable, reliable, and less flexible than the cognitive acts that depend on them. There is considerable research, however, that undermines this assumption (e.g., Medin, Goldstone, & Gentner, 1993). The three experiments to be reported here provide evidence that assessing the similarity of two things depends on the context of the judgment in several important and specific ways. The contextualized nature of similarity is shown by demonstrating systematic violations of several assumptions associated with standard models of similarity. Furthermore, the results show that "context" is not limited to the items actually present in a choice situation; two compared objects may create, or "recruit," their *own* context, which in turn influences judged similarity.

This research was funded by National Science Foundation Grant SBR-9409232 awarded to the first author, National Science Foundation Grant 95-11757 to the second author, and a Javits Predoctoral Fellowship to the the third author. We wish to thank Jerry Busemeyer, Evan Heit, Arthur Markman, Robert Nosofsky, Richard Shiffrin, and Linda Smith for their useful comments and suggestions. Experiment 1 was based on a suggestion by Jerry Busemeyer. Correspondence concerning this article should be addressed to R. Goldstone, Psychology Department, Indiana University, Bloomington, IN 47405 (e-mail: rgoldsto@indiana.edu). Further information can be found at http://cognitrn.psych.indiana.edu/

Contextual Effects in Similarity

The notion of context dependence stands in contrast to what we will call *fixed-set* approaches to similarity. In this view, similarity is computed by integrating evidence from a fixed set of features or dimensions. Each of the things to be compared is first assigned a representation, in terms of features (Tversky, 1977), or values along continuous dimensions (Carroll & Wish, 1974; Shepard, 1962). The features or dimensions may also be assigned weights that indicate their salience. The representations are then compared in terms of overlap (Tversky, 1977), distance in psychological space (Caroll & Wish, 1974), or transformational distance (Imai, 1977; Wiener-Ehrlich & Bart, 1980). Importantly, the featural or dimensional representations are determined before the comparison process takes place, although in some methods, most notably Tversky's contrast model, the context of a comparison may influence the weights assigned to the features of an object's representation.

The assumption that fixed-object representations are the inputs to a similarity computation is useful in several ways. It is a necessary precondition for many of the similarity techniques to operate. For example, in multidimensional scaling (MDS), objects are represented by points in multidimensional space. Context or intrinsic variability may alter an object's position in the space (Ennis, 1992), and attentional changes may stretch or shrink the space in one or more dimensions (Nosofsky, 1986). However, if one allows that different object properties are systematically considered when the object is compared with different objects, then consistent and appropriate point locations for

the object cannot be determined and a legitimate MDS space cannot be constructed. More generally, if one wishes to explain categorization, problem solving, or memory retrieval in terms of similarity, then a stable notion of similarity is desirable. If similarity itself is as difficult to explain as these higher level cognitive phenomena, as some philosophers have contended (Goodman, 1972), then it cannot provide an informative account or stable ground.

Although the notion of similarity would be simple if the representation and weighting of object properties were context invariant, a substantial amount of evidence argues against this idea (Medin, Goldstone, & Markman, 1995). For example, several researchers have discussed general context effects in judgments as a result of previous trials. According to Parducci's (1965; Wedell, 1994, in press) "range-frequency" theory, judgments represent a compromise between a range and a frequency principle. The range principle states that the range of stimulus values will be divided into equally wide subranges. The frequency principle states that the stimuli will be divided into intervals that contain equal numbers of stimuli. According to the frequency principle, two objects will receive a higher similarity rating when the presentation frequency of highly dissimilar pairs is increased. In fact, Sjöberg (1972) has shown that the similarity of falcon to *chicken*, for example, increased when the entire set of items (mostly birds) to be compared included *wasp* rather than *sparrow*. Similarly, Helson's (1964) adaptation-level theory predicts that earlier trials may create standards by which later trials are compared. In a classic demonstration, a moderately heavy weight was judged to be heavier when preceded by light, rather than heavy, weights (Helson, Michels, & Sturgeon, 1954).

In addition to these general judgment phenomena, other context effects peculiar to similarity judgments have been proposed. Krumhansl (1978) argued that similarity between objects decreases when they are surrounded by many close neighbors—neighbors that were presented on previous trials (also see Wedell, 1994). Tversky (1977) obtained evidence for an *extension effect*, according to which features influence similarity judgments more when they vary within an entire set of stimuli. In one experiment, some participants rated the similarity of pairs of South American countries, others rated the similarity of European countries, and still other participants rated the similarity of two South American countries on some trials and two European countries on other trials. The similarity ratings from this last group were significantly higher than from the other groups, presumably because the features "South American" and "European" became important when they were not held constant across the entire set of stimuli.

Items presented within a particular trial also influence similarity judgments. Perhaps the most famous example of this is Tversky's (1977) diagnosticity effect, discussion of which will be deferred until Experiment 2. More recently, Medin et al. (1993) have argued that different comparison standards are created depending on the items that are present on a particular trial. In one of their experiments, participants rated the similarity of pairs of words that were either presented separately ("separated context") or simultaneously ("combined context"). The pairs of words were either related antonymically (e.g., *white* and *black*) or by association/category (e.g., *skin* and *hair*). Words that were related antonymically received lower similarity ratings than the other words, but only when the words were presented in the separated context. For example, in the separated context, the group of participants who saw the *sunrise-sunset* comparison gave lower similarity ratings than participants who saw the *sunrise-sunbeam* comparison, but this trend was reversed when *sunbeam* and *sunset* were simultaneously compared with *sunrise* for a participant in the combined context group. Medin et al. (1993) argued that the most salient standard of comparison for antonyms in isolation was their dimensional difference (e.g., time of day for *sunrise-sunset*), which is quite large, because they occupy opposite poles. However, when other terms are considered simultaneously, the standard of comparison is enlarged to include the many features shared by antonyms.

In summary, the aggregate of evidence suggests that similarity is not just simply a relation between two objects; rather, it is a relation between two objects and a context. Similarity appears to depend on contexts defined by the set of stimuli used in a particular experimental setting as well as by the context defined by the other alternatives present on a particular trial.

Plan of Experiments

The present experiments examine three distinct types of context effect. The first of these experiments presents evidence against any approach to similarity that assumes that "choose the most similar object" judgments operate on fixed similarities between pairs of objects. The form of this evidence is a violation of transitivity in two-choice similarity judgments. The second experiment explores the cause of the best known context effect in similarity—Tversky's (1977) diagnosticity effect—which is evidenced by violations of choice independence of irrelevant alternatives. The third experiment uncovers additional evidence against choice independence, evidence that is explainable in terms of different contrast sets' being evoked for different comparisons.

Transitivity and choice independence are among the most fundamental measurement assumptions associated with standard models of similarity. Although our experiments find evidence against these assumptions, they will not be interpreted as showing the hopeless vagaries of similarity judgments; instead, the studies are unified in their support of dynamic, on-line judgment processes that determine the importance of particular attributes or dimensions. We argue that stability must be understood in terms of similarity processes, not in terms of outcomes.

EXPERIMENT 1

Assumptions of transitivity are found in most general models of judgment, including models of similarity judg-

ment (Coombs, 1983; Mellers & Biagini, 1994; Shepard, 1962). Because of the intuitiveness and widespread use of the transitivity assumption, Tversky's (1969) demonstration of intransitivities in preference judgments was notable (for similar violations associated with voting aggregation, see Arrow, 1951). Specifically, Tversky found intransitivities in five gamble choices varying on two dimensions, probability of winning and dollar amount won, whose values are shown in Table 1. Participants showed a tendency to prefer A over B, B over C, C over D, D over E, but E over A, giving rise to reliable intransitivities. Payoff probabilities were represented by pie graphs, and consequently subtle discriminations between probability values were difficult to make. Tversky argued that participants ignored probability differences that fell below a certain criterion, but paid attention to these probabilities when given the items A and E that had highly different probabilities. In a second experiment, Tversky also observed asymmetries, arguing that greater weight was given to dimensions that had the largest differences.

The purpose of the present experiment was to look for corresponding intransitivities in forced-choice similarity judgments. The experiment also explored whether intransitivities could be obtained only when there were differential changes in dimension importance due to scaling, the form of explanation provided by Tversky (1969). In the current experiment, a standard object was paired with two comparison objects, and participants were instructed to select the object that was most similar to the standard. For example, the stimuli from Figure 1 were shown to participants on three trials. Participants were asked to choose between A and B, between B and C, or between A and C, selecting the object most like the standard. Systematic intransitivities, if found, would be incompatible with any model of forced-choice similarity judgments that assumed initial determination of individual similarities of A, B, and C to the standard, and then probabilistic selection of the choice with the greatest similarity.

Experiment 1 was designed to detect either of two varieties of intransitivity. The first variety of intransitivity would be for A to be systematically selected over B as more similar to T, B to be selected over C, and C to be selected over A. This pattern will be called the "dimension counting strategy," because it amounts to counting the number of dimensions on which an object is more similar to the standard than are its alternatives. This strategy is an example of the "majority of confirming dimensions" choice rule in judgment. By this strategy, the object that is most similar to standard on most dimensions will tend

Table 1
Gambles Employed in Experiment 1 of Tversky (1969)

Gamble	Probability of Winning	Payoff (in $)
A	7/24	5.00
B	8/24	4.75
C	9/24	4.50
D	10/24	4.25
E	11/24	4.00

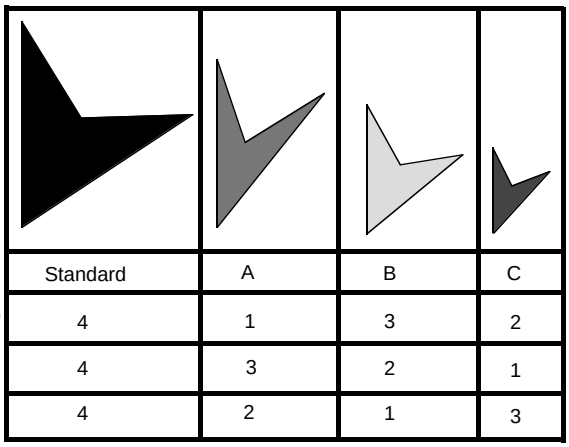


Figure 1. Sample stimuli from Experiment 1. Hue differences are represented by shading differences. The majority of obtained intransitivities were of the following pattern: B was more similar to the standard than was A, C was more similar to the standard than was B, and A was more similar to the standard than was C.

to be selected. In Figure 1, A is more similar than B to the standard on size and color, B is more similar than C on angle and size, and C is more similar than A on angle and color. A second variety of intransitivity arises if B is selected over A, C is selected over B, and A is selected over C. This "diagnostic-dimension strategy" emphasizes the dimension that serves to distinguish the two choices most clearly (a similar rule was proposed by Mellers & Biagini, 1994). By this rule, B should be chosen over A because B is much more similar to the standard than is A on the angle dimension, C would be chosen over B because of C's substantial superiority on the color dimension, and A would be chosen over C due to A's substantial size advantage.

The experiment examined whether the probability of making a particular choice varied significantly in accord with either the diagnostic-dimension or the dimension-counting rules. The data could also reveal no systematic violation of transitivity. Patterns of random responses, or patterns of responses that indicate a consistent weighting of dimensions will not be taken as evidence of intransitivity. For example, if A > B > C with respect to similarity to the standard, then we can simply hypothesize that size is the most important dimension in the experiment.

Method

Subjects. Fifty-eight undergraduate students from Indiana University served as subjects in order to fulfill a course requirement.

Materials. The basic objects to be compared were wedge-like shapes, as shown in Figure 1, displayed on color Macintosh IIsi computers. The objects varied on three dimensions: angle, size, and hue. Four values were possible on each dimension. The angle dimension refers to the angle of the right arm of the wedge and varied from 16º to 30º relative to the left edge of the left arm of the wedge, which always pointed directly upward (0º). The size dimension refers to the total vertical length of the objects, and that varied from 2.6 to 6.3 cm. The color dimension refers to the amount of red hue in the objects. The 1976 CIE (Commission Internationale de L'Éclairage)

values for the hues varied from u′ = .5463 and v′ = .3696 to u′ = .5266 and v′ = .3764. All of the objects had luminances of 27.6 cd/m^2 as measured by a Spectrascan 714 chromometer.

Design. A standard was compared with three comparison objects. Sets of comparison objects were created such that each object was slightly closer to the standard than was another object on two dimensions and further from the standard than was the other object on the remaining dimension. Figure 1 shows one example of this structure. In this example, the standard has the largest possible value on all three dimensions (widest angle, largest size, deepest red). Object A is closer to the standard than is object B on two of the three dimensions; on the third dimension (angle), object B is two values closer to the standard than is A. Similarly, object B is closer to the standard than is object C on two dimensions, and object C is closer to the standard than is object A on two dimensions. In this manner, each of the three comparison objects was always more similar to the standard than was another comparison object on two dimensions, and was less similar to the standard than was the third comparison object on two dimensions.

Eight replications of this structure were tested; in each, the standard was defined by different combinations of extreme values on the three dimensions: 444 (highest values on all three dimensions, as shown in Figure 1), 441, 414, 411, 144, 141, 114, and 111 (lowest values on all three dimensions). Each comparison object was separated from the standard by one value on one dimension, by two values on another dimension, and by the three values on the third dimension. The separation intervals are arranged in the equivalent of a Latin-square design in the three objects and in the three dimensions.

Customization of materials. Opportunities for discovering intransitivities are available only if the three dimensions have approximately equal saliences. For example, if one participant found color to be the most salient dimension in Figure 1, then he or she would choose A over B, B over C, and A over C, without violating transitivity. Consequently, over the course of the experiment, we equated the salience of the dimensions for each participant. In addition, it was necessary to customize the interval between values on a dimension, because participants might disagree with respect to the subjective difference between two values on a dimension.

In order to achieve this participant-specific customization of the materials, the function that translated the abstract dimensional characterization of an object into its physical instantiation was altered depending on a participant's previous choice. If a participant selected the object that was more similar on only one dimension (dimension X) then the difference between the standard value on dimension X and the choice's value on dimension X was increased by two units by changing the choice's value, and the difference between the standard value and the alternative not selected was decreased by one unit on each of the other two dimensions. If a participant selected the object that was more similar on two dimensions, then the difference between the standard and the choice's value on both of these dimensions was increased by one unit, and the difference between the standard and the alternative not selected was decreased by two units on the remaining dimension. These alterations preserve the overall dissimilarity of the three choices from the standard in terms of number of units. Initially, abstract values of 1, 2, 3, and 4 on a dimension were given unit values of 0, 20, 40, and 60, respectively. For the color dimension, one unit roughly corresponded to a .007 change in the u′ and v′ CIE hue coordinates of an object. For the size dimension, one unit corresponded to 0.11 cm. For the angle dimension, one unit corresponded to 0.8º.

Procedure. The standard was always placed at the top of the screen, and the two comparison objects were placed to the lower left and lower right of the standard. Subjects were instructed that they would see three wedge-like objects on the screen and would be asked to decide which of the two lower objects was more similar to the top object. The subjects pressed the "1" or "2" key on the computer if they judged the left or right object, respectively, to be most similar to the standard. The objects remained on the screen until the judgment was made.

Each subject completed 288 trials in all. The first 144 trials were used to customize dimension values for the subjects; the results were analyzed for only the last 144 trials. These consisted of six replications of the 24 basic comparisons. In turn, each set of 24 comparisons consisted of 8 replications of 3 trials; each of the 8 replications had a different standard defined in the manner described above. The three repeated trials displayed standard-A-B, standard-B-C, and standard-A-C. The order of the trials and the left/right positions of the two alternatives were randomized.

Results

Similarity choices can be coded in terms of whether they are predicted by the dimension-counting or diagnostic-dimension rules. Evidence for systematic intransitivities exists if either rule receives significantly greater support than the other rule. In fact, 61% of the 2,815 responses were consistent with the diagnostic-dimension rule, and 39% were consistent with the dimension-counting rule [$t(57) = 5.8$, $p < .01$], where the null hypothesis is that the true proportion of diagnostic-dimension responses is 50%. Given that objects A, B, and C were chosen equally often overall [51%, 49%, and 50%, respectively; $F(1,57) = 2.2$, $MS_e = 0.21$, $p > .3$], this preponderance of diagnostic-dimension choices cannot be explained by unequal similarities between the choices and the standard. Rather, the difference seems to be due to the context (the other choice) in which the objects are presented.

The standard method for finding violations of transitivity is to test the assumption of weak stochastic transitivity (WST), which states that if $P(A,B) \geq 0.5$ and $P(B,C) \geq 0.5$, then $P(A,C) \geq 0.5$, where $P(A,B)$ is the probability of object A being chosen over object B. WST is the easiest form of transitivity to satisfy, and hence, finding violations of this form of intransitivity is the most difficult and significant. In testing WST, the A-B, B-C, and A-C pairs for each participant were separately tabulated. Treating each of three response proportions as binary valued (e.g., either A is selected more often than B or it is not), there are eight (2^3) different outcomes for the three choice pairs, shown in the eight rows of Table 2. The top and bottom rows reflect intransitivities, and the remaining six rows reflect transitivities. The top row reflects the type of intransitivity that would occur if a participant were adopting a diagnostic-dimension rule for selecting most similar choices. Overall, the percentage of participants producing response profiles that fall into one of the two intransitive patterns is greater than would be expected by a chance rate of 25% (binomial $Z = 5.63$, $p < .01$). In addition, more participants produce diagnostic-dimension intransitivities than produce dimension-counting intransitivities ($Z = 6.36$, $p < .01$).

One might argue that the large number of diagnostic-dimension intransitivities can be explained by a transitive choice model with response noise added. For example, if a participant's $P(A,B) = 0.9$, $P(B,C) = 0.8$, and $P(C,A) = 0.45$, then the participant would produce data in accord with WST.[1] Random noise would be more likely

Table 2
Frequencies of Eight Choice Outcomes Among 58 Subjects

A > B	B > A	C > A	Frequency
Yes	Yes	Yes	25
Yes	Yes	No	5
Yes	No	Yes	5
Yes	No	No	4
No	Yes	Yes	4
No	Yes	No	3
No	No	Yes	4
No	No	No	8

Note—A "Yes" in the column "A > B" indicates that object A was chosen over object B greater than 50% of the time. The top row indicates the type of intransitivity that would occur by following a "diagnostic dimension" rule. The bottom row reflects a "dimension-counting" intransitivity profile. The remaining six rows indicate transitive profiles.

to flip P(C,A) than P(A,B) or P(B,C) to less than 0.5, and thus intransitivities could be produced by added random noise. However, this possibility can be ruled out because the number of participants producing diagnostic-dimension intransitivities is greater than the number of participants producing any response profile that differs from this intransitivity by only one choice (e.g., rows 2, 3, and 5 in Table 2). In fact, the number of participants producing these intransitivities significantly exceeds the *sum* of the most similar response profiles ($Z = 3.38$, $p < .01$).

In addition to conducting frequency counts of participants, further analyses explored whether significant departures from WST were found for individual participants. Following Tversky (1969), likelihood ratio tests of WST and the two varieties of intransitivity were conducted for each participant. The quantity

$$Q(M_1, M_0) = -2 \ln \frac{L^*(M_1)}{L^*(M_0)}$$

provides a measure of the increase in probability that an unrestricted model has over a more restricted model. $L^*(M_0)$ is the maximum value of the likelihood function of the observed data under the unrestricted model, and is given by the product of the binomial probabilities. $L^*(M_1)$ is the same function under the restricted model, and is obtained by substituting a value of 0.5 in the binomial product for those choice probabilities that were incompatible with the restricted model. Violations of the restricted model are shown when $Q(M_1, M_2)$ reaches significance at $p < .05$ under a chi-square distribution with degrees of freedom equal to the number of restricted values.

The restricted model that tests WST is: whenever the three choice probabilities exhibit an intransitivity, the value closest to 0.5 is set to 0.5. With this model as M_1, the null hypothesis that the unrestricted and restricted models fit the data equally well could be rejected for 22 of the 58 subjects. As such, restricting a model so as to predict transitivities resulted in a significantly worse fit to the observed choice probabilities for roughly one third of the subjects. For these subjects, 16 showed intransitivities according to the diagnostic-dimension rule, and the remaining 6 showed dimension-counting intransitivities.

Discussion

Participants displayed a greater number of diagnostic-dimension intransitivities than would be predicted if their similarity choices were not influenced by the alternatives simultaneously presented. If the similarity of each choice to the standard were invariant, then participants either would prefer one alternative over the others or (if all three choices were approximately equally similar) would randomly choose between alternatives. The results indicated that each alternative within a triad was selected approximately 33% of the time by a participant. Still, participants systematically selected some alternatives over other alternatives more than 50% of the time when given two alternatives from which to choose.

The results indicate that similarity assessments are influenced by the context within a single trial. It should be noted, however, that the customization procedure, in which the saliences of the dimensions were roughly equated for each participant, made it likely that even subtle intransitivities would be uncovered. Consequently, it is difficult to estimate the magnitude of these context effects. Pilot testing showed no violations of intransitivities when the customization procedure was not used.

The majority of participants demonstrated diagnostic-dimension intransitivities, in which they tended to select the alternative with one clearly more similar dimension than the other alternative. There are similarities between this result and the intransitivities of preference reported by Tversky (1969). In Tversky's Experiment 1 (see Table 1), when alternatives (gambles) differed by only a small amount, participants based their judgments on payoffs rather than probabilities. When they differed by larger amounts, participants based their judgments on probabilities instead. Tversky argued that small probability differences were treated as inconsequential. Similarly, in the current experiment, participants may have ignored a dimension if the alternatives did not differ greatly on it. The difference between the current results and Tversky's findings is that our results do not seem to depend on particular salience characteristics of dimensions. Tversky explained his effects by positing that particular dimensions (probabilities in his Experiment 1, intelligence in Experiment 2) tend to be particularly ignored if there are small differences along them, and heavily weighted if differences along them are large. To the extent that our participants are following a diagnostic-dimension rule, intransitivities can be explained without positing that different dimensions' saliences are differentially affected by scaling.

The predominant diagnostic-dimension strategy is also consistent with other recent research on choices and similarity. Goldstone, Medin, and Gentner (1991) found that adding a feature match along one abstract dimension increased similarity more when there were other feature matches along the same abstract dimension. Mellers, Change, Birnbaum, and Ordonez (1992) described a similar effect that they called "contrast weighting," in which attributes with similar levels between alternatives re-

ceived less weight than did attributes with dissimilar levels. Mellers and Biagini (1994) also found evidence that similarity along one dimension enhanced differences on another dimension. All three of these effects are consistent with diagnostic-dimension intransitivities in that each predicts that differences between the alternatives along the most discrepant dimension will be particularly heavily weighted during choice selection. Apparently, in addition to influencing their preferences, participants' tendency to selectively weight diagnostic dimensions also influences their forced-choice similarity judgments, even with materials comprising logically equivalent stimuli (unlike the stimuli used by Tversky, 1969).

EXPERIMENT 2A

Experiment 1 demonstrated that dimensions that were diagnostic for distinguishing between choices were used as the basis for selecting choices as being more similar to a standard. The best known example of a within-trial context effect is also based on the diagnosticity of the choice attributes. Tversky (1977) argued that features that were diagnostic for relevant classifications would have disproportionate influence on similarity judgments. In one experiment, participants were asked to choose one out of three countries as being most similar to a fourth country (the standard). Participants tended to choose Sweden over Hungary as being more similar to Austria when the third alternative was Poland. However, participants tended to choose Hungary over Sweden when the third alternative was Norway. Tversky argued that the third alternative influenced similarity judgments by altering the categories that were likely to be created. Participants who were given the countries Austria, Sweden, Hungary, and Poland to sort were more likely to group Austria and Sweden together than were participants who were given Austria, Sweden, Hungary, and Norway.

By changing one of the choices, Tversky altered the featural commonalities between the alternatives. However, Tversky did not distinguish between two types of featural commonalities that might behave quite differently. Two choices may share a common feature that is also possessed by the standard. This situation will be called "shared match," where "match" refers to the match between the alternatives and the standard. Alternatively, the two choices may share a common feature that is *not* possessed by the standard. This situation will be called "shared mismatch." Figure 2 shows how making alterations to one item (B) can affect whether features of items A and C share a match or mismatch with other alternatives. In the first row of Figure 2, face A shares a match with face B: faces A and B have the same eyes, which are also shared by the standard. In the third row, face A shares a mismatch with face B because they share a common smile, and that smile is not shared by the standard.

In Experiment 2, participants were given a forced-choice similarity judgments among three alternatives, such as those in Figure 2. Different participants received the same sets except that the appearance of one item (B in Figure 2) was manipulated to alter the shared feature matches and mismatches. For all triads, B was dominated in the sense that either A or C had greater similarity to the standard than did B (for the shared-match trials, both A and C dominated B). B's role was to alter the uniqueness of A's and C's features within a triad.

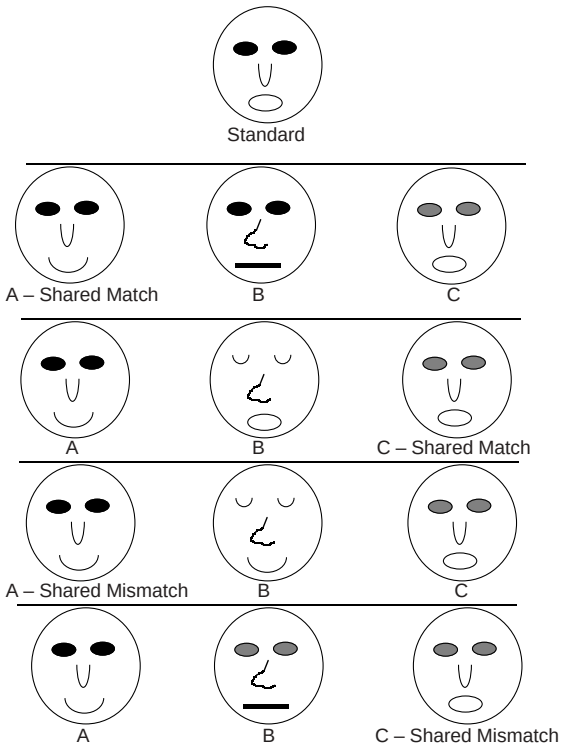


Figure 2. Sample stimuli from Experiment 2. The four triads of choices are each compared to the same standard. A choice shares a feature match with another alternative if the two alternatives share a feature that is also possessed by the standard. Two choices share a mismatching feature if they share a feature that is not possessed by the standard. Within a given set, objects A and C do not change at all.

Disentangling the effects of shared matches and mismatches allows one to test several different theories of context-based similarity that make different predictions for the two situations. First, consider Tversky's suggestion that categorization determines feature diagnosticity, which in turn determines choice selection. This theory makes the clear prediction that items that share mismatches should be chosen less frequently than items that do not share mismatches. As an example, consider the third row of Figure 2. If these three items and the standard were given to participants to sort into two categories, more participants would place A and B in one category and C and the standard in the other category than would place A and the standard in one category and B and C in the other category. This was confirmed in a pilot experiment ($n = 10$) testing several of the sets from Experiment 2 and using the same sorting task used by Tversky (1977), except that we did not force participants to cre-

ate equal-sized categories. The observed sorting was preferred by all 10 subjects, presumably because it allowed for equal-sized groups that could be determined by examining only one feature (smiles in one group, open mouths in the other group). Category-based diagnosticity predicts that once C and the standard are placed in the same category, attention to the mouth dimension increases because it distinguishes between the two categories, and therefore participants should be more likely to select C, the choice that has its mouth in common with the standard. This is also the choice that does not share any mismatches with either of the other alternatives.

The predictions of category-based diagnosticity for the shared-match situation are less clear, because participants' sorting performance is not as clear. There appears to be some tendency (12 of 18 pilot participants) to group items that share a matching feature with the standard. For example, face A in the first row of Figure 2 is more likely to be sorted with the standard (and face B) than is face A in the second row. Consequently, this predicts a slight tendency for participants to choose an item as most similar more often when it shares a matching feature. Thus, as Table 3 summarizes, Tversky's categorization-based diagnosticity principle predicts that, all else being equal, an item that shares with another item a mismatching feature with the standard, should be selected relatively rarely; an item that shares a matching feature should be selected relatively frequently.

Another intuitive strategy that participants might use is to select the "odd value out," that is, to base their choices on the dimension that distinguishes one alternative from the other two. This strategy predicts that an alternative that shares either matches *or* mismatches with another item will be selected less frequently as being most similar to the standard. For example, in the top row of Figure 2, because A and B have the same eyes, C's eyes stand out as being distinct, and would be selected as the basis for choice. In the third row of Figure 2, A and B have the same mouth, and thus there would be a tendency to select C because its mouth matches the standard. This strategy is plausible given that distinctiveness may capture attention.

Finally, a third possible strategy, "variability-based diagnosticity," is to determine similarity by weighting dimensions according to how many different values they take within their sets. Dimensions that have many different values would get greater weight in determining the similarity between alternatives and the standard. The motivation for weighting variable dimensions is similar to the motivation behind diagnostic-dimension strategy in Experiment 1. Both are based on the premise that dimensions that distinguish between choices well are particularly valuable. Dimensions that take many different values carry more information than dimensions that take few values, and consequently better allow alternatives to be distinguished from one another. On average, dimensions that have many different values will also have a greater degree of difference between dimension values than will dimensions that have fewer different values.

The variability-based diagnosticity strategy predicts that, all else being equal, alternatives that share matches will be selected infrequently and alternatives that share mismatches will be selected frequently. For example, in the top row of Figure 2, there are three types of mouth and two types of eyes. The more variable mouth dimension will receive greater weight in a similarity assessment than will the eye dimension. Accordingly, C will tend to be selected over A, because C and the standard have a common mouth feature. By the same token, an item with shared mismatches will be chosen relatively frequently because the dimension that is shared will have relatively few values. In the third row of Figure 2, there are three different eye types and two mouth types. Now, the more variable eye dimension will receive greater weight, and A (the choice with a shared mismatch) will tend to be selected over C.

Table 3 contains a summary of the predictions made by these three strategies of context-dependent choice selection. In addition to these strategies, it is also possible that shared matches and mismatches both increase selection probability or that no context dependence is found (e.g., B may be so different that it no longer is effectively a part of the choice context).

Method

Subjects. Eighty-eight undergraduate students from Indiana University served as subjects in order to fulfill a course requirement.

Materials. Figure 2 shows an example of one full set of four triads. There were five such sets involving faces, two separated geometric shapes, single geometric shapes defined by their shape and color, two connected circles with different colors, and contiguous "blobs" with three protrusions. The items within a set varied on two or three dimensions. The sets were created under the following constraints: the standard and items A and C were not changed from trial to trial; A and C were approximately equally similar to the standard, such that each had an identical feature in common with the standard that the other item did not; item B was less similar to the standard than A or C, and four variations of B shared matching or mismatching features with A and C, as in Figure 2.

All of the items were approximately 4 cm^2 in area. When an item consisted of two detached shapes, the shapes were separated by no

Table 3
Predictions and Outcomes for Experiment 2

	Selection Probability	
Theory	Shared Match	Shared Mismatch
Categorization-based diagnosticity	Increases	Decreases
Select "odd value out"	Decreases	Decreases
Variability-based diagnosticity	Decreases	Increases
Actual results	Decreases	Increases

more than 0.6 cm. The objects were shown on a color Macintosh IIsi screen.

Procedure. On each trial, participants were shown a display that consisted of the standard item at the top of the screen and the three choice items below it. Participants were instructed to select the item that was most similar to the standard. They made their selections by using a mouse to move a cursor until the cursor was on top of an item and then pressing the button on top of the mouse.

Each set of items produced four trials of the three alternative forced-choice task, as illustrated in Figure 2. There were five sets of items and eight repetitions of every trial, yielding 160 trials in all for every subject. Trial order was randomized. In addition, the positions of items A,B, and C were randomized on every trial and were always equally spaced and horizontally level.

Results

Although there are four types of trials within a set of items, only two types are logically distinct—trials that involve shared matching features and trials that involve shared mismatching features. There is no logical difference between items A and C. Consequently, the two trials involving shared matches (likewise for mismatches) were collapsed together.

On trials with a shared match, the choice that shared a matching feature with B was selected 48.5% of the time, the choice that did not share a matching feature with B was selected 50.8% of the time, and B was selected 0.7% of the time. Although the effect size is not large, a preplanned paired t test indicated that the alternative that shared matches with B was selected less often than the other alternative [$t(87) = 2.6$, $p = .01$]. Of the five sets, three produced this pattern unambiguously, and the other two produced ambiguous results in which one item was selected more often when it shared matches; the other item in the set was selected less often when it shared matches. Thus, 8 of the 10 items supported the trend that was significant in the subject analysis.

On shared mismatch trials, the choice that shared a mismatching feature with B was selected on 50.2% of the trials, the choice that did not share a mismatch with B was selected on 49.0% of the trials, and B was selected on 0.8% of the trials. The difference between the shared mismatch choice and the choice without a shared mismatch was marginally significant for the subject analysis [$t(87) = 1.8$, $p = .08$]. In the item analysis, 6 of the 10 items showed this trend.

To assess whether these results were an artifact of combining across qualitatively different response patterns, two measures were calculated for each participant that reflected the degree and direction of the context effect: (1) percent choices of object with shared matching feature − percent choices of object without shared matching feature, and (2) percent choices with shared mismatching feature − percent choices without shared mismatching feature. There was a modest, significant correlation between these measures across subjects [$r(87) = -0.15$, $p < .05$], but this simply reflects the general effect that some participants were more influenced by the irrelevant choice, item B, than others. For each of the two derived measures, the null hypothesis that the measures were distributed normally could not be rejected by a goodness-of-fit test [$\chi^2(7) < 5.4$, $p > 0.3$]. As such, there was nothing in the results to suggest that the average context effects represented a blend across a multimodal distribution of different classes of participants.

In sum, the results indicate that a shared feature match decreases selection probability, while there is a nonsignificant trend for a shared mismatch to increase selection probability. When the magnitudes of these two effects are compared for each participant (average effect sizes due shared matching and mismatching features were 2.3% and 1.2% of choices, respectively), there is a significantly stronger influence of shared matches than of shared mismatches [$t(87) = 2.2$, $p < .05$].

Before discussing these findings, we report the results of a replication, Experiment 2B, which was conducted because of the marginally significant effect of shared mismatching features. In the replication, an attempt was made to increase the contextual influence of the middle choices in Figure 2, by having these items appear on the screen before the other two alternatives. In this manner, it was hoped that attention would initially be focused on these items, and that changes to them would consequently have a greater impact on people's choices.

EXPERIMENT 2B

Method

Subjects. Sixty-two undergraduate students from Indiana University served as subjects in order to fulfill a course requirement.

Procedure. The materials were the same as those used in Experiment 2A. The procedure was also identical except for the timing of the displayed items. At the beginning of a trial, the standard and choice B were presented. After 500 msec, choices A and B also appeared on the screen. As before, the spatial positions of the choices A, B, and C were randomized. Thus, choice B was not always presented in the center position. The subjects were not allowed to select an object as most similar to the standard until all three choices were present.

Results

On trials with a shared match, the choice that shared a feature match with B was selected 48.5% of the time, the choice that did not share a feature match with B was selected 51.0% of the time, and B was selected 0.5% of the time. A paired t test indicated that the alternative that shared matches with B was selected less often than the other alternative [$t(61) = 3.2$, $p < .01$]. For 9 of the 10 items that tested the influence of a shared match, the shared match decreased the likelihood of choosing the item.

On shared mismatch trials, the choice that shared a feature mismatch with B was selected on 50.2% of the trials, the choice that did not share a mismatch with B was selected on 49.2% of the trials, and B was selected on 0.6% of the trials. The difference between the shared mismatch choice and the choice without a shared mismatch was again only marginally significant for the subject analysis [$t(61) = 1.863$, $p = .10$]. In the item analysis, 7 of 10 items showed a trend for shared mismatches to increase choice probability. When the results from Experi-

ments 2A and 2B are combined, a significant effect does emerge for shared mismatches to increase choice probability, by an analysis of variance (ANOVA) with experiment (2A vs. 2B) as a between-subject variable [$F(1,146) = 4.7$, $MS_e = 0.07$, $p < .05$], but there was no interaction with experiment; the effect of shared mismatches was not affected by the timing of the displayed items.

Discussion of Experiments 2A and 2B

The results of Experiments 2A and 2B are most consistent with a variability-based diagnosticity strategy. According to this strategy, the influence of a dimension increases as the variability (or number of values) along this dimension among the choices increases. Although the effect sizes were smaller than the effect sizes reported by Tversky (1977), they indicated that items that shared matching features with other alternatives were less likely to be selected than items that did not share matching features. Although the results were more ambiguous for shared mismatch trials, if anything, shared feature mismatches seem to increase selection probability. In Experiment 2A, the influence of shared matches was significantly greater than the influence of shared mismatches.

It should be noted that another strategy that predicts the pattern of results is "rarity-based diagnosticity," in which similarity is determined by weighting features by how rare they are within their sets. Shared rare features might receive more weight than typical features because they are highly informative. For example, the similarity of zebras and tigers is increased substantially by their shared possession of the feature "striped," which is a fairly rare feature among animals. In the same manner, a distinctive but typical feature might decrease similarity to a lesser extent than a distinctive rare feature. The similarity of pigeons to cows is not decreased much by the pigeons' possession of the feature "only one stomach." Again, the rarer feature has more influence on (decreasing) similarity because it is more informative. Consistent with our results, rarity-based diagnosticity predicts that alternatives that share matching features will be chosen relatively infrequently, and that alternatives that share mismatching features will be chosen relatively frequently.

Neither the variability-based nor the rarity-based principle, by itself, predicts the asymmetry in Experiment 2A between the influences of shared matching and mismatching features. One possible account of the asymmetry is that, because of an attenuating effect of an "odd value out" strategy, shared mismatches increase selection probability less than shared matches decrease selection probability. Overall, however, the variability-based diagnosticity hypothesis can account for the results more effectively than the "odd value out" strategy, because the latter does not predict an asymmetry (unless supplemented by the variability-based diagnosticity strategy) and also predicts the opposite influence of shared mismatching features than was obtained.

Although we have distinguished Tversky's original category-based diagnosticity from the other theories tested in the current experiment, the strategies do have much in common. All of the strategies assume that the weight associated with items' features depends on the context defined by the other alternatives. Although the contextual effects were not strong, they confirmed Tversky's basic finding that within-trial context can alter selections. Furthermore, all of the strategies assume that the contextual effects are based on features' or items' being emphasized because of their informativeness. The primary difference is that Tversky's diagnosticity hypothesis is based on the informativeness of a feature for determining categorizations that involve all of the items that are displayed on a given trial. The principles that received support here are based on the informativeness of a dimension (variability-based diagnosticity) or dimension value (rarity-based diagnosticity).

These results, combined with those from Experiment 2A, indicate significant context effects, but on a smaller scale than the 40% differences reported by Tversky (1977). Several possible explanations for this difference were pursued. First, one factor leading to the attenuated influence of context may be the use of a within-subject design with many repeated trials, encouraging participants to adopt consistent, context-independent, choice preferences. In a test of this hypothesis, the two measures of degree of contextual influence were regressed on blocks but were not found to be significant [$F(1,61) < 1.3$, $MS_e = 1.4$, $p < .05$]. Accordingly, there is no evidence that later trials were less sensitive to contextual manipulations than were early trials. Second, we may have obtained smaller context effects than Tversky (1977) because, in several of his examples, a high percentage of choices were for B items, and he did not control for these. Third, with the small context effects obtained here, the effect size is roughly the same as (or larger than) those obtained in other studies of diagnosticity effects in similarity (James Hampton, personal communication, March 1991). Fourth, the influence of diagnosticity may be in competition with other factors that drive context effects (see, for example, the attraction effects discussed in Medin et al., 1995).

Ultimately, the comparison between these results and Tversky's (1977) data may be difficult to make, given the variety of materials that Tversky used to test his principle. In fact, some of the materials (e.g., Tversky's Figure 4) do unambiguously involve shared mismatches, but these materials are combined with materials that seem to test shared matches. Tversky's original formulation of the diagnosticity hypothesis makes different predictions in these two cases, even though the net result for both shared matches and mismatches is to increase the similarity between two alternatives. One of the central purposes of Experiments 2A and 2B was to disentangle the effects of shared features that increase versus decrease similarity to a standard. We do find discrepancies between subjects' sorting of objects into categories and their choice judgments that would not be expected by Tversky's category-based diagnosticity premise. The importance of dimensions does seem to be dynamically based

on their diagnosticity, but it appears not to be based on their diagnosticity for grouping objects together, but rather on their diagnosticity for distinguishing among alternatives.

In sum, the results of Experiments 2A and 2B are consistent: There seems to be a strong influence of shared matching features in decreasing choice probability and a weaker influence of shared mismatching features in increasing choice probability. These results are consistent with either of two strategies—placing emphasis on dimensions that show large amounts of variability within a stimulus set or placing emphasis on dimension values that are rare. Both of these strategies, particularly the first, are consistent with the intransitivities from Experiment 1; dimensions or dimension values seem to be dynamically weighted as a function of their informativeness within the set of items presented on a single trial.

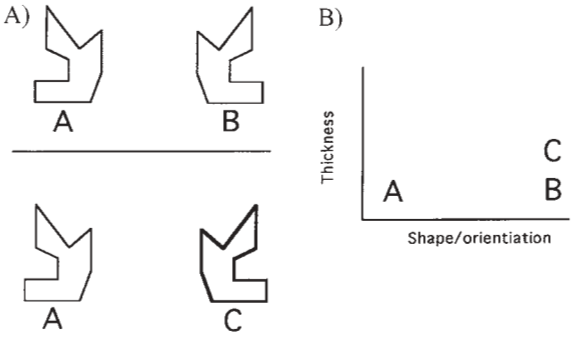


Figure 3. (A) Sample stimuli from Experiment 3. Nonmonotonicities are obtained if objects A and C are judged to be more similar than objects A and B. The only difference between objects B and C is on line thickness, and A and B have the same value on this dimension. (B) The abstract, multidimensional representation of this, and every other, set is shown.

EXPERIMENT 3A

The first two experiments demonstrate that the context in which a similarity judgment is made can produce violations of transitivity and of choice independence. In both cases, the context was defined by the items presented on a particular trial. Experiment 3 explored the possibility that even when the context of the comparison remains constant, comparisons themselves can evoke their own context in the form of a *contrast set* (Kahneman & Miller, 1986; Lehrer & Kittay, 1992). Even in comparisons that involve only two items, different contrast sets may emerge, depending on which dimensions are foregrounded or highlighted by the comparison.

Garner (1962) argued that dimensions are foregrounded when there is variation along them (see also Bransford, Franks, Vye, & Sherwood, 1989; Pomerantz & Lockhead, 1991). On seeing a circle drawn in black ink in the center of a card, one often does not imagine that it might have been colored differently, moved to a corner, printed with thicker lines, or drawn three-dimensionally, unless these variations are explicitly mentioned. People naturally created an "inferred set" of possible alternatives when describing a stimulus. Similarly, Kahneman and Miller's (1986) norm theory assumes that a presented situation evokes a set of alternatives that are used in subsequent evaluations of the original situation. Evoked alternatives, inferred sets, and contrast sets all have in common the notion that objects or situations spontaneously call to mind other related objects.

Adding unique features may change the salience of a previously backgrounded dimension (i.e., a dimension ignored because of lack of variation) of the comparison pair, and therefore also the contrast set evoked by the alternatives. This spontaneous evocation of contrast sets can potentially lead to situations where adding a unique feature to one of a pair of objects does not decrease their similarity. Consider the top pair of shapes labeled A and B in Figure 3. The contrast set for this comparison would likely be other shapes, perhaps with similar regularity and angularity. The dimension on which the objects differ (orientation) is salient, while the many ways they are similar (e.g., thickness of lines, color, size on the page, texture, etc.) are backgrounded: they are not considered relevant for the judgment (Garner, 1962). In the second pair of curves, A and C, a unique feature, line thickness, has been added to C, and the similarity relative to the first pair should therefore decrease in that a featural similarity has been removed. However, the contrast set for the comparison may also have changed, expanding to include shapes of different thicknesses. In the context of the thick line, it becomes apparent that the shapes could vary on a second dimension—line thickness. Yet, on this new dimension, the two shapes are relatively close; their lines vary only slightly within the range of possible line thicknesses defined by the contrast set. The first pair lies at the extremes of its contrast set because the shapes have opposite orientations, but the second pair is relatively similar in the expanded set. Thus, although a unique feature has been added in the second comparison, it is possible to predict that A and C will not receive a lower similarity rating than will A and B. In fact, A and C may receive a higher similarity rating, in violation of an assumption of monotonicity, according to which adding a common feature to two items should never decrease their similarity, and adding a unique feature to one of the items but not the other should never increase their similarity (Tversky, 1977).

An abstract characterization of the experimental logic is shown on the right side of Figure 3. Items A and B differ considerably on the horizontal dimension; items B and C have the same value on this horizontal dimension. Items A and B have identical values on the second, vertical dimension. Items A and C have slightly different values on this dimension. If this vertical dimension is a normally backgrounded dimension such as line thickness, then participants who are given only A and B to compare may not even consider their similarity on this dimension when evaluating their similarity. However, when given the comparison between A and C, a second group of participants may consider the vertical dimension because of variation along it between the compared items, and increase their similarity estimates accordingly (reasoning

that "A and C may be far on the horizontal dimension, but they are quite close on the vertical dimension, so I will give them an intermediate similarity rating").[2]

In Experiment 3, we employed the design shown in Figure 3 to construct sets of three stimuli that we predicted would produce potential nonmonotonicities. Participants judged the similarity of pairs of these triples in one of two conditions. In one condition (three-way comparison), objects A, B, and C were simultaneously displayed. In the second condition (two-way comparison), A and B were displayed on some trials and A and C were displayed on others (Medin et al., 1993).

For the three-way comparison, the same contrast set that is used for the A-B comparison is likely to be used for the A-C comparison because the same three items are present during both comparisons. If the same contrast set is used for both comparisons, then it is expected that A-B pairs will receive higher similarity ratings than A-C pairs, reflecting the physically greater similarity of the A-B pairs. For the two-way comparison condition, however, it was predicted that the similarity of the A-C pairs would be not be greater (and may be less) than their corresponding A-B pairs, because different contrast sets would be invoked in the two comparisons. As argued above, the addition of a unique feature in C is predicted to increase the size of the contrast set for the comparison and to foreground a dimension backgrounded in the A-B comparison.

Even if context-dependent foregrounding of typically backgrounded dimensions occurs, it may be difficult to find materials that reliably produce nonmonotonicities—situations in which the A-C pair actually receives higher similarity ratings than A-B pairs. However, obtaining equivalent ratings for the two pairs in the two-way comparison condition will provide only weak evidence for contextual influences, because the two-way condition may not offer sufficient sensitivity to detect the small physical difference between B and C. It is reasonable to expect more differential responding to A-B and A-C pairs in the three-way context because they are simultaneously present, and it is common for relational judgments to be more sensitive than absolute ones (Miller, 1956). To make certain that any observed differences between A-B and A-C judgments in the two different comparison conditions are not simply due to sensitivity effects, an additional set of control stimuli were designed that did not involve backgrounded dimensions. For these stimuli, a large difference between the two comparison conditions was not expected.

In sum, the current experiments tested the common assumption that the same dimensions were used to describe an object whenever it was compared to other objects. Violations of this assumption were expected when certain, typically backgrounded, dimensions were foregrounded by variation along them in a set of objects.

Method

Subjects. Ninety-eight undergraduate students from Indiana University, divided evenly into the two-way and three-way conditions, served as subjects in order to fulfill a course requirement.

Materials. Sets of three stimuli were designed in accordance with five constraints (see Figure 3). First, the three objects varied on two dimensions (although the dimensions sometimes were composed of two dimensions that varied together), dimensions X and Y. Second, one of the objects (A) had exactly the same value on dimension Y as another object (B), and B had exactly the same value on dimension X as a third object (C). Third, an attempt was made to make the difference between A and B on dimension X larger than the difference between A and C on dimension Y. Fourth, an attempt was made to design dimension Y such that it would be likely to be "backgrounded" or ignored when no variation along it was present. Fifth, the simultaneous presence of dimensions X and Y did not create emergent features that could be responsible for the similarity of A-C to be greater than the similarity of A-B (see note 2). Appendix A shows the nine sets of "backgrounded dimension" stimuli used in the experiment.

Appendix B shows the six sets of control stimuli used in the experiment. The logic for constructing these items was different from that used for the nine experimental items and involved three constraints. First, object B was more similar than object C to object A on one dimension, and object C was more similar than object B to object A on the other dimension. Second, the similarity between A and B was chosen to be somewhat greater than the similarity between A and C. Third, both of the dimensions along which B and C differed were chosen so as to be foregrounded. For example, in the top set, B is more similar than C is to A on shape, and C is slightly more similar to A on shading, and both of these dimensions are likely to be noticed and used by participants in the two-way comparison condition.

The objects, approximately 5 × 5 cm, were presented on a Macintosh II SI monitor. The objects appeared side by side, separated horizontally by 8 cm. Subjects viewing distance was not controlled, but was approximately 11 cm.

Procedure. On each trial in the two-way rating condition, two objects appeared on the screen, either A and B or A and C, from one of the nine stimulus sets. No subject in the two-way rating condition ever received the A-B and A-C comparison from the same set. In the three-way rating condition, there were two rows of objects. One of the rows contained A and B, and the other row contained A and C. Subjects in the two-way rating condition were instructed to estimate the similarity of the two displayed objects. Subjects in the three-way condition were instructed to first rate the similarity of the top two objects and then to rate the similarity of bottom two objects.

Subjects gave their similarity ratings by moving a cursor on the screen with a mouse. A 20-cm horizontal line was drawn along the bottom of the screen. The left and right edges of the line were labeled "Not Similar At All" and "Highly Similar," respectively. Participants were instructed to press a button on the mouse when the cursor was positioned along the line at their subjective similarity estimate.

Several factors were randomized: which of the nine stimulus sets was presented on a particular trial, the spatial positioning (left or right) of the two objects within a comparison, and the spatial positioning (top or bottom) of the two comparisons for the three-way condition.

Results

Because of the rating technique used, similarity ratings were obtained on a scale of 1 to 550 (the total number of different positions on the horizontal line that was used as a scale). The results for both the backgrounded dimensions and control items are shown in Figure 4. For the items with backgrounded dimensions, the average similarities for A-B and A-C pairs in the three-way comparison condition were 274 and 224, respectively, showing overall monotonicity [unpaired $t(48) = 6.7$, $p < .001$]. The average similarities for A-B and A-C pairs in

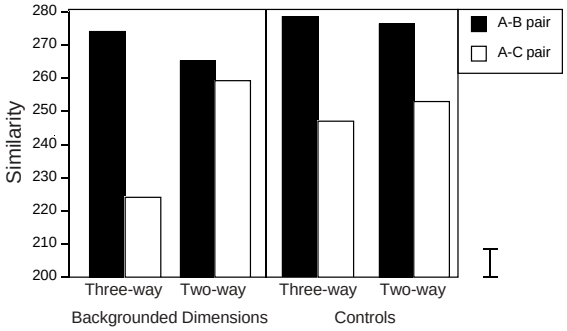


Figure 4. Results from Experiment 3A. The sets with backgrounded dimensions are shown in Appendix A; the control sets are shown in Appendix B. In the three-way comparisons, all three objects were presented simultaneously. In the two-way comparisons, only objects A and B or objects A and C were presented.

the two-way comparison condition were 265 and 259, respectively, nonsignificantly different in the direction of a monotonicity [paired $t(48) = 0.6$, $p > .3$]. These results reflect a significant comparison condition (two-way vs. three-way) by pair (A-B vs. A-C) interaction [$F(1,48) = 5.4$, $MS_e = 12$, $p < .01$], indicating that the similarity ratings for the pairs are significantly closer for the two-way than for the three-way condition.

For the control set of items, the average similarities in the two-way comparison condition for the A-B and A-C items were 278 and 247, respectively. For the three-way comparison condition, A-B and A-C similarities were 276 and 253, respectively. Thus, the comparison condition × pair interaction that was found for the backgrounded dimensions sets was not found for the control sets [$F(1,48) = 0.7$, $MS_e = 15$, $p > .3$], although there was a main effect of A-B pairs being rated as being more similar than A-C pairs [$t(48) = 6.3$, $p < .01$]. The difference between the control and backgrounded dimension items is evidenced by a significant three-way interaction between items (control vs. backgrounded dimensions), pair, and comparison condition [$F(1,48) = 5.6$, $MS_e = 21$, $p < .01$], accounted for by the fact that the A-B sets received significantly higher similarities than did the A-C sets at all times except in the two-way condition with backgrounded dimensions.

Discussion

Experiment 3A showed strong context effects on similarity ratings, but only for those stimuli that had dimensions that were likely to be backgrounded. Backgrounded dimensions are dimensions that are not noticed and used when there is no variation along them among the compared items. For these sets, shown in Appendix A, the relative similarity of A-B and A-C pairs is quite different, depending on whether the pairs are displayed simultaneously or in isolation. When displayed together, a strong monotonicity is observed, with A-B pairs receiving higher similarity ratings than A-C pairs, presumably because of their greater physical similarity on the dimension along which B and C differ. This dimension is foregrounded because variation along it is apparent in the context of the three items. However, when A and B are compared in isolation, this dimension is more likely to be backgrounded, and thus the similarity of A and B along the dimension is not increased much by this shared dimension. Consequently, in the two-way comparison condition, A-B pairs are not rated as being more similar than A-C pairs, despite their physically greater similarity.

It might be argued that differences between A-B and A-C pairs could be greater in the three-way, relative to the two-way, condition simply because this condition is more sensitive to dimension information in general. However, the control sets argue against this. The A-B pairs of the control sets were designed to be slightly more similar to the A-C pairs, and this difference is equally reflected in the two- and three-way conditions. In fact, for the isolated, two-way comparisons, the difference between A-B and A-C ratings is much greater for the control items than for the items with backgrounded dimensions, but this relationship is reversed for the three-way comparisons. Thus, the large context effect that is found for the critical items is probably not due to a general sensitivity difference between the conditions. The context effect is found only when dimensions that are unlikely to be considered in the two-way comparison between A and B are considered in the three-way comparison.

EXPERIMENT 3B

Experiment 3B was conducted to replicate the differences between the two- and three-way comparisons of Experiment 3A. In addition, the experimental task was altered in an attempt to promote nonmonotonicities. Experiment 3A revealed no increase in similarity due to the shared backgrounded dimension in the two-way comparison. However, in order to show a "true" violation of monotonicity, it is necessary to demonstrate significantly greater similarity estimates for pairs that do not share the backgrounded dimension than for pairs that do share this dimension.

To increase the probability of finding violations of monotonicity, it is important that the backgrounded dimension in the A-C pair be detected by participants and be influential in their judgments. To make variation on that dimension more salient, subjects were asked, prior to making their similarity judgments, to list the ways in which each pair of objects was similar. This technique was modeled after Wilson's "reasons listing" procedure (see Wilson, Dunn, Kraft, & Lisle, 1989), which has been used extensively in research on the effects of verbalizing knowledge.

An additional effect of these instructions is to increase the use of dimensions that are explicitly noticed, as opposed to using the overall similarity of the stimuli. This is potentially useful in creating nonmonotonicities, because it has been shown that requiring people to justify their judgments causes them to focus on a subset of dimensions that varies from trial to trial (Levine, Halberstadt, & Goldstone, 1996). Context effects can emerge

only if people differentially weight dimensions on different trials. To the extent that a general, overall consideration of dimensions occurs on each trial, context effects are unlikely to occur. Asking participants to justify their similarity assessments seems to be one way to disengage overall similarity assessments across the entire set of available dimensions.

Methods

Subjects. One hundred and eight undergraduate students from Indiana University served as subjects in order to fulfill a course requirement. Sixty and 48 participants were assigned to the two-way and three-way comparison conditions, respectively.

Materials and Procedure. The procedure was the same as that used in Experiment 3A, with the following exceptions. First, only the nine sets of items with backgrounded dimensions (shown in Appendix A) were used; no control sets were used. Second, prior to each similarity rating, subjects in both the two- and three-way conditions were instructed to "describe the features that these two objects have in common." The subjects described each common feature by typing the description on a line by itself and pressing the "return" key. They were required to list at least one common feature. After they were finished describing the commonalities, they typed "end" on a line by itself and proceeded to give their similarity rating.

Results

In the three-way comparison condition, the average similarities for A-B and A-C pairs were 278 and 232, respectively [unpaired $t(58) = 4.7$, $p < .001$]. In the two-way comparison condition, the average similarities for A-B and A-C pairs were 269 and 283, respectively [paired $t(47) = 1.5$, $p = .14$]. Thus, for the three-way comparison, a strong monotonicity was found, and for the two-way comparison, a nonsignificant trend in the direction of nonmonotonicity was found. In an item analysis, for the three-way comparison, all nine sets produced monotonicities, with A-B similarities being greater than A-C similarities. For the two-way comparison, five of the nine sets produced nonmonotonicities.

The results also indicated an asymmetry in similarity ratings that depended on the spatial positions on the screen of the compared objects. This asymmetry is shown in Figure 5. Collapsing across both conditions, when object A was on the left side of screen, average similarity was 262; when A was on the right, similarity was 269 [unpaired $t(107) = 2.4$, $p < .05$]. For the two-way comparison condition, there was a significant position (A on left vs. A on right) × displayed objects (A and B vs. A and C) interaction on similarity assessments [$F(1,107) = 4.1$, $MS_e = 8.2$, $p < .05$]. This same interaction for the three-way condition did not approach significance [$F(1,107) = 1.8$, $MS_e = 7.7$, $p > .1$]. If we restrict our attention to the trials in which A is on the right, the nonmonotonic trend observed in the two-way condition becomes significant, with B-A and C-A comparisons obtaining similarity assessments of 271 and 290, respectively [unpaired $t(58) = 2.2$, $p < .05$].

The two- and three-way comparison conditions can also be compared by treating the results as though they were obtained from a forced-choice similarity judgment.

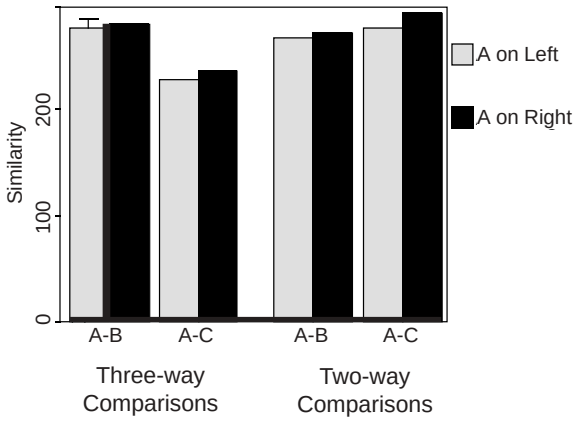


Figure 5. Results from Experiment 3B.

Trials from the same stimulus set were randomly yoked across subjects in the two- and three-way comparison conditions. In the three-way condition, A-B displays were rated as being more similar than their randomly yoked A-C displays for 86% of the sets (A-B displays were rated as more similar than their naturally accompanying A-C displays for 93% of the sets). In the two-way condition, only 43% of A-B similarity ratings were greater than the yoked A-C ratings. A condition (two- vs. three-way comparison) × comparison (A-B vs. A-C) ANOVA on the randomly yoked sets revealed a significant interaction [$F(1,107) = 7.2$, $MS_e = 9.4$, $p < .01$], indicating that the difference between A-B and A-C comparisons was greater in the three-way than in the two-way condition.

Discussion

Although Experiment 3B did not find an overall significant nonmonotonicity in the two-way comparison condition, strong context effects were found. First, even though the addition of a unique feature to C decreased similarity (in accordance with the monotonicity assumption) in the three-way comparison, it tended to increase similarity when the A-B and A-C comparisons were judged in isolation. This difference reached significance when A appeared on the right side of screen (a result discussed below). When the data were treated as forced-choice similarity judgments, participants adhered to the monotonicity assumption significantly less often in the two-way than in the three-way condition.

The effects in this experiment and Experiment 3A can be understood in terms of compared items eliciting different contrast sets that are used for evaluating the items' similarity. When objects that differed widely on only one of two dimensions (objects A and B) were compared, subjects tended to neglect their common dimension (particularly if the common dimension could be backgrounded). When objects differed widely on one dimension and slightly on another dimension (objects A and C), then both dimensions were considered relevant to the comparison. When subjects were encouraged to consider

the same dimensions in both comparisons (in the three-way comparison condition), the same comparison standards were used, and A and C were judged to be much less similar than A and B, in accordance with the monotonicity assumption.

Some alternative accounts for the different patterns of A-B and A-C comparisons in the two- and three-way comparison conditions can be rejected. First, the failure to obtain monotonicity in the two-way condition was not simply due to overly subtle dimension Y differences between objects A and C. Although differences along this dimension were designed to be small, they were large enough to be noticed 93% of the time in three-way comparisons.

Second, the obtained nonmonotonicity from the two-way condition (when analysis was restricted to trials where object A was displayed on the right) cannot simply be explained by greater weighting of dimension X than dimension Y. Dimensional stretching and shrinking (Nosofsky, 1986) cannot produce nonmonotonicities, and also cannot explain the importance of dimension Y for the three-way comparison. If the dimensions are dynamically weighted according to comparison condition, then the results can be explained, but such a weighting account would probably involve an explanation similar to the one advocated here, based on the backgrounding of nonvarying dimensions.

Third, the results cannot be explained simply in terms of a process of averaging dimensional similarities, although, again, such a process can provide an explanation in conjunction with dimensional backgrounding. Research indicates that judgments may sometimes be made by averaging evidence from different aspects (Anderson, 1982; Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993). An averaging process can produce nonmonotonicities in that positively valenced information can lower subjective utilities. For example, a student with a superb GPA but unknown references may be rated 9 on a scale of 1–10, but the same student with a superb GPA and good references may be rated 8 if the information from GPA (subjective value = 9) is averaged with the information from references (subjective value = 7). An averaging process could account for the observed nonmonotonicities, but only if it is additionally assumed that the same dimension that is averaged into A-C judgments is not averaged into A-B judgments. That is, A-C pairs may be judged as fairly similar due to an averaging of one highly similar dimension and one moderately similar dimension, whereas A and B are judged as less similar because they have one moderately similar dimension (and one backgrounded dimension that is not noticed). In short, an averaging process explains the relatively high A-C similarity ratings, but only if our claim of context-driven encoding of dimensions is assumed.

One post hoc account for why departures from monotonicity were greater when object A was on the right is that the object on the left was the first object noticed, and it established the initial set of dimensions that were used for comparison. Furthermore, object A often contained a "default" value on dimension Y that would make the dimension more likely to be backgrounded. For example, for the *thickness* dimension, the default value is *1-pixel line*, the default for *length of object's bottom line* is *just long enough to connect to object*, the default for *position of face* is *centered*, and the default for *orientation of suspended ball* is *upright*. When the object that contains the default (A) is on the left, subjects may not encode the backgrounded dimension. As such, when the object on the right (C) is subsequently viewed, it seems quite different, because it appears to require the postulation of a dimension not previously considered. When C is on the left, dimension Y is not as likely to be backgrounded, and dimension Y will be considered before participants examine A on the right. In short, when C is on the left, the A-C difference appears to be only a slight difference of dimension values; when A is on the left, the A-C difference appears, at first, to be a difference of what dimensions are needed to describe the objects. Accordingly, the nonmonotonicity, because it requires high similarity estimates for A-C comparisons, is most significant when C is the first object considered.

Because the materials used in the experiment were selected for their tendency to produce nonmonotonicities, conclusions can be drawn only with regard to the existence, rather than frequency, of nonmonotonicity. Nevertheless, our results corroborate other findings of nonmonotonicities in judgment and reasoning (e.g., Kahneman et al., 1993). In many cases, the same notion of context-driven contrast sets seems to be at work. Slovic (1985) asked participants to rate the attractiveness of two bets separately: (1) a 7/36 chance to win $9, and (2) a 7/36 chance to win $9 and a 29/36 to lose 5¢. Participants rated the second bet as more attractive, even though it added only a negative outcome. As in the current results, adding a small negative outcome may make participants invoke a contrast set in which larger losses would have been possible, a possibility that they do not consider in the first scenario. Birnbaum, Coffey, Mellers, and Weiss (1992) found a similar nonmonotonicity with assessments of lotteries in which dimensions that possessed values of zero were assigned relatively low weight. Their findings are consistent with the idea that dimensions are neglected when they are given default values.

GENERAL DISCUSSION

The results of the three reported experiments are important for several reasons. First they indicate the contextualized nature of similarity judgment. In all three experiments, the alternatives present during a particular comparison influenced which dimensions were foregrounded, and therefore considered and weighted, in judgment. In Experiment 1, dimensions that were especially relevant for distinguishing between alternatives were foregrounded. In Experiment 2, dimensions that had greater variability within the set of presented items were foregrounded. In Experiment 3, dimensions were foregrounded by introducing variation along them among the compared items. Together, the experiments provide evi-

dence that dimensional salience is modified on line, at the time of a comparison, rather than being fixed by a priori physical properties.

Several specific levels of context appear to influence dimensional salience. Earlier trials create a context within which later trials are assessed. Contexts can also be defined by the alternatives present within a single trial. Tversky's (1977) diagnosticity effect and Experiments 1 and 2 demonstrate cases in which the diagnosticity of information within a set of alternatives changes the weighting of the information. Experiment 3 argues that even when other alternatives are not present to provide a context for the judgment, the compared items themselves establish their own context (see Kahneman & Miller, 1986). Isolated comparisons seem to be made by recruiting standards of comparison that define the dimensions and alternatives relevant for a particular trial. Overall, these context effects parallel those found for other types of judgment (Medin et al., 1995).

Although the three experiments explore quite different context effects, preliminary work can be made toward a process model of similarity (and perhaps of judgment more generally) that can account for all three results. In fact, many of the processing mechanisms that are required for some of the context effect are also required for others. In particular, all three experiments suggest the existence of a process that highlights dimensions that exhibit variability (mutability) within a context. When the magnitude of difference between choices' values on a dimension is great, then the dimension becomes relatively important (Experiment 1). When the number of qualitatively distinct values that a dimension takes is large, then the dimension similarly becomes important (Experiment 2). When variation along a dimension is introduced, then a dimension that may otherwise have been ignored comes to influence judgments (Experiment 3). As such, a full processing account of similarity judgments should contain a mechanism that dynamically alters the importance of dimensions as a function of their variability. In this way, the different violations of standard assumptions of choice and similarity models may be produced by the same basic process—a process that does not use the same dimension weights across comparisons but, rather, adjusts the weights to reflect the transitory diagnosticity of the dimension.

The experiments also provide evidence against some common assumptions made by models of similarity and choice. Assumptions of transitivity (Experiment 1), choice independence of irrelevant alternatives (Experiments 2 and 3), and perhaps monotonicity (Experiment 3) were violated. We do not wish to claim that these assumptions are necessarily commonly violated—customization of dimensions was used to find violations of transitivity, and materials were carefully designed to produce nonmonotonicities. Nor do we intend to leave the impression that context effects always and inevitably affect judgments. For example, although participants do change their similarity judgments due to the context provided by task instructions (Heit & Rubinstein, 1994; Melara, Marks, & Lesko, 1992), such shifts are often incomplete (Goldstone, 1994a). Subjects attend to stimulus properties even when they are required to ignore them (Egeth, 1966; Stroop, 1935), and overall similarity across many properties is used even when subjects are told specific properties to use for comparisons (Allen & Brooks, 1991; Sadler & Shoben, 1993). Although explicit instructions are one of the strongest contextual pressures for altering responses to properties, even these contextual influences have limits.

Nonetheless, the fact that we can observe systematic violations of such fundamental assumptions as transitivity and monotonicity is important, not because it undermines the concept of similarity, but because it emphasizes the need to consider the processing side of similarity. There is a temptation to conclude that, by shifting contexts, one can make any two objects have any degree of perceived similarity. This conclusion is misguided, because it continues the unfortunate practice of focusing solely on outcomes or judgments rather than the processes that give rise to them. The motivation for our studies is to develop processing underpinnings to combine with the structural underpinnings associated with MDS or featural models of similarity. In that sense, our work is in the spirit of alignment-based similarity models (e.g., Gentner, 1989; Gentner & Markman, 1994; Goldstone, 1994b; Holyoak & Thagard, 1989; Medin et al., 1993; Markman & Gentner, 1993a, 1993b) that tend to be more explicit about processes underlying comparisons.

CONCLUSION

The goal of the three experiments has been to explore three notions of context, three processes for determining what properties will be considered important for a comparison. At a general level, the processes are based on informational diagnosticity, but differ from Tversky's diagnosticity principle, which is related to the usefulness of features for creating potential categories. The processes involve the diagnosticity of a feature for distinguishing between candidate choices, the diagnosticity of a dimension in terms of its informativeness within a set, and the diagnosticity of a feature for an implied contrast set.

In previous work (Medin et al., 1993), we suggested that the properties that are used for evaluating similarity are fully fixed only after the comparison process has begun. Likewise, the current work argues that similarity is not determined solely by object representations that are fixed prior to the comparison episode. The first two experiments suggest that the alternatives present on a trial influence how salient particular features will be within an object. The last experiment goes further, suggesting that even when no alternatives are explicit, a comparison of two objects will evoke other dimensions and objects. Alternatives, when present, influence item representations and, when absent, are spontaneously generated. Contextual effects may be virtually inevitable, and so it behooves us to try to understand them.

REFERENCES

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, **120**, 3-19.

Anderson, N. H. (1982). *Methods of information integration theory*. New York: Academic Press.

Arrow, K. J. (1951). *Social choice and individual values* (Cowles Comission Monograph 12). New York: Wiley.

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 401-485). Hillsdale, NJ: Erlbaum.

Birnbaum, M. H., Coffey, G., Mellers, B. A., & Weiss, R. (1992). Utility measurement: Configural-weight theory and the judge's point of view. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 331-346.

Bransford, J. D., Franks, J. J., Vye, N. J., & Sherwood, R. D. (1989). New approaches to instruction: Because wisdom can't be told. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 470-497). Cambridge: Cambridge University Press.

Carroll, J. D., & Wish, M. (1974). Models and methods for three-way multidimensional scaling. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 57-105). San Francisco: Freeman.

Coombs, C. (1983). *Psychology and mathematics*. Ann Arbor: University of Michigan Press.

Egeth, H. E. (1966). Parallel versus serial processes in multidimensional stimulus discrimination. *Perception & Psychophysics*, **1**, 245-252.

Ennis, D. M. (1992). Modelling similarity and identification when there are momentary fluctuations in psychological magnitudes. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 279-298). Hillsdale, NJ: Erlbaum.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. New York: Wiley.

Gati, I., & Tversky, A. (1982). Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception & Performance*, **8**, 325-340.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity, analogy, and thought* (pp. 199-241). New York: Cambridge University Press.

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

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

Goldstone, R. L. (1994b). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 3-28.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relations, attributes, and the non-independence of features in similarity judgments. *Cognitive Psychology*, **23**, 222-264.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects* (pp. 23-32). New York: Bobbs-Merrill.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 411-422.

Helson, H. (1964). *Adaptation-level theory*. New York: Harper & Row.

Helson, H., Michels, W. C., & Sturgeon, A. (1954). The use of comparative rating scales for the evaluation of psychophysical data. *American Journal of Psychology*, **67**, 321-326.

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

Imai, S. (1977). Pattern similarity and cognitive transformations. *Acta Psychologica*, **41**, 433-447.

James, W. (1950). *The principles of psychology: Volume I*. Dover: New York. (Original work published 1890)

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, **4**, 401-405.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, **93**, 136-153.

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, **85**, 450-463.

Lehrer, A., & Kittay, E. F. (1992). *Frames, fields and contrasts*. Hillsdale, NJ: Erlbaum.

Levine, G. M., Halberstadt, J. B., & Goldstone, R. L. (1996). Reasoning and the weighting of attributes in attitude judgments. *Journal of Personality & Social Psychology*, **70**, 230-246.

Markman, A. B., & Gentner, D. (1993a). Splitting the differences: A structural alignment view of similarity. *Journal of Memory & Language*, **32**, 517-535.

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

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, **100**, 254-278.

Medin, D. L., Goldstone, R. L., & Markman, A. (1995). Comparison and choice: Relations between similarity processes and decision processes. *Psychonomic Bulletin & Review*, **2**, 1-19.

Melara, R. D., Marks, L. E., & Lesko, K. (1992). Optional processes in similarity judgments. *Perception & Psychophysics*, **51**, 123-133.

Mellers, B. A., & Biagini, K. (1994). Similarity and choice. *Psychological Review*, **101**, 505-518.

Mellers, B. A., Change, S., Birnbaum, M., & Ordonez, L. (1992). Preferences, prices, and ratings in risky decision making. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 347-361.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, **63**, 81-97.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, **72**, 407-418.

Pomerantz, J. R., & Lockhead, G. R. (1991). Perception of structure: An overview. In G. R. Lockhead & J. R. Pomerantz (Eds.), *The perception of structure* (pp. 2-23). Washington: American Psychological Association.

Sadler, D. D., & Shoben, E. J. (1993). Context effects on semantic domains as seen in analogy solution. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 128-147.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. Part I. *Psychometrika*, **27**, 125-140.

Sjöberg, L. (1972). A cognitive theory of similarity. *Göteborg Psychological Reports*, **2**(10), 1-23.

Slovic, P. (1985). *Violations of dominance in rated attractiveness of playing bets* (Decision Research Report 84-6). Eugene, OR: Decision Research.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, **18**, 643-662.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, **76**, 31-48.

Tversky, A. (1977). Features of similarity. *Psychological Review*, **84**, 327-352.

Wedell, D. (1994). Context effects on similarity judgments of multidimensional stimuli: Inferring the structure of the emotion space. *Journal of Experimental Social Psychology*, **30**, 1-38.

Wedell, D. H. (in press). Contrast effects in paired comparisons: Evidence for both stimulus-based and response-based processes. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 1158-1173.

Wiener-Ehrlich, W. K., & Bart, W. M. (1980). An analysis of generative representation systems. *Journal of Mathematical Psychology*, **21**, 219-246.

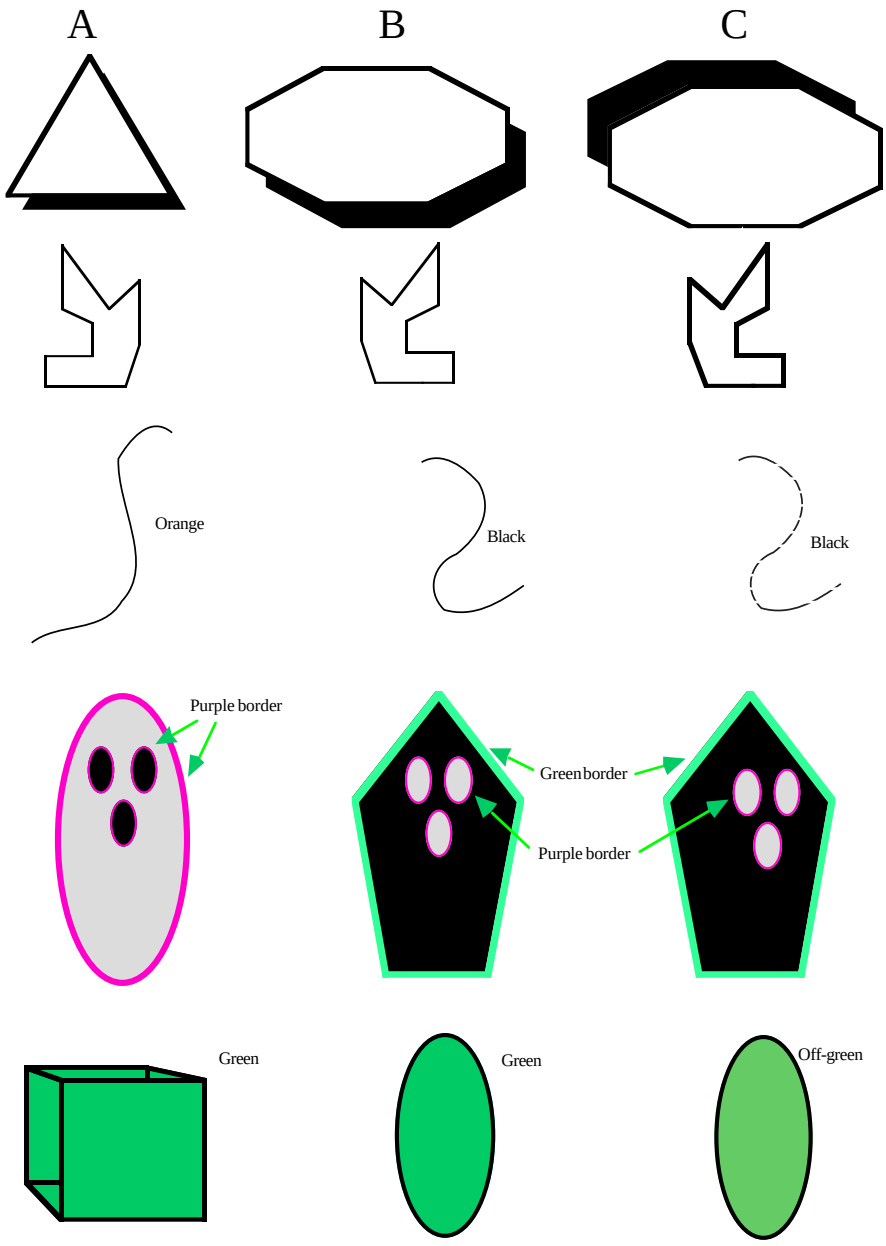
Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123-205). Orlando, FL: Academic Press.

NOTES

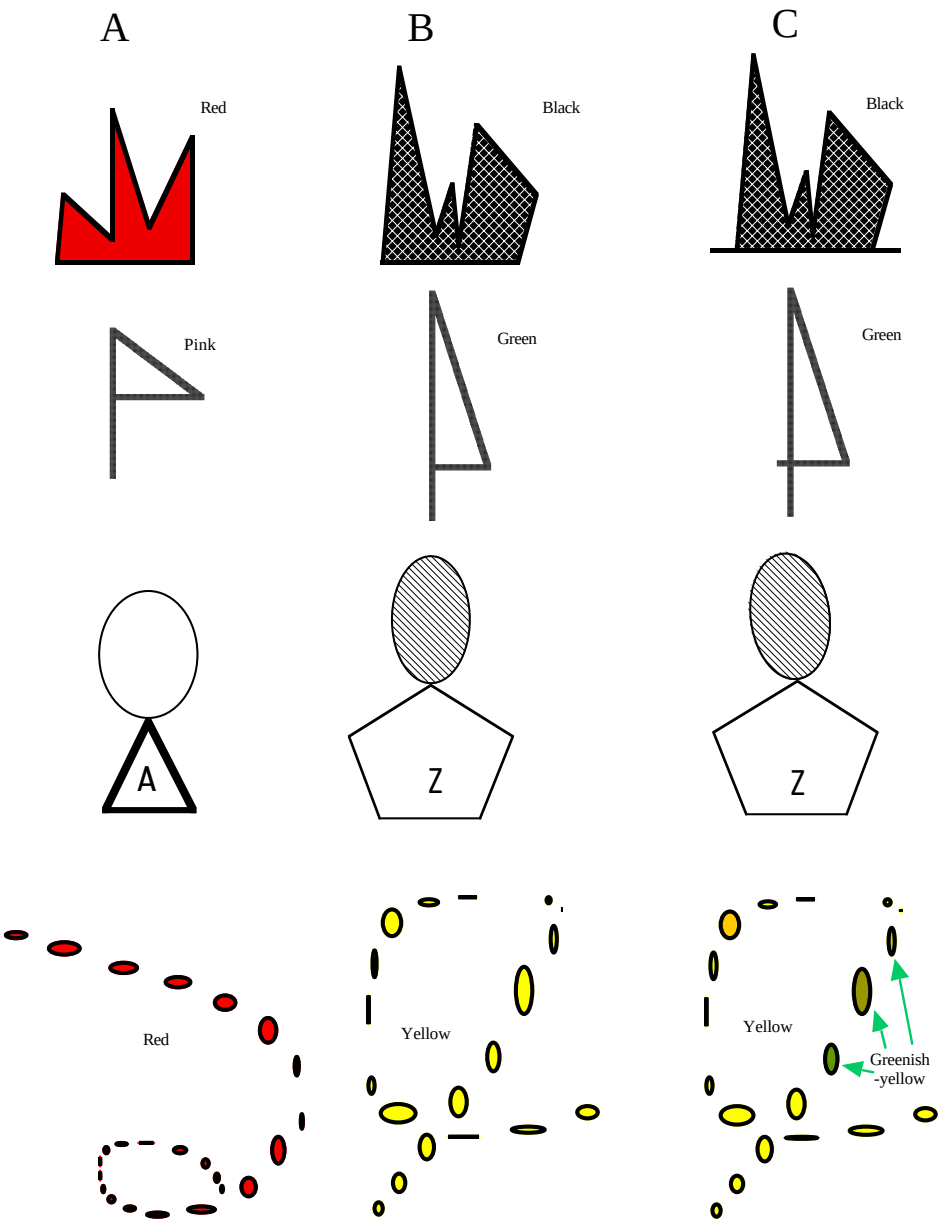
1. These data do violate moderate stochastic transitivity, according to which $P(A,C)$ should be at least as large as the minimum of $P(A,B)$ and $P(B,C)$.

2. This potential effect based on expandable contrast sets is theoretically distinct from "false" nonmonotonicities that can be produced by the emergence of a more abstract dimension in the stimuli. For example, Goldstone, Medin, and Gentner (1991) found that an item similar to "XX" is judged to be more similar to "YY" than it is to "XY." Although "XX" and "XY" share a feature that "XX" and "YY" do not (an "X" in the first position), this is not a true nonmonotonicity because "XX" and "YY" have an emergent common feature that "XX" and "XY" do not: two identical letters. This would be a genuine nonmonotonicity only if this abstract, relational feature were not psychologically important, but existing evidence strongly suggests otherwise (Gentner, 1989; Markman & Gentner, 1993a, 1993b). The nonmonotonicity tested in the current experiment is not based on the emergence of new dimensions, but rather on changes in the salience of existing dimensions due to changing contrast sets.

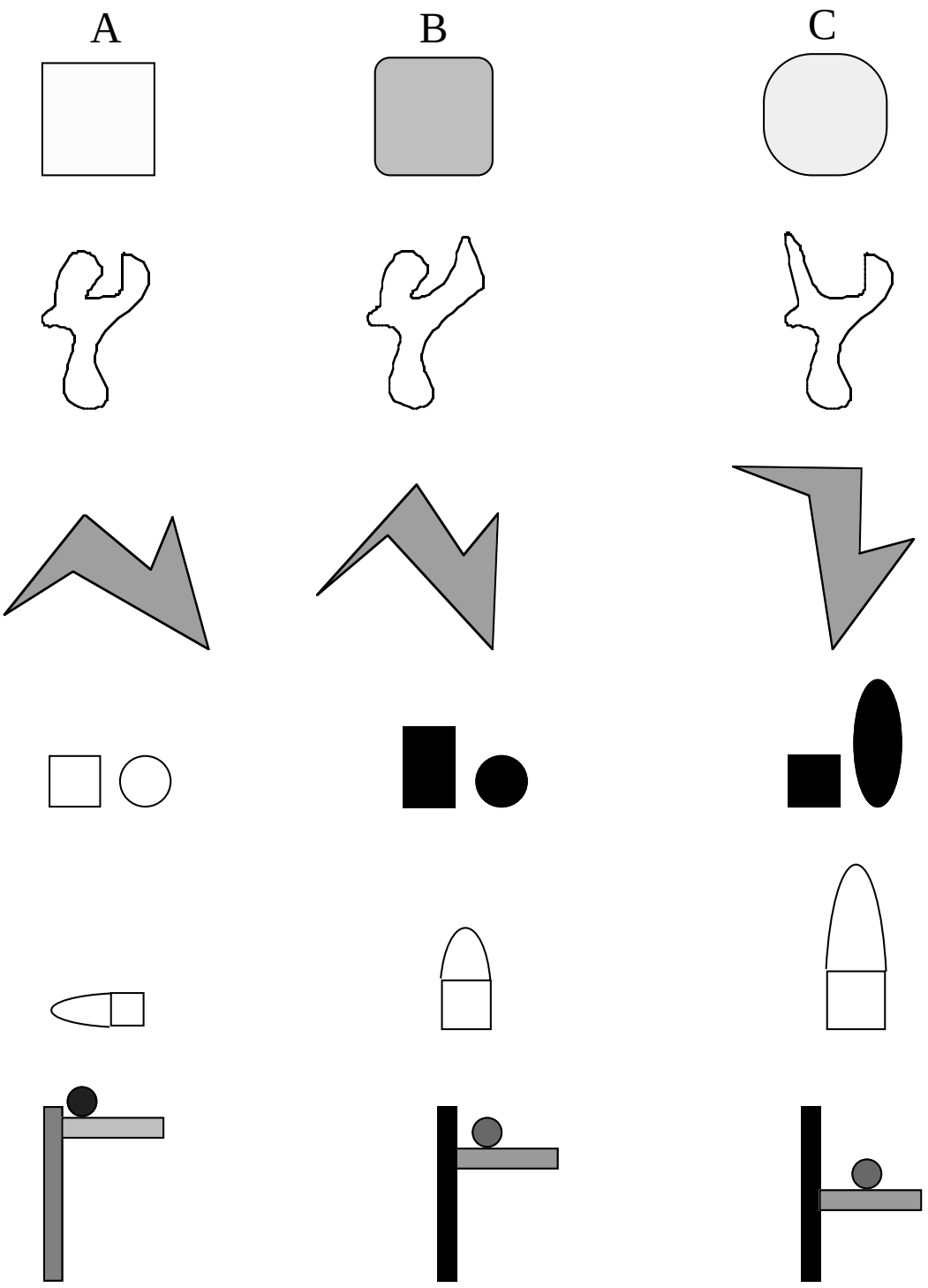
APPENDIX A
Stimuli for Experiment 3



APPENDIX A (Continued)



APPENDIX B
Control Stimui for Experiment 3



(Manuscript received June 28, 1995;
revision accepted for publication March 21, 1996.)